e-ISSN: 2148-7456

a peer-reviewed online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 9 Special Issue 2022

https://dergipark.org.tr/en/pub/ijate

https://ijate.net

IJATE

e-ISSN 2148-7456

https://dergipark.org.tr/en/pub/ijate
https://www.ijate.net

Volume 9 Special Issue 2022

Editor : Dr. Omer KUTLU
Address : Ankara University, Faculty of Educational Sciences,
Cebeci Yerleşkesi, Çankaya, Ankara, Türkiye

E-mail : ijate.editor@gmail.com; omerkutlu@ankara.edu.tr

Publisher Info : Dr. Izzet KARA
Address : Pamukkale University, Faculty of Education,
Kinikli Yerleskesi, 20070, Denizli, Türkiye
Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ikara@pau.edu.tr

Frequency : 4 issues per year (March, June, September, December)
Online ISSN : 2148-7456
Website : http://www.ijate.net/
http://dergipark.org.tr/en/pub/ijate

Journal Contact : Dr. Eren Can AYBEK
Address : Department of Educational Sciences, Pamukkale University,
Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye
E-mail : erencanaybek@gmail.com
Phone : +90 258 296 1050

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed and academic online journal.

The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).

International Journal of Assessment Tools in Education

International Journal of Assessment Tools in Education (IJATE) accepts original research on the design, analysis and use of evaluation along with assessment to enhance comprehension of the performance and quality of stakeholders in educational settings. IJATE is pleased to receive discriminating theoretical and empirical manuscripts (quantitative or qualitative) which could direct significant national and international argumentations in educational policy and practice.

IJATE, as an online journal, is hosted by DergiPark [TUBITAK-ULAKBIM (The Scientific and Technological Research Council of Türkiye)].

In IJATE, there are no charges under any procedure for submitting or publishing an article.

Indexes and Platforms:

- Emerging Sources Citation Index (ESCI)
- Education Resources Information Center (ERIC)
- TR Index (ULAKBIM),
- EBSCO,
- SOBIAD,
- JournalTOCs,
- MIAR (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib
- Index Copernicus International

International Journal of Assessment Tools in Education dedicates this special issue to **William James Popham** and **Thomas Haladyna**, who have contributed to classroom assessment with their publications.

Foreword to special issue / Özel sayıya sunuş

Omer Kutlu[1]

[1]*Editor*;
Ankara University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

Dear reader/audience,

This special issue of IJATE was prepared under the title of "Classroom Measurement and Assessment". This title has its own place as it is the very first attempt to be addressed in an international peer-reviewed journal in Türkiye. It is my wish that this issue be focused on more and more in the coming days. First of all, I would like to express my heartfelt thanks

to the special guests of this special issue: Dr. William James Popham and Dr. Thomas Haladyna. Both scientists have presented original views in their books and articles they have written since the 1990s that will affect classroom achievement. These two scientists have made critical contributions to this field while the change and innovative transformations regarding classroom learning have been happening. These publications shed light not only on the present but also on the future with the power they derive from their past knowledge. I came across their work in the mid-1990s. Today, when I stop and think, I can see their impact on my views regarding learning, students, teachers, schools, measurement, and assessment processes and even education systems. Lucky me, all my PhD students, (now academics) whom I supervised and with whom I had the opportunity to work got the chance to know Popham and Haladyna. For this reason, in the person of scientists who have contributed to measurement and assessment and psychometrics, this special issue is dedicated to Dr. W. James Popham and Dr. Thomas Haladyna. Thank you Popham, thank you Haladyna…

My second thanks go to all my academic friends who are interested in this special issue as I would like to see their work appropriate for this issue, and of course my special thanks go to our reviewers who evaluate these studies. I would like to state that this cooperation, which we carry out with care, is very valuable. We are passing through the days when Turkey's understanding of measurement and assessment is narrowed to

Değerli okuyucu,

IJATE bünyesinde hazırlanan bu özel sayı “Sınıf içi Ölçme ve Değerlendirme” başlığı altında hazırlandı. Bu başlık uluslararası hakemli bir dergide ele alınması açısından Türkiye’de bir ilk olma özelliği de taşıyor. Dileğim bu konunun önümüzdeki günlerde artan ölçülerde daha fazla ele alınmasıdır. Öncelikle teşekkürlerimi iletmek isterim.

İlk teşekkürüm bu özel sayının iki özel konuğuna; Dr. William James Popham ve Dr. Thomas Haladyna. Her iki bilim insanı da özellikle 1990’lı yıllardan itibaren yazdıkları kitaplar ve makaleleriyle sınıf içi başarıyı etkileyecek özgün görüşler sunmuşlardır. Sınıf içi öğrenmelerle ilgili değişimin ve yenilikçi dönüşümlerin yaşandığı bu yıllarda bu iki bilim insanının bu alana oldukça kritik katkıları olmuştur. Bu yayınlar geçmiş bilgi birikiminden aldığı güçle yalnızca bugüne değil geleceğe de ışık tutmuşlardır. Ben onların bu çalışmalarıyla 1990’lı yılların ortasında karşılaşmıştım. Geldiğim noktada, durup düşündüğümde öğrenmeye, öğrenciye, öğretmene, okula, ölçme ve durum belirleme süreçlerine ve hatta eğitim sistemlerine dair onların üzerimdeki etkisini görebiliyorum. Ne mutlu ki danışmanlığını yürüttüğüm ve çalışma fırsatı bulduğum tüm doktora öğrencilerim (ki şu an da hepsi birer akademisyen) Popham ve Haladyna ile tanışma fırsatı buldular. Bu nedenle bu özel sayı ölçme ve durum belirlemeye, psikometriye emek vermiş bilim insanlarının şahsında, Dr. W. James Popham ile Dr. Thomas Haladyna’ya ithaf edilmiştir. Teşekkürler Popham, teşekkürler Haladyna…

İkinci teşekkürüm bu özel sayıya ilgi duyan ve çalışmalarını bu sayı için uygun gören tüm akademisyen arkadaşlarıma ve tabii ki bu çalışmaları değerlendiren hakemlerimize. Özenle yürüttüğümüz bu işbirliğinin çok değerli olduğunu belirtmek isterim. Türkiye’nin ölçme ve durum belirleme anlayışının istatistiksel çözümlemelere indirgendiği günlerden geçiyoruz. İçinde bulunduğumuz 21. yüzyılda bilim alanımız ön statistical analysis. In the 21st century we are in, our field has come to the fore. Every nation has a wide variety of educational problems. The deficiencies of students in basic life skills are the leading problem. It is important for schools to diversify their functions in eliminating these deficiencies, strengthening assessment for monitoring purposes, prioritizing feedback, and disseminating the use of items based on higher-order thinking processes. For these practices that will strengthen education systems, Turkish scientists should focus on realistic education problems. In this context, these studies will guide how within-class measurement and assessment practices should be structured in a way that they enrich student achievement. Sincere thanks to my authors and reviewers for this very first attempt…

National programs such as Monitoring and Assessment of Academic Skills (ABIDE), which provide information about student achievement in Turkey, and the international ones such as Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and The Project of International Reading Language Skills (PIRLS) clearly reveal that there are problems in the development of student achievement. For example, when the PISA findings are analyzed as a whole, it is seen that Turkish students have both performed below the Organization for Economic Co-operation and Development (OECD) average since 2003, and 50% of the students have remained at or below the 2nd proficiency level, which is accepted as the basic level in PISA (MEB, 2005; MEB, 2007; MEB, 2010; OECD, 2014; OECD, 2016a; OECD, 2018; OECD, 2019). According to the OECD (2018) report, the percentages of Turkish students at proficiency levels 5 and 6 are as follows: Science literacy is 2.5%, mathematical literacy is 4.8%, and reading skills are 3.3%.

This situation should be accepted as a report card of the Turkish education system. It is possible to change it and carry students to higher proficiency levels. The findings of the national and international student assessments that Turkey has held or participated since the 2000s provide important clues in this regard. Özer et al., (2020) state in their study that there are inequalities in education in Turkey as in other countries. In the study, attention is drawn to the effect of socioeconomic background on academic achievement. It is emphasized that inequalities in

plana çıkmıştır. Her ulusun çok çeşitli eğitim sorunları vardır. Bu sorunların başında öğrencilerin temel yaşam becerilerindeki eksiklikleri gelmektedir. Okulların öğrencilerin bu eksiklerini gidermedeki işlevlerini çeşitlendirmesi, izlemeye dayalı durum belirleme yaklaşımlarını güçlendirmesi, geribildirime öncelik vermesi, üst düzey düşünme süreçlerine dayalı maddelerin kullanımının yaygınlaştırılması önemlidir. Eğitim sistemlerini güçlendirecek bu yaklaşımlar için Türk bilim insanlarının gerçekçi eğitim sorunlarına eğilmesi gerekmektedir. Bu kapsamda sınıf içi ölçme ve durum belirleme anlayışlarının öğrenci başarısını zenginleştirecek biçimde nasıl yapılandırılması gerektiğine bu araştırmalar yol gösterici olacaktır. Bu ilk girişim için yazarlarıma ve hakemlerimize içten teşekkürler…

Türkiye'de öğrenci başarısı hakkında bilgi veren Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABIDE) gibi ulusal, Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), The Project of International Reading Language Skills (PIRLS) gibi uluslararası çalışmalar öğrenci başarısının gelişiminde sorunlar olduğunu açıkça ortaya koymaktadır. Örneğin PISA bulguları bir bütün olarak incelendiğinde Türk öğrencilerin 2003 yılından itibaren hem Organisation for Economic Co-operation and Development (OECD) ortalamasının altında performans gösterdiği hem de öğrencilerin %50'sinin PISA'da temel düzey olarak kabul edilen 2. düzey ve altında kaldığı görülmektedir (MEB, 2005; MEB, 2007; MEB, 2010; OECD, 2014; OECD, 2016a; OECD, 2018; OECD, 2019). OECD (2018) raporuna göre, Türk öğrencilerin 5 ve 6 yeterlik düzeyinde bulunma yüzdeleri şöyledir: Fen okuryazarlığı %2.5, matematik okuryazarlığı %4.8 ve okuma becerileri %3.3'tür.

Bu durum Türk eğitim sisteminin de bir karnesi olarak kabul edilmelidir. Bu karneyi değiştirmek ve öğrencileri üst yeterlik alanlarına taşımak olanaklıdır. 2000'li yıllardan itibaren Türkiye'nin yaptığı ulusal ve katıldığı uluslararası öğrenci başarısını belirleme sınavlarının bulguları bu konuda önemli ipuçları vermektedir. Özer et al., (2020) çalışmalarında, diğer ülkelerde olduğu gibi Türkiye'de de eğitimde eşitsizlikler bulunduğunu belirtmektedir. Çalışmada, sosyoekonomik arka planın akademik başarıya etkisine dikkat çekilmektedir. Eğitimdeki eşitsizliklerin temel eğitim öncesinden başladığını belirtmektedir.

education start before basic education/primary education. Socioeconomically disadvantaged families have limited access to preschool education; therefore, the education life of a first grader starts with disadvantages from the beginning. It is also stated that since these inequalities are not sufficiently compensated in basic education, the gap in achievement has been gradually widening.

Studies draw attention to the factors that lower the students' achievement in the classroom. For example, Sarıer (2016) states three major factors that have a role on student achievement, which are "student" (motivation, self-efficacy, self-esteem, study habits), "school" (attitude towards the lesson, teacher leader, school culture, school principal's leadership) and "family" (socioeconomic level, family involvement in education, attitudes and behaviors of the family, education level of the mother and father). OECD (2016b), on the other hand, drew attention to the variable of "education system" (physical, educational resources, student selection, school management and financials) in addition to the student and school variable. Akey (2006), Rumberger, and Rotermund (2012) noted that affective characteristics such as perseverance, motivation, courage and self-efficacy play an important role in using learning opportunities effectively.

Within-class assessment processes are very important in many educational aspects such as measuring, evaluating, monitoring student learning, and giving feedback to the student. Kutlu and Altıntaş (2021) state that students should have three indispensable features such as learning, understanding, bringing together what has been learned and using them in life. For this reason, approaches that will improve students' cognitive, internal and interpersonal skills should be given importance in the classroom assessment processes. Waugh and Gronlund (2013) noted that assessment processes not only improve students' cognitive capacities but also their metacognitive skills, making them more independent learners.

Although assessment process has been discussed for over 30 years, Bennett (2011) states that assessment practices are still evolving. Stiggins (2006) argues that assessment processes should not be known as just giving feedback; instead, they must keep up with the changing life, measured characteristics and learning styles and should teach students how to be a better achiever.

Sosyoekonomik yönden dezavantajlı ailelerin okulöncesi eğitime erişimlerinin kısıtlıdır; bu nedenle, birinci sınıf öğrencisinin eğitim yaşamı dezavantajlarla başlar. Ayrıca temel eğitimde bu farklar yeterince telafi edilmediği için başarı farkının giderek açıldığı da belirtilmektedir.

Yapılan çalışmalar öğrencinin sınıf içi başarısını düşüren etkenlere dikkat çekmektedir. Örneğin Sarıer (2016) "öğrenci" (motivasyon, öz yeterlik, benlik saygısı, ders çalışma alışkanlığı), "okul" (derse yönelik tutum, lider öğretmen, okul kültürü, okul müdürünün liderliği) ve "aile" (sosyoekonomik düzey, ailenin eğitime katılımı, ailenin tutum ve davranışları, anne ve baba eğitim düzeyi) olmak üzere üç temel etkenin başarı üzerinde rolü olduğunu belirtmiştir. OECD (2016b) ise öğrenci ve okul değişkenine ek olarak "eğitim sistemi" (fiziksel, eğitsel kaynaklar, öğrenci seçme, okul yönetimi ve gerekli para) değişkenine de dikkat çekmiştir. Akey (2006), Rumberger ve Rotermund (2012) kararlılık, güdü, cesaret ve özyeterlik gibi duyuşsal özelliklerin öğrenme fırsatlarını etkili kullanmada önemli rolünün olduğunu dile getirmektedir.

Sınıf içi durum belirleme süreçleri; öğrenci öğrenmelerinin ölçülmesi, belirlenmesi, izlenmesi, öğrenciye geribildirim verilmesi gibi eğitsel anlamda birçok açıdan oldukça önemlidir. Kutlu ve Altıntaş (2021) içinde bulunduğumuz yüzyılda öğrencilerin; öğrenme, anlama, öğrenilenleri bir araya getirerek yaşamda kullanma gibi vazgeçilmez üç özelliğe sahip olmaları gerektiğine dikkat çekmektedir. Bu nedenle sınıf içi durum belirleme sürecinde öğrencilerin bilişsel, içsel ve kişilerarası becerilerini geliştirecek yaklaşımlara önem verilmelidir. Waugh ve Gronlund (2013) durum belirleme süreçlerinin öğrencilerin yalnızca bilişsel kapasitelerini değil, aynı zamanda bilişötesi becerilerini de geliştirerek onları daha bağımsız öğrenenler konumuna getirdiğini belirtmektedir.

Durum belirleme süreci her ne kadar 30 yılı aşkın süredir tartışılsa da, Bennett (2011) uygulamaların hâlâ geliştiğini belirtmektedir. Stiggins (2006) durum belirleme süreçlerinin değişen yaşama, ölçülen özelliklere ve öğrenme biçimlerine ayak uydurarak yalnızca dönüt vermek olarak bilinmesinden uzaklaşılıp öğrencilere nasıl başarılı olabileceklerini de öğretmesi gerektiğini savunmaktadır.

Türk eğitim sistemi durum belirleme uygulamalarından elde edilen bulgulara dayanarak sınıf içi öğrenmeleri zenginleştirmeli ve öğrenci

The Turkish education system should enrich within class learning based on the findings obtained from assessment practices and take measures to increase student achievement at the international level. Countries which are developed in terms of socioeconomic level variables have entered the 21st century with new expectations since the last quarter of the 20th century from education. The century we live in attaches great importance to students having skills that can be used in real-life situations. For this reason, it is much more important to structure educational processes that enable the development of students' high-level thinking skills, rather than within classroom learning activities and measurement and assessment practices that keep students at a level of knowledge. (Haladyna, 1997; Kutlu, & Altıntaş, 2021; Kutlu et al., 2017; Kutlu, & Kartal, 2018; Nitko, 2001; Popham, 2000).

In this sense, I hope that this special issue will be interesting and instructive for all academics who are interested in classroom assessment processes and also for teachers who play the leading role in classroom assessment processes.

başarısının uluslararası düzeyde artmasını sağlayacak önlemler almalıdır. Sosyoekonomik düzey değişkenleri bakımından kalkınmış ülkeler, 20. yüzyılın son çeyreğinden itibaren, 21. yüzyıla, eğitimden yeni beklentilerle girmişlerdir. İçinde bulunduğumuz yüzyıl öğrencilerden edindikleri bilgileri gerçek yaşam durumlarında kullanabilecekleri becerilere sahip olmalarını önemsemektedir. Bu nedenle öğrencileri bilgi düzeyinde tutan sınıf içi öğrenme etkinlikleri ile ölçme ve durum belirleme uygulamaları yerine öğrencilerin üst düzey düşünme becerilerinin gelişimini sağlayan eğitsel süreçlerin yapılandırılması çok daha önemlidir. (Haladyna, 1997; Kutlu, & Altıntaş, 2021; Kutlu et al., 2017; Kutlu, & Kartal, 2018; Nitko, 2001; Popham, 2000).

Bu anlamda bu özel sayının sınıf içi durum belirleme süreçlerine ilgi duyan tüm akademisyenler ve sınıf içi durum belirleme süreçlerinin başrolü olan öğretmenler için ilgi çekici ve öğretici olacağını umut ediyorum.

Orcid

Omer Kutlu https://orcid.org/0000-0003-4364-5629

REFERENCES

Akey, T. M. (2006). *School context, student attitudes and behavior, and academic achievement: An exploratory analysis.* MDRC.

Bennett, R.E. (2011). Formative assessment: A critical review. *Assessment in Education Principles Policy and Practice, 18*(1), 5 25. http://dx.doi.org/10.1080/0969594X.2010.513678

Haladyna, T.M. (1997). *Writing test items to evaluate higher order thinking.* Allyn & Bacon.

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2017). *Measurement and evaluation: Performance and portfolio-based assessment [Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme]* (5th edition). Pegem Akademi Yayıncılık.

Kutlu, O., & Kartal, S.K. (2018). The prominent student competences of the 21st century education and the transformation of classroom assessment. *International Journal of Progressive Education, 14*(6), 70-82. https://doi.org/10.29329/ijpe.2018.179.6

Kutlu, Ö., & Altıntaş, Ö. (2021). A brief history of psychological measurements and an approach of classroom assessment in the 21st century [Psikolojik ölçmelerin kısa tarihi ve 21. Yüzyılda sınıf içi durum belirleme anlayışı]. *Trakya Eğitim Dergisi, 11*(3), 159 9-1620. https://doi.org/10.24315/tred.896121

MEB (MoNE) (2005). *PISA 2003 project: National final report [PISA 2003 projesi: Ulusal nihai rapor].* TC. Milli Eğitim Bakanlığı, Eğitimi Araştırma Geliştirme Dairesi Başkanlığı.

MEB (2007). *PISA 2006 international student assessment program: National preliminary report [PISA 2006 uluslararası öğrenci değerlendirme programı: Ulusal ön rapor].* TC. Milli Eğitim Bakanlığı, Eğitimi Araştırma Geliştirme Dairesi Başkanlığı.

MEB (2010). *PISA 2009 international student assessment program: National preliminary report [PISA 2009 uluslararası öğrenci değerlendirme programı: Ulusal ön rapor].* TC. Milli Eğitim Bakanlığı, Eğitimi Araştırma Geliştirme Dairesi Başkanlığı.

Nitko, A.J. (2001). *Educational assessment of students* (3rd edition). Uper Saddle River.

OECD (2014). *PISA 2012 results: What students know and can do- student performance in mathematics, reading and science* (Volume 1). OECD Publishing.

OECD (2016a). *PISA 2015 results* (Volume I): Excellence and equity in education. OECD Publishing.

OECD (2016b). *Low-performing students: Why they fall behind and how to help them succeed*. OECD Publishing.

OECD (2018). *The future of education and skills: Education 2030*. OECD Publishing.

OECD (2019). *PISA 2018 results (Volume I): What students know and can do*. OECD Publishing.

Özer, M., Gençoğlu, C., & Suna, E. (2020). Policies for alleviating educational inequalities in Turkey [Türkiye'de eğitimde eşitsizlikleri azaltmak için uygulanan politikalar]. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 39*(2), 294-312.

Popham, W.J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd Edition). Allyn and Bacon.

Rumberger, R.W., & Rotermund, S. (2012). Student engagement determinants and student outcomes. In S.L. Christenson, A.L. Reschly & C. Wylie (Eds.). *Handbook of research on student engagement.* Springer.

Sarıer, Y. (2016). Türkiye'de öğrencilerin akademik başarısını etkileyen faktörler: Bir meta-analiz çalışması [The Factors That Affects Students' Academic Achievement in Turkey: A Meta-Analysis Study]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(3), 1-19.

Stiggins, R. (2006). Assessment for learning: A key to motivation and achievement. *Edge, 2*(2), 3-19.

Waugh, C.K., & Gronlund, N.E. (2013). *Assessment of student achievement* (10th edition). Pearson.

CONTENTS

Original Articles

Research Articles

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 1–5

https://doi.org/10.21449/ijate.1196705

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Original article*

Formative assessment: Stalled by too few right-size reports

William James Popham [1,*]

[1]University of California, Graduate School of Education and Information Studies, Los Angeles, USA

ARTICLE HISTORY

Received: Jun. 16, 2022

Accepted: Aug. 10, 2022

Keywords:
Assessment,
Formative assessment,
Reporting,
Right size reporting.

Abstract: After providing key definitions well as substantial supportive evidence for the instructional process under consideration, this analysis identifies a serious shortcoming in the way that many U.S. educators are currently encouraging teachers' adoption of the formative-assessment process--a teaching approach informed by students' en route test performances during instruction. After identifying the basics of the formative assessment process, and the manner in which reports of students' en route assessment performances should be built, it is claimed that formative assessment will attain its much-lauded learning payoffs only when short reports, easily used by both teachers and students, are routinely employed.

1. INTRODUCTION

Formative assessment works. When teachers routinely test their students' emerging mastery of instructionally emphasized knowledge or subskills, then make any needed adjustments based on students' measured progress, such adjustments typically pay off. Indeed, classroom formative assessment might well represent education's most successful example of properly conceived ends-means thinking. When en route tests indicate that a teacher's instructional procedures (the means) aren't satisfactorily moving students toward mastery of designated curricular aims (the ends), then different instructional tactics are employed.

1.1. In Praise of Formative Assessment

Not only is formative assessment a potent analytically identifiably instructional strategy, but for more than 25 years, we have possessed heaps of hard, empirical evidence indicating that formative assessment works—and works well.

In 1998, Paul Black and Dylan Wiliam, two British researchers, published a comprehensive review of almost 10 years' worth of empirical research dealing with classroom assessment (Black and Wiliam, 1998). The following conclusion from their in-depth review succinctly sums up that analysis: "The research reported here shows conclusively that formative assessment does improve learning (Black and Wiliam, 1998)."

*Corresponding Author: W. James POPHAM ✉ wpopham@g.ucla.edu University of California, Graduate School of Education and Information Studies, Los Angeles, USA

e-ISSN: 2148-7456 /© **IJATE** 2022

But were those demonstrable learning improvements substantial or, perhaps, merely modest? Well, Black and Wiliam (1998) concluded from a welter of empirical studies that the student gains in learning triggered by formative assessment were "amongst the largest ever reported for educational interventions." That's high praise indeed.

Meanwhile, in the U.S., the enactment of the federal No Child Left Behind Act of 2002 called for American educators to promote substantial—and definitely measurable—improvements in their students' performances. Circumspect educators realized that formative assessment might well be the chief means for promoting the federal call for improved student test scores. Summing up, then, more than a decade's worth of research focused on instructional uses of students' testing—plus a congressional mandate demanding higher test scores—meant that many American teachers seemed nearly certain to hear about, and even install this relatively new assessment-rooted strategy advocated by early proponents of formative assessment.

By the late 1990s, then, many American educators began learning about the nature of formative assessment, while also encountering numerous research reports indicating that formative assessment was a sure-fire instructional winner. In addition, a major federal law had been enacted urging American educators to employ instructional tactics capable of bringing about substantial improvements in their students' test scores.

Rarely, in the history of U.S, public schooling, have research evidence and legislative demands meshed so fortuitously. Educational commentators of that era often opined that formative assessment, an empirically demonstrable assessment-rooted strategy for instructional improvement, would soon be seen in most American schools.

Yet, many of those optimistic prophecies were issues more than 20 years ago. And, although it was widely believed back then that formative assessment would be installed in many U.S. schools, this oft-voiced prophecy simply failed to flower. What went wrong?

The following analysis will identify one repairable shortcoming in our thinking about formative assessment that represents an important reason formative assessment, apart from the few early years of interest it drew from U.S. educators, has fallen far short of widely foreseen usage hopes. It will be argued that if this single shortcoming were to be rectified, the long-promised learning dividends of properly formulated formative assessment will have a far better chance of being realized.

1.2. Definitions: A Pair

Although, these days, most educators possess a general notion of what formative assessment is, and many of those educators understand that this strategy represents a measurement-spurred instructional approach, it is always useful to define the central focus of any analytic commentary. Accordingly, then, bedecked in boldfaced italics below, is a formal definition of what most educators mean these days when they employ the descriptor "formative assessment:"

Formative assessment, an ongoing process seeking intermittent evidence of students' emerging learning, is used by teachers to adjust their instructional procedures and/or by students to adjust their current learning tactics.

But there's one more label that needs defining. This is because what's being defined above overlooks a missing ingredient in most formative-assessment dissemination strategies. It is an ingredient that, if lacking, decisively limits the expanding implementation of formative assessment. At least in the U.S., regrettably, formative assessment is rarely accompanied by "right-size reporting." So, in a bow to even-handed definitions, what's meant by "right-size is presented below—predictably, in boldfaced italics.

Right-size reporting describes efficient methods of describing students' test performances so that report-users can easily arrive at defensible decisions regarding next-step instructional actions consonant with the test's intended use.

The advocacy of right-size reporting's use during formative assessment usually stems from a belief that the more teachers who employ formative assessment, the better taught will be those teachers' students and, therefore, the better those students will learn. Although properly conceived formative assessment can be employed by students as well as teachers, the following remarks apply chiefly to teachers' needs for right-size reporting.

Nonetheless, ask any teacher who has made a serious commitment to employing formative assessment for an extended period to comment on that experience. What you'll often hear in response from the teacher is that (1) the formative-assessment process was effective and (2) it required too much work from the teacher to frequently implement it. We are not surprised by the "effective" response, of course, but the "too much work" replies often come as a surprise. Yet, when we think hard about the most distinctive feature of the entire formative-assessment process, it is the use of ongoing tests to collect evidence indicating whether instructional modifications are needed.

Although, depending on the curricula aims being pursued and, of course, the particular students being taught, teachers typically determine how often to measure their students' progress. Typically, there will be one or two short-duration assessments (called en route tests) used during a week or so of instruction. If the instructional period at hand is at all lengthy, for instance, five or six weeks, this quickly translates into a hefty number of en route tests that must be administered, scored, and then employed to arrive at appropriate instructional decisions regarding instructional next steps. Where do those tests come from?

Putting aside for the moment the who-creates and who-scores issues, what attributes should en route tests possess if they provide right-size reports and, therefore, contribute to improved instructional decisions by teachers? Here, then, are three features that, if present, optimize the instructional contributions of en route tests employed during the formative-assessment process.

- Balanced Representation. The evidence reflecting the content, i.e., the knowledge and/or skills assessed, provides an accurate representation of this content.

- Suitable Numbers of Items. For whatever knowledge and/or skills are tested, sufficient but not excessive numbers of items are present.

- Actionability. Content of each item on a formative assessment's en route test, depending on a student's responses, suggests next step(s) for teachers.

Let's briefly consider these three attributes of the en route tests used during formative assessment because, as we will see, the pressures on teachers to incorporate truly first-rate tests have, surprisingly, led many educators to completely abandon use of the formative-assessment process. More about this shortly.

1.2.1. *Content representativeness*

First, students' responses to formative assessments' en route tests supply teachers with the evidence needed to make any necessary adjustments in ongoing instruction. Such evidence is, arguably, the essence of the formative-assessment process. It is clearly necessary, therefore, for formative assessment's en route tests to be accompanied by evidence, perhaps judgmental in nature, indicating the degree to which students' responses to a test's items will provide a sufficiently representative reflection of students' status regarding the en route targets being sought. Accordingly, credible evidence of some sort—perhaps gathered from a teacher's colleagues—should be routinely provided to indicate the representativeness of an en route test's items.

Through the last few decades, many teacher-review panels have been employed to judge the adequacy of test-items' content for the intended use of test's results—particularly for high-stakes tests. We have learned that formatively focused en route tests should also, if possible, have their items reviewed for content representativeness. For any particularly important en route tests, a suitable rating form plus a systematic orientation for reviewers' use is normally required. For less significant tests, a content-representativeness judgment from one or two content-knowledgeable colleagues is often sufficient.

What is being recommended here is that, whenever feasible, the content representativeness of formative assessment's tests be determined so that the teacher (as decision-maker) can determine how much confidence should be based on the evidence garnered by different en route tests.

1.2.2. *Item numbers*

One of the most vexing requirements facing teachers who use formative assessment hinges on a seemingly small problem, namely, how many items to employ in a teacher's en route tests. The potential mistakes made here are usually "too few" or "too many." If too few items are employed in an en route test, then it is unlikely that teachers can draw a valid inference about a test-taker's mastery of the content represented by such a tiny collection of a test's items. Conversely, if far too many items are included in en route tests, then students' performances on those tests may, in fact, accurately reflect students' content mastery but, because of an excessive number of en route items, such along-the-way testing takes far too much time—time that might otherwise be profitably spent on instruction.

1.2.3. *Actionability*

Teachers engage in formative assessment to help them discern whether instructional changes should be enacted and, if so, to decide which changes to make. In some instances, of course, students' performances will indicate that no instructional modifications are needed—because the teacher's students are learning wonderfully. But if some students' less than lustrous en route performances make it clear that instructional alterations are necessary, the teacher must then identify what instructional changes to make and determine when to make them.

Realistically, there are three main options to consider when dealing with next-step options based on the rests used as part of the formative-assessment strategy. First, if teachers are working alone, then such teachers will need to come up with—on their own—one or more next-step instructional options. Second, a group of teachers working collaboratively in the same school—or affiliated with the same school district—could also devise a set of potential instructional activities for students that would address test-isolated content or subskills in need of an instructional re-do. But if you were to ask many teachers who have taken part in such collaborative test-building to comment on such endeavors, you're almost certain to learn that these sorts of collaborative instruction-building efforts are, obviously, dependent on the individuals involved, particularly time-consuming and, often, not all that effective.

The third major source of potential next-step instructional activities are the many commercial products now being sold by both profit-making and non-profit organizations. Spurred often by the positive results from long-term formative assessment, numerous formatively oriented systems are currently being marketed so that both en route tests, as well as suggested instructional alternatives aimed at such targets, are now purchasable. However, in the attempts of commercial vendors to market their products to a sufficiently large and heterogeneous array of potential purchasers, almost all such purveyors of these sorts of ready-made materials for formative-assessment materials are obliged to create materials far too general to satisfy the needed accuracy of truly on-target instruction. The effect of educators' adopting such too-general instructional and/or assessment materials is that much of today's en route testing—and subsequent instructional amelioration—leads to far less successful learning than hoped.

1.3. Formative Assessment: An Obstacle and A Solution

Summing up, then, although we now possess ample evidence from many quarters indicating that formative assessment, a potent marriage of ends/means assessment and instruction, is capable of producing substantial improvements in students' learning (Black and Wiliam, 1998) we see far less real-world usage of classroom formative assessment than had been widely prophesied (Popham, 2008, 2011). It was claimed in this analysis that a prominent deterrent to teachers' expanded employment of classroom formative assessment was that its implementation requires a raft of classroom formative practices that are simply too difficult for most teachers to undertake.

Fashioning en route tests so that they do not require Herculean efforts to employ, yet provide evidence needed for making adroit next-step instructional decisions, is what's needed. Although the most significant factor in the formative-assessment process is the quality of the en route assessments being used and the evidence they provide, sufficient attention has simply not been given to how to provide teachers with formative tests, or how to report right-size results. Putting it more tersely, we need to make it easy for teachers to employ formative assessment. That's right, easy.

Scrutiny of the many introductory books devoted to formative assessment reveals scant attention given to the necessity of creating right-size reports and, moreover, little heed to sharpening the way in which right-size reports will be provided so that formative assessment makes a meaningful improvement in students' learning. It is, clearly, time to change our ways.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

William James Popham https://orcid.org/0000-0001-5822-0303

REFERENCES

Black, P., & Wiliam, D. (1998). Assessment and classroom learning, *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-73. https://doi.org/10.1080/0969595980050102

Popham, W.J. (2008). *Transformative assessment.* ASCD.

Popham. W.J. (2011). *Transformative assessment in action.* ASCD.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 6–18

https://doi.org/10.21449/ijate.1196701

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Original article*

Creating multiple-choice items for testing student learning

Thomas Haladyna[1,*]

[1]Arizona State University, Arizona, USA

ARTICLE HISTORY

Received: Sep. 10, 2022

Accepted: Sep. 28, 2022

Keywords:

Multiple choice items,
Classroom assessment,
Measuring student learning.

Abstract: The use of multiple-choice items for classroom testing is firmly established for many good reasons. The content any unit or course of study can be well sampled. Test scores can be reliable (trusted). And time spent administering and scoring can be minimized. This article provides a current review of best practices in the design and use of a variety of multiple-choice formats for classroom assessment of student learning. One of the most serious problems facing current educators is developing test items that measure more than simple factual recall. It is important to measure understanding, comprehension, critical thinking, and problem solving. Not only are these types of higher-level thinking described, but items are presented that illustrate how this is done. These best practices are continually evolving. The objective is always to use tests to measure validly what students have learned as well as help students learn what they have not yet learned. We call this formative and summative assessment. Guidelines are presented showing good and bad practices. What may be surprising to readers is the extensive variety of formats and methods for gathering or generating new test items. Readers are encouraged to experiment with these formats. Some formats can very efficiently measure what students were supposed to learn.

1. INTRODUCTION

This article concerns the writing and use of multiple-choice (MC) test items for evaluating student learning in a classroom or course of study. Best practices are described that lead to the development and validation of MC items that can be part of an inventory of test items for formative and summative evaluation of student learning. Formative refers to the use of test items to help students learn. Think of formative as practice and feedback on how much learning has occurred. Summative refers to a piece of evidence used with other evidence to assign a student grade.

This article is based my extensive study of the origin of item development, related research, and considerable experience both with testing programs and testing in the classroom in elementary and secondary schools and in universities and professional schools. Much of the background for this article comes from references provided at the end of this article.

*CORRESPONDING AUTHOR: Thomas Haladyna ✉ tmh@asu.edu 🖃 Arizona State University, Arizona, USA

e-ISSN: 2148-7456 /© **IJATE** 2022

1.1. Objectively-scorable Items

To start, the term *MC item* is limiting. A better term is objectively-scorable item (OSI). In this section, a family of OSIs is introduced and illustrated, one of which is the MC item format. Each is introduced, illustrated, and a brief comment is offered. The most comprehensive source of information about OSIs is in our book (Haladyna & Rodriguez, 2013).

1.1.1. *Conventional MC (CMC)*

The most basic OSI has a stem and three choices. Note that traditionally, the CMS has consisted had four or five choices. Extensive research over many years has led us to conclude that three choices are sufficient (Haladyna, Raymond, & Stevens, 2019; Rodriguez, 2016). These fourth and fifth choices typically fail to discriminate or are so implausible that no student would choose it. Creating fourth and fifth options is usually a waste of time for the item writer.

Which type of travel from Ankara to Istanbul is most economical? (STEM)

A. Bus (Correct)*

B. Train (Distractor)

C. Plane (Distractor)

The biggest objection to this recommendation is that guessing might influence the accuracy of a test score. A student might be a lucky or unlucky guesser. A recent review and study show that guessing is overrated as threat to the accuracy of any test score (Haladyna, submitted for publication). The average score for random guessing on a three-option item test is 33%. One would have to be extremely lucky to get a score much higher than deserved.

1.1.2. *Alternate choice (AC)*

Which type of travel is slowest on a trip from Ankara to Istanbul?

A. Personal auto

B. Bus

This format is most useful for higher-achieving students who ordinarily can narrow down any OSI test item to one or two plausible choices. Also, this item does not take up space and much reading time. The more items in a test, the more reliable the test score will be. With more items, you can cover more content as well.

An AC item can be modified under some circumstances in this way:

Which type of travel from Ankara to Istanbul is considered very economical?

A. Personal auto

B. Bus

C. Both A and B

D. Neither A nor B

This modification transforms the item into a four-option CMC.

1.1.3. *True-false (TF)*

This format has a questionable reputation that has limited it usefulness. However, stating a set of 30 declarative statements is convenient, about half of which are true and half of which are false. The efficiency of TF format is unmatched. Items are easy to write. Administration time is short. Scoring is easy. Reliability of test scores can be very high. If the items represent student learning outcomes, the test score is accurate.

The main problem with the TF format is the tendency to test factual recall instead of higher types of learning. Another criticism is random guessing. We must recognize that the floor of the test score scale is 50%, so performance at this level would show a lack of student learning. Our standards for evaluating student performance should be much higher than 50%.

Mark A if true and B if false.

1. Ankara is west of Istanbul.

2. Istanbul has a greater population compare with Ankara.

3. The climate of Ankara is generally warmer than Istanbul.

4. The climate of Ankara is rainier than Istanbul.

5. Rahmi M. Koç Museum is one of the best tourist attractions in Istanbul.

6. Gülhane Park is on the grounds of the Topkapi Palace.

1.1.4. *Multiple true-false (MTF)*

The MTF format is useful for testing a family of related characteristics or examples of a concept. The MTF format has a lead question or an open-ended statement followed by a list of choices. Each choice is marked true or false by the student.

Which of the following are true regarding travel from Ankara to Istanbul?

1. Bus travel is very slow.

2. Air travel is the most expensive.

3. Bus travel is the least expensive of all options.

4. Train travel is the most comfortable.

5. Train travel to Istanbul leaves you in the city center.

6. Car travel through Bursa is slower and more scenic.

Like TF items, the MTF is very easy to write, administer, and score. Test scores can be very reliable. As with TF, we have the annoying interference of lucky or unlucky guessing, but as previously noted, random guessing is overrated as a threat. A major limitation of the MTF is a tendency to focus content narrowly instead of broadly. The examples in the above item deal with travel between two cities.

1.1.5. *Matching*

This format is underutilized. Little research exists on its use. Nonetheless, it should not be left out of your collection of OSI formats. As you can see, matching has one set of choices and many stems. So, it is an efficient type of MC.

A. Ankara	*1. The most populous city*
B. Istanbul	*2. The national capital*
C. Izmir	*3. The fastest growing city*
D. Bursa	*4. The most beautiful city of the four listed above*
	5. The most popular city for visitors
	6. Near the sea of Marmara

Many more items can be added using the same four choices above. This format is very efficient regarding administration. Also, items are easier to write. Finally, test scores tend to have high reliability because many items are used.

1.1.6. *Extended matching*

The extended matching format is used in situations where many choices are available. The example shown below comes from a medical test of cardiovascular symptoms and signed.

Options:

A. Radiofemoral delay

B. Pan-systolic murmur
C. Systolic blood pressure of 220 mmHg
D. Tapping apex beat
E. Chest pain eased by glyceryl trinitrate in five minutes
F. Third heart sound
G. Splinter hemorrhages
H. Breathlessness eased by lying flat
I. Slow-rising carotid pulses
J. Bardycardia with pulse rate 20 per minute
K. Chest pain eased by glyceryl trinitrate after an hour

Which of the choices above best describes the patients below?

1. *65-year-old man collapsed when running. He has a sustained heaving apex beat that is slightly displaced and an ejection systolic murmur.*
2. *An 80-year- old woman has an excruciating pain between the shoulder-blades. You palpate the right radial pulse but not the left.*
3. *A 70-year-old man had a myocardial infarction two years ago. He now has gradually increasing breathlessness worse on lying flat with crepitations in the lung bases.*
4. *A 65-year-old woman has been increasingly breathless over the last few years. On an auscultation, she has a loud first heart sound and a mid-diastolic murmur.*
5. *A 60-year-old man who smokes 20 cigarettes per day. He complains of a tight pain in the center of his chest, which comes on when he walks up stairs.*

With more items, all the choices listed above can have associated stems. The benefit of the extended matching is the wide coverage given to a variety of problems involving the heart.

1.1.7. *Testlets*

The testlet is the most useful and desirable of all OSI formats. It is often used in tests that measure reading comprehension. A short vignette or story is introduced and a series of items follows that provide indications of the student's comprehension. In science, an experiment or scientific observation is presented. Then a series of items follows that comprise the testlet. In teaching statistics to graduate students, my tests were designed around problems where a statistical procedure was applied. A set of generic CMC test items was used. All I had to do was change values of the problem to generate a new testlet. This approach to writing items and testlet has grown more popular now that automated item generation is a reality (See Gierl & Haladyna, 2015). Testlets are widely used in virtually all testing situations where complex use of knowledge and skills is required.

An excellent source of examples of testlets can be found on the following website:

https://www.act.org/content/act/en/products-and-services/the-act/test-preparation/reading-practice-test-questions.html?page=0&chapter=0. A Google search of testlets will yield a wealth of examples.

Testlets are typically more than a page in length, so one will not be presented here. However, understanding the structure is important, so a skeletal version of a testlet is presented in Table 1.

Table 1. *A sample skeleton version of a testlet*

Planning a Family Vacation

Passage: You are planning travel for your family from Ankara to Istanbul for a four-day vacation. Your father and mother trust you to provide useful information in planning this exciting trip. For each item, pick the correct response.

Items (only stems are provided)

1. How many miles is the distance from Ankara to Istanbul?

2. Which type of transportation is least costly to travel

3. What kind of climate might expect for this time of year in Istanbul?

4. How long will the trip be if we travel by car?

5. How expensive is air travel?

6. What is the cost of train travel?

With any testlet, any OSI format can be used. Also, the number of items can be quite long for each testlet. I once observed an entire test consisting of one testlet involving a group of teenagers going to a fair in their village.

1.1.8. *Completion Items*

There is one OSI that has no choices. It is the simplest of the family of OSI formats. The completion item is simply a question or prompt where a correct answer or performance is noted.

What is the most valuable natural resource of Turkey?

About how many hours is a train trip from Ankara to Istanbul?

With the completion item, a single right answer or a small set of right answers exists. The completion item is often used for measuring skills. This is an application of an item format for performance.

1.1.9. *Complex MC*

Here is one format is that not recommended. It looks like MC but that combinations that make it more challenging. The strike-out shows that this item type should NEVER be used.

~~*Which of the following modes of transportation are very slow?*~~

~~*1. Bus*~~

~~*2. Car*~~

~~*3. Walking*~~

~~*A. 1 and 2*~~

~~*B. 1 and 3*~~

~~*C. 2 and 3*~~

~~*D. 1, 2, and 3*~~

There are many bad variations of this format (sometimes called Type K). This format is widely rejected. With the exception of the complex MC, the other OSI formats have attractive features that recommend their use.

1.2. Validity and Reliability

The most important concept in the measurement of student learning is validity. We have extensive discussions of validity in various sources. For the sake of brevity, these principles are address validity. This brief section is intended to provide more context for choosing and using OSI formats for measuring student learning.

Validity refers to the accuracy of an interpretation of a test score. The term *valid test* is inappropriate. We consider the evidence supporting the creation of that test score as an accurate measure of student learning. By carefully creating OSI items and using these items in a test to obtain a test score fairly, we make a claim that the test score is a valid (accurate) measure of student learning.

That said, reliability comes into play. Reliability refers to the degree of random error represented in a set of test scores. We cannot have a validity interpretation of a test score, if reliability is low. Let us disregard how to compute reliability. Any measure of student learning should have a low degree of random error (thus high reliability). To ensure this valuable piece of validity evidence, OSIs MUST be well written and representative of the domain of knowledge and skills a test is supposed to represent. Longer tests tend to have less random error. Items of appropriate difficulty for the students tend to reduce random error. That is, items should not be too hard or too easy.

Students need to be informed about what they are about to learn. They much need to have a way to identify and learn what you are teaching. This content may be a lesson, unit, topic, course, curriculum, textbook, other written materials. Think of content as existing in domain that consists of knowledge and skills. Students learn the content in that domain. A test is a fair, unbiased sample from that domain to ensure high validity. If you guarantee the students have received adequate instruction and the test fairly represents this content, valid test score interpretations are achieved.

1.3. Content

We can categorize all content that is taught into four convenient categories.

1.3.1. *Facts*

Are true statements verifiable by all. The square root of nine is three. The area of a square or rectangle is the length of one side times the length of the adjacent side. Ankara is 445 kilometers from Istanbul. The opposite of East is West. Earth is a planet. Facts are notoriously over tested. We might say that facts are over taught. Focusing on facts does not leave room for more important types learning.

1.3.2. *Concepts*

A concept is an idea. For example, love, peace, fruit, car, money, television are some examples of concepts. Each concept has a definition, distinguishing characteristics, and examples. Thus, testing for a concept involves distinguishing among concepts, definitions of a concept, characteristics of the concept, or examples and non-examples of the concept.

1.3.3. *Principles*

Principles are relationships that are causal. Some principles are absolute (axiomatic), and some principles are probabilistic.

> *The first step in trauma injury is to ensure the airway is open. (Axiomatic)*
>
> *The density of air depends on its elevation. (Axiomatic)*
>
> *As temperature declines, at some point, water turns to ice. (Axiomatic)*
>
> *What is the chance of survival in an car accident if a passenger is wearing a seat belt. (Probabilistic)*
>
> *Which factors contribute to heart disease? (Probabilistic)*
>
> *Wheat tends grow optimally under what conditions? (Probabilistic)*

All formats presented previously can be useful for testing principles, but the testlet is the most highly recommended. Unfortunately, the testlet is difficult to design. However, many testlets can be designed to have interchanging values that provide more usability. That is, we can vary values in a problem and create new problems and use a set of standardized questions as previously shown.

1.3.4. *Procedures*

A procedure is a set of mental or physical steps. OSIs are not suitable for measuring physical procedures. For mental procedures, we might ask a student to identify correct or incorrect sets of steps, or to identify a key feature of a procedure. Only the completion item is useful for measuring physical skills. The other OSI formats apply best to measuring knowledge.

We have plenty of understanding that facts are taught too much. Learning concepts is useful. Applying principles is more complex and very desirable in everyday life. Procedures are things we do every day and over time that have many steps. When you create a MC item, you will choose which of these four types of content will fit your purposes. Ultimately, we can to use facts, concepts, principles, and procedures in some combination that is complex. This leads us to mental complexity.

1.4. Mental Complexity

For every item, we assign a judgment of what type of mental complexity is required to choose the correct option. Of course, this is speculation, because every student has a different reaction to a MC item. The low-achieving student must use a higher degree of mental complexity in choosing a correct choice. The high-achieving student usually uses previous knowledge. Nonetheless, there is a premium of writing MC items with greater mental complexity because we want our students to use knowledge and skills in complex ways to solve problems, evaluate alternatives, decision-making, and thinking critically. Simply memorizing facts does not take us very far. Three types of mental complexity are briefly illustrated using the CMC format.

1.4.1. *Recall*

In Turkey, which river is the longest?

*A. Kizilirmak**

B. Euprhates

C. Tigris

This item may also be considered a trick item, because B and C are very long rivers but are shared by other countries. A is correct.

Items of this type are very easy to write and use. We have an abundance of recall items. Most educators admit that we tend to teach and test for recall instead of teaching for deeper and more complex types of student learning. Thus, recall items should be used sparingly.

1.4.2. *Understand (Comprehend)*

The focus here is a concept, which is an idea or mental picture of a group or class of objects formed by combining all their aspects. To measure a student's understanding of a concept we can ask them to identify the correct definition, the distinguishing characteristics, or examples of the concept.

OSI formats can also be designed to understand a principle or procedure.

Which of the following best defines the educational term assessment?

A. A student's test score

*B. A judgment based on a variety of valid information**

C. An evaluation of the student's mental, physical, and social conditions.

A is wrong because but many misuse this term. B is correct. C is too inclusive

Which of the following is an axiomatic principle?

*A. Longer tests tend to yield more higher test scores than shorter tests.**

B. A student test score is likely to be more accurate if the item difficulty matches the achievement level of the student.

C. The chances of correctly answer five CMC items correctly via random guess is very small.

A is correct because it is absolute. B and C are probabilistic therefore not axiomatic.

Which of the following influences the warming of the earth?

Mark A if true and B if false

1. The earth is closer to the sun.

2. Burning fossil fuels

3. Nuclear energy

4. Agriculture

5. Solar energy

6. Hydroelectric energy

1.4.3. *Application of knowledge and skills*

This category of mental complexity is most needed in modern education, because it requires students to use knowledge and skills in coordinated and complex ways. The most common examples are seen in testlets. In fact, the testlet designed to measure the application of knowledge and skills in complex ways. However, the truest form of application comes with a performance test where a checklist or rating scale is used and human judgment determines how well the student performs. Economies are gained by using OSIs for test items that measure application. Some examples of application testlet items are presented here is abbreviated form:

1. Reading. The student reads a passage and responds to three to 12 items probing various aspects of reading comprehension.
2. Mathematics. The often-used story problem initiates a testlet. As with reading, OSIs are used in a coordinated set.
3. History. A passage from a textbook is presented for student analysis. OSIs are presented as a set probe the students' ability to combine knowledge and skills to draw a conclusion, evaluate the merits of a decision, or extract a defensible analysis of the event.
4. Science. An experiment or a vignette introduces something the student was supposed to learn. The vignette might contain data, a chart, a graph, or a report. The OSIs probe who well the student understands and applies knowledge and skills.

1.5. Guidelines for Creating OSIs

Please explain the method, sample or study group, data collection tools, data collection process, and data analysis procedures in this section. This section should indicate the study's design, the sampling, the data collection tools, and the data analysis. Clarification is essential in this part.

In this section, some guidelines are highlighted to guide in the creating or evaluating OSIs. The basis for this section is a popular taxonomy has been published long ago and updated (Haladyna & Rodriguez, 2013). A list of guidelines appears on the internet and is widely shared and used. As a service to readers, poorly written items will be illustrated here as instruction for what not to do. These are really bad items.

Opinion Items. Which country offers the best kebabs? It might be factual, but it looks like an opinion.

Trick Items. In what country, do Panama hats originate? The correct answer is Chile. If one option is Panama, the student is tempted to choose that option.

Format Items Vertically, not Horizontally.

The Ankara Central Station represents which school of architecture?

A. Classical

B. Ottoman

*C. Modernism**

This is the clearest presentation a CMC item. However, in the interest of saving space, some test designers like to place option in the same line.

The Ankara Central Stations represents which school of architecture?

*A. Classical   B. Ottoman C. Modernism**

This horizontal formatting may be confusing to some students.

Edit and Proof Items. All items should be grammatically correct and proofed. Common errors in sentence construction should be avoided. If an item is not well edited and proofed, it leaves a bad impression with the student. Also, lacking editing, the syntax of the item might be clumsy and by that confuse the student.

Linguistic Complexity, Window Dressing, Length. The reading level of any test items should be suitable for the reading level of the class. For those whose first language is different than the language used in a test, the linguistic complexity of an item stem might challenge the student unfairly. I am reminded of a licensing test for police where item stems were very long and linguistically very complex. Much of the information in the stem was irrelevant (window dressing). These factors led to very low performance on the licensing test. Remember that each item has a scoring weight of one. We should attempt to make each item as brief as possible yet retain the content and mental complexity needed.

Avoid Negation in the Stem and the Options.

Which is not true of cardiopulmonary resuscitation (CPR)?

*A. Closed chest massage is as effective as open chest message.**

B. The success rate for out-of-hospital resuscitation may be as high as 30% to 60%.

C. The most common cause of sudden death is ischemic heart disease.

Put the main idea in the stem of the item, not the options. The stem usually has more words than the options. However, one item-writing fault is the unfocused stem.

Agriculture

A. is an important part of the Turkish economy.

B. is an important part of the Turkish economy.

C. shows a decline in avocado production in Turkey.

For this kind of item, options may wander all over the place and even might not be grammatically equivalent.

All choices should be plausible? In writing or evaluating distractors, as the content expert, you are best suited to decide if a distractor is plausible. If it is not plausible, even a student who has not learned will eliminate that distractor and improve the chance of a lucky correct guess. Another way to find out if a distractor is implausible is to ask your students!

Avoid options such as none-of-the-above, all-of-the-above, and I-don't know. Such options offer clues for the clever student.

Longest option is correct. The weary item writer may write question and a long correct answer and then make the other choices due to lack of effort. The longer choice is the correct one.

Avoid absolute words. In the choices, certain absolute words are seldom correct. These absolutes include absolutely, always, completely, outright, never, perfect, without exception and ultimate. Again, clever students will avoid choices with extreme.

Repeating a word or phrase in the both the stem and one choice. This is a clue that the repeating word or phrase is correct. If it is not the right answer, then we have a trick.

What is are Mediterranean avocados principally grown in Turkey?

A. Mediterranean coastal region

B. Southern region

C. Northern region

Pairing terms that presents a clue.

Which condiments are best on kebabs?

A. Salt and pepper

B. Sugar and spice

C. Salt and spice

D. Spice and pepper

If a student does not know the answer, the choices may offer a clue. Spice appears three times. Salt and pepper twice. Sugar once.

Ridiculous Choices. In a hurry to find a third or fourth option, you might insert a choice that no student will choose.

In growing avocados, what is the most important factor?

A. Adequate water

B. An ideal climate

C. Good luck

D. A green thumb

An item like this one has essentially two plausible options.

Format Options in Numerical Order and Observe Place Value. In a test, most students can be very anxious and feel stress, putting numbers of numerical order with clear place value helps the student.

What is the speed of sound?	*What is the speed of sound in kilometers per hour?*
A. 120 km/h	*A. 120*
B. 1200 km/h	*B. 400*
C. 400 km/h	*C. 700*
D. 700 km/h	*D. 1200*

1.6. Creating a Collection of Items for Future Testing

This activity is difficult and time-consuming. Honestly, it takes years to develop a useful collection. This collection will also be subject to review: keep, revise, discard.

We have at least three ways to create a useful collection: (1) Free available items, (2) cloned items, (3) creating you own items. All three methods have advantages and disadvantages.

1.6.1. *Free items*

Depending upon the subject matter and students taught, the worldwide WEB provides many sources of free items. These items are open source. You can obtain such items easily and incorporate them into your item collection judiciously. Each item MUST represent suitable content and have a desirable mental complexity that is appropriate for your students. The problem is that such items lack the close connection with actual instruction. Nonetheless, the price is right. If you can obtain some items at no cost, that might help you develop new, similar items, as suggested in the next strategy.

1.6.2. *Cloned items*

If you find items copyrighted, one strategy is to take the general form of the item and create a model. This is briefly illustrated with an item obtained from the Worldwide WEB.

Painting a wall that measure 15 square meters. One pint of paint covers 7 to 9 square meters. A pint of paint costs 100 Lira.

> *How much paint should I buy?*
>
> *What will it cost?*
>
> *If I have to use two coats, how much paint should I buy?*
>
> *If I have to use two coats, how much will it cost?*
>
> *A painter charges _____ per hour. She estimates the job to take _____ hours.*

The above example is actually an outline for a testlet. It shows that with an item that contains area and cost for a product, many useful items can be generated. Automated Item Generation (Gierl & Haladyna, 2013) has many examples of item models that will produce many items. The limitation is that the items may measure a narrow band of content that is taught.

1.6.3. *Item shells*

Long ago, when helping pharmacists write useful test items for their national pharmacy licensing test, we came upon an idea that still works today (Haladyna & Shindoll, 1989). The approach we found useful is to identify items that had the same syntactic structure and create a shell of the item. The shell consisted on the stem followed by a blank where the content was inserted. Here are some examples of item shells (Haladyna & Rodriguez, 2013, p. 145). These are very generic.

> *Which is the best definition_____? Which is an example of _____? What is the meaning of _____? What is like _____? What are the distinguishing characteristics of _____?*
>
> *Which is the principle of _____? What is the cause/reason for _____? What is the relationship between _____ and _____? Which is an example of the application of this principle ______? What would happen if _____? Which is better/worse, higher/lower, nearer/farther, heavier/lighter, _____? What is the difference/similarity between _____ and _____? Which principle best applies _____? What is the best way to ____?*

One problem with item shells is that items generated from shells get to be repetitious. So, the use of any specific item shell should be limited. Nonetheless, the item shell gets item writers started if they have “writers’ block.” Clearly, it speeds up the item-writing process.

1.6.4. *Creating items*

The old-fashion way to create items is simply to select which format to use and write the item. Teacher/instructor-made test items are notoriously bad item writers. This tendency is true because most teachers/instructors do not have adequate training or have not been exposed to the formats, guidelines, and techniques found in this article and in the references are the end of this article.

Writing your own items is tedious and time-consuming. As pointed out previously, we often refer to your collection of items as an item bank. So, writing and placing items in your bank yields benefits in the future, just like a savings account in a bank.

1.7. Evaluating Items

Once items are created for measuring student learning in a classroom or course in a university or professional school, evaluating items is challenging. For large-scale testing programs, we have very sophisticated methods for evaluating test items (Haladyna, 2015; Haladyna & Rodriguez, 2021). These methods are inappropriate for student testing in the classroom.

In the classroom or in a course of study, how students respond to items is the best way to evaluate each item. A review of any summative test should reveal if items are working as intended. High-achieving students should choose correctly, and low-achieving students should choose incorrectly. If all students choose correctly, teaching has been effective and student learning has also been effective. If an item has a low degree of correct choice (less than 50% for a CMC item), we have a problem. Here are some questions that should help you evaluate whether your students are being given fair treatment in measuring what they have learned.

1. Is the item irrelevant regarding content?

2. Is the item flawed? Review the guidelines for writing items.

3. Does the item have two correct choices? This can happen.

4. Does the item have no correct choices? This can happen.

5. Was the content taught? Testing students on content not taught is not fair.

6. Did most of students dismiss what was taught? Students have to accept responsibility for a lack of study.

As we evaluate our test, we also evaluate our teaching. Honest discourse with students following the administration of a summative test, a meeting with students to go over test results reveals answers to the many questions just posed. Also, a chance for students to discuss what they learned and have not learned can be a valuable learning experience. It also helps you (the teacher/instructor) improve the quality of your collection of test items for future use.

1.8. Closing

The advice offered in this article is intended to guide you and your students toward a positive experience when it is time to measure what students have learned and help them continue on the path to future learning. Having a collection of useful test items is a start. Using these items in formative and summative ways is important as we guide each student to a successful end of their brief educational experience.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Thomas Haladyna https://orcid.org/0000-0003-3761-6979

REFERENCES

Gierl, M., & Haladyna, T.M. (Eds.). (2013). *Automatic item generation: Theory and practice.* Routledge.

Haladyna, T.M. (Submitted for publication). *How much of threat to validity is random guessing?*

Haladyna, T.M. (2015). Item analysis for selected-response test items. In Lane, S., Raymond, M.R., & Haladyna, T.M. (Eds.). *Handbook of test development* (pp. 392-409). Routledge.

Haladyna, T.M., & Rodriguez, M.R. (2021). Using full-information item analysis to evaluate multiple-choice distractors. *Educational Assessment, 26*(3), 198-211. https://doi.org/10.1080/10627197.2021.1946390

Haladyna, T.M., & Shindoll, L.R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation in the Health Professions, 12*(1) 97-106. https://doi.org/10.1177/016327878901200106

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items.* Routledge.

Haladyna, T.M., Raymond, M.R., & Stevens, C. (2019). Are multiple-choice items too fat? *Educational Assessment, 32*(4), 350-364. https://doi.org/10.1080/08957347.2019.1660348

Lane, S., Raymond, M., & Haladyna, T.M. (Eds.) (2015). *Handbook of test development* (2nd ed.). Routledge.

Rodriguez, M.C. (2016). Selected-response item development. In Lane, S., Raymond, M.R., & Haladyna, T.M. (Eds.). *Handbook of test development* (pp. 259-273). Routledge.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 19–27

https://doi.org/10.21449/ijate.1127958

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of the article.

Classroom Assessment: The Psychological and Theoretical Foundations of the Formative Assessment

Seval Kula-Kartal[1,*]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Türkiye

ARTICLE HISTORY

Received: June 08, 2022
Revised: Aug. 19, 2022
Accepted: Aug. 28, 2022

Keywords:
Formative assessment,
Classroom assessment,
Educational measurement.

Abstract: The change of the learning and teaching definitions in psychology has also changed the nature of classroom assessment implementations. One of these important changes is that the targeted skills in the classroom assessment, and item structures utilized to measure these skills have changed. Teachers have started to use items and tasks that can represent the significant knowledge and skills of subject better and provide rich information and evidence for students' development rather than utilizing items measuring only remembering the bits of knowledge or simple comprehension. The second important change has taken place in terms of why and how teachers and students use classroom assessment. Although there are a few studies on the change of targeted skills and item structures used in the classroom assessment in the related national literature, the number of studies providing information on how to apply classroom assessment in line with its recent definitions, aims and implementation is still limited. Thus, the current study aims at providing guiding information on why and how to implement classroom assessment to develop students' learning, motivation, and self-assessment skills.

1. INTRODUCTION

The studies in the field of educational sciences widely aim at revealing the classroom implications that will contribute to the students' academic achievement. The main target of the studies in this field is to provide findings for specifying the necessary developments and applications to increase students' success in all elements of the educational system. In addition, the studies in the educational sciences are closely related to psychology, and they follow the contemporary perspectives suggested in psychology. Similarly, the studies carried out in the field of educational measurement and assessment have progressed aligning with the recent developments in psychology.

1.1. How does the Psychology Affect the Classroom Assessment?

The psychological approach that has dominated both psychology and education for a long time is behaviorism. In this approach, it is accepted that learning occurs by acquiring bits of

*Corresponding Author: Seval Kula-Kartal ✉ kulasevaal@gmail.com Pamukkale University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

knowledge. The center of teaching is to teach the bits of information within a sequential and hierarchical process. Therefore, the focus of classroom assessment is to measure and evaluate to what extent students can remember the concepts and definitions of a specific subject matter rather than assessing if students can transfer their knowledge to new problems and situations. Moreover, the source of students' motivation is attributed to external factors, and it is agreed that students' motivation depends on reinforcing their small steps. In accordance with this approach to student motivation, the whole information obtained from classroom assessment is mostly used to enable students to get motivated externally instead of using the classroom assessment to contribute to students' learning and motivation. In addition, classroom assessment is regarded as a separate activity following the completion of a teaching and learning process rather than an activity integrated with the whole teaching and learning process (Shepard, 2000).

Since the 1960s, together with the development of cognitive and constructivist approaches in psychology, the definition of learning has extended by including implementing knowledge to new situations, problem-solving and other complex and higher order learning outcomes. The point of view underlying the learning has transformed to that student makes meaning of new knowledge and connects prior knowledge to new knowledge and experiences (Brookhart, 2020). The student has been accepted as a person who not only reacts to the stimulus coming from his/her environment, but also reorganizes the new information to fit it into the pre-existing cognitive schema and develop new schemas (Senemoğlu, 2015). In accordance with the new definition of learning, the approaches to students' role in the learning process have also changed. The students have been accepted as active participants and regulators of their own learning process. This transformation became clear with that the regulation has gained importance in most contemporary learning approaches (Bandura, 1977, 1993; Zimmerman, 2000; Zimmerman & Schunk, 2001).

The change of the learning and teaching definitions in psychology has also changed the nature of classroom assessment implementations. One of these important changes is that the targeted skills in the classroom assessment and item structures utilized to measure these skills have changed. Teachers have started to use items and tasks that can represent the significant knowledge and skills of the subject better and provide rich information and evidence for students' development rather than utilizing items measuring only remembering the bits of knowledge or simple comprehension. The second important change has occurred related to why and how teachers and students use classroom assessment (Brookhart, 2020; McMillan, 2020; Shepard & Penuel 2018; Shepard, 2000). Although there are a few studies on the change of targeted skills and item structures used in the classroom assessment in the related national literature (Berberoğlu, 2006; Kutlu et al., 2014), the number of studies providing information on how to apply classroom assessment in line with its recent definitions, aims and implementation is still limited. Thus, the current study aims at providing guiding information on why and how to implement classroom assessment to develop students' learning, motivation, and self-assessment skills.

1.2. How does the Cognitive Approach Define the Classroom Assessment?

According to the behavioristic approach to educational measurement, the most important aim of the classroom assessment is to measure and evaluate students' degree of learning at the end of a specific teaching unit or process. However, the new perspectives on learning and teaching processes in cognitive approach define the primary aim of the classroom assessment as contributing to the development of students' learning, motivation, and self-regulation. Under this contemporary point of view, the one of the most comprehensive definitions of classroom assessment has been made by McMillan (2013). In this definition, the classroom assessment is regarded as a process in which teachers and students attain, use, and evaluate the information and evidence of students' learning for various purposes. These purposes include specifying students' strengths and weaknesses, monitoring their learning processes and providing feedback for their development and grading. The classroom assessment is a tool utilized by teachers to attain necessary information to make inferences supported by the evidence regarding what students can know, understand, and do. This tool can also develop students' learning and motivation if it can be used effectively.

There are two important points emphasized in the classroom assessment definition made by McMillan (2013). One of these points is that it becomes possible to make inferences regarding students' learning based on the information and evidence attained from the classroom assessment implementations. Another point is to utilize the information and evidence obtained from assessments to develop students learning and motivation. Accordingly, the fundamental aim of the classroom assessment is to use the information to adjust students' learning. The term "formative assessment" emphasizes this primary aim of the classroom assessment. A more recent term, which is "assessment for learning", is another concept used to put an emphasis on this aim of the classroom assessment. In the current study, formative assessment (FA) is preferred to emphasize the formative purposes of classroom assessment.

1.3. What is the Formative Assessment?

One of the common points emphasized in the FA definitions is that the results obtained from the assessments with formative purposes provide evidences informing the decisions teachers and students must make during the teaching and learning processes (Black & William; 1998; Panadero et al., 2018). Another significant feature of the FA is that its general goal is to develop students' learning. As stated by Wilson (2016), the primary function of the FA is to collect detailed information that can be used to improve teaching and student learning. The FA not only provides the necessary information and evidence regarding student learning but also improves instruction and learning by enabling students and teachers to decide what to do in the next steps based on that information (McMillan, 2020; Panadero et al., 2018; William, 2010).

In order to achieve the two goals mentioned, which provide information and adjusting teaching and learning, the FA must be integrated with the whole teaching and learning process. This is another feature of the FA commonly emphasized by the researchers (Brookhart & Helena, 2003; Nitko & Brookhart, 2014). Education researchers state that the actions taken by the teachers and students to develop student learning are at the center of the FA. The classroom assessment with formative purposes requires teachers and students to take actions based on the information provided by the assessments (Ferrara et al., 2020). Therefore, teachers and students should reach the necessary information and evidence in time to take an action with the students who generated the results (Chappuis, 2009; Chappuis et al., 2013). To call assessment as formative, the results provided by the assessments should be able to inform decisions made by students and teachers (Moss & Brookhart, 2009, 2015). Lastly, the FA should be a well-planned process so that it can develop student learning by providing momentary and daily information regarding the teaching and learning process.

To sum up, there are four keys commonly emphasized in the FA definitions: it provides information and evidence for student learning, its primary aim is to develop student learning, it proceeds in an integrated way with the teaching process, and it is planned in detail before the instruction starts. A comprehensive definition involving the four keys has been made by Popham (2011). According to that definition, the FA is a planned process in which the information and evidence obtained from assessments are used by teachers and students to improve instruction and learning.

The comprehensive definition of the FA also clarifies some misconceptions regarding classroom implementations of the FA. For example, the FA is not a unit test or a type of test, that is conducted on students to measure their learnings at the end of a specific unit. Instead, it is a planned process that proceeds in an integrated way with the teaching. Therefore, it includes the assessments conducted on students while they are still in the process of learning in the related unit. To call an assessment as formative, the information and evidence provided by the assessment should be used to develop current students' learning or instruction given to the current students. If students do not have any chances to use the information attained from the assessment, or that information is only used to adjust instruction for the future students, then, these assessments cannot be considered as parts of the FA process. In addition, the FA comprises all assessments providing information that can be primarily used to develop student learning. Therefore, it differs from the assessments that are primarily used for accountability, ranking or assessment of learning purposes (Ferrara, 2020; Moss & Brookhart, 2015; Popham, 2011). The FA definition also clarifies its characteristics and keys that should be considered when implementing the FA in the classroom.

1.4. What are the Characteristics of the Formative Assessment?

There are six important components of the FA process enabling it to develop student learning and instruction when implemented effectively: 1) defining the learning outcomes, learning progression and performance criteria clearly, 2) sharing the outcomes and performance criteria with students in a student-friendly language and by using samples, 3) attaining information and evidence showing students' current state of learning, 4) giving formative feedback to students based on the information and evidences, 5) students' self-assessment of their own learnings based on the performance criteria and information attained from the assessments, 6) creating a classroom assessment climate enabling assessments to improve learning (Chappuis et al., 2013; Moss & Brookhart, 2009; Panadero et al., 2018). Those components are the most effective characteristics of the FA in developing students learning, motivation and self-regulation.

Defining performance criteria and sharing them with the students: The first two components of the FA is the basis of planning and implementing the FA. Researchers consider that the FA is especially effective in developing and monitoring the skills that take a long time to develop and needed by the students during their whole life (Popham, 2011). Therefore, teachers should specify those kinds of information and skills and related sub-skills at the beginning of the FA process. In addition, they should also define the possible learning progressions followed by the students when they acquire those skills and success criteria that will be used to determine if students acquired the targeted skills. Sharing plans and definitions with the students in a student-friendly language have the equal importance with the planning. Researchers suggest teachers to use tasks embodying the learning outcomes to help students discover and develop conceptions of the learning outcomes and success criteria (Moss & Brookhart, 2009). It is necessary to share good and weak examples of work with students and make a discussion with them on the features that make those works good or weak. The examples of work and discussions on those examples enable students to transform learning outcomes from abstract outcomes to more concrete success criteria to be met to accomplish the task.

Attaining information and evidence for student learning: Defining the targeted learning outcomes, learning progression and success criteria also reveal at which points of the learning progression teachers should attain information to monitor student learning. The students should work on the tasks embodying the learning outcomes first with the guidance of teacher, and then independently. Based on the information and evidence attained from assessments, it is necessary to specify the strengths and weaknesses of students' performance by comparing students' performance with the success criteria (Moss & Brookhart, 2015; Popham 2011). The tasks used in the assessments during the learning progression should be able to provide rich information for the development of students' learning and the success criteria of the task should be defined clearly (Brookhart & Helena, 2003; Shepard & Penuel, 2018).

Formative feedback: Teachers should give feedback to the students so that the results obtained from the assessment can be used to develop students' learning. The findings of the studies reveal that the quality of feedback matters, and all kinds of feedback do not develop student learning (Black & William, 1998; Kluger & DeNisi, 1996; Shepard, 2020). The related studies revealed that the feedback that is task dependent, compares student performance with the success criteria of the task, not only informs the strengths and weaknesses of the performance but also includes some suggestions for the next steps that should be taken to develop performance, helps students develop their learning (Brookhart, 2008, 2020; Sadler, 1989). Thus, teachers should give feedback to the students in which they define what to develop in student learning and suggest some strategies and methods that can be used by the students to develop their learning. In fact, it can be stated that the key point enabling the FA process to develop student learning is the formative feedback given by the teachers to the students.

Self-assessment: In the FA, it is very important that students are in an active role, and they monitor and take the responsibility for their own learning process (Popham, 2011). It requires students to use their self-assessment and goal setting skills to use the information and evidence for their learning obtained from the assessments (Chappuis et al., 2013). Self-assessment is defined as a student-centered activity in which students evaluate their own performance on the assessment task (McMillan, 2020). Self-assessment includes the three steps that should be taken by the students. Firstly, students should clearly understand the success criteria of the task to be able to evaluate their own learning or performance. This puts an emphasis on sharing the targeted learning outcomes and success criteria with the students in a clear language one more time. Secondly, students should monitor their own performance and specify inadequacies of their performance by comparing their performance with the success criteria of the task. Lastly, they should set related, short-date and clear goals for themselves by defining the future steps to overcome inadequacies based on the evaluations they made in the second step (Brookhart & Helena, 2003; Wylie & Lyon, 2020).

Classroom assessment climate: The FA requires teachers to embrace an appropriate classroom assessment approach and create a classroom assessment climate so that the FA process can develop students learning, motivation and self-regulation. The social and emotional dimensions of the classroom are closely related to how students are going to use the information and feedback provided by the assessments and teacher. Students are more willing to monitor and evaluate their own learning in a classroom characterized by interpersonal trust and in which mistakes are accepted as natural components of the learning process (Chappuis et al., 2013; Leighton, 2020; Shepard & Penuel, 2018). When teachers can create this assessment climate in their classrooms, students are not punished or rewarded for their wrong or correct answers. On the contrary, they are inside of a learning process during which their strengths and weaknesses are revealed thanks to continuous and personalized feedback. They are allowed to reach deeper and sophisticated learning by using feedback. Within this assessment climate, it is possible for students to focus on mastering targeted skills without feeling any anxiety to perform better than the others or being punished because of mistakes (Kutlu & Kula-Kartal, 2018).

The mentioned characteristics of the FA enable teachers and students to answer the three questions they ask themselves momentarily, daily, weekly, and monthly in the classroom: "Where am I going?" "Where am I now?" "What should I do to close the gap between my current and targeted status?" (Chappuis et al., 2013; Moss & Brookhart, 2009; Sadler, 1989). Defining the learning outcomes, success criteria and sharing them with the students make the learning targets more concrete and clearer both for teachers and students. The results attained from formative assessments provide information for the current levels of student learning. The formative feedback, self-assessment and goal setting enables both teachers and students to specify the necessary future steps to develop current performance. Thanks to those characteristics, the FA both provide answers to the three critical questions which adjust students' learning process and develop students' learning, motivation, and self-regulation.

1.5. Why does the Formative Assessment Develop Learning, Motivation and Self-Regulation?

The FA, when it is implemented effectively, increases students' academic achievement by changing their interactions with classroom assessments (Black & William, 1998; Chappuis et al., 2013). Students take the feedback given to them by the teacher more seriously and concentrate on assessment tasks more effectively when they perceive that the assessment is related to and consistent with learning outcomes, learning progression and success criteria. In addition, the FA enables students to develop their performances by providing them continuous feedback regarding the inadequacies of the performance that need to be developed and future steps should be taken to develop them (Brown et al., 2009; McMillan, 2018).

Students can evaluate their own competencies realistically thanks to the clearly defined and shared success criteria, teacher's formative feedback and their self-assessments. When students set goals consistent with the learning outcomes and their competencies, they can believe that they can accomplish those goals and trust their competencies more. If students can understand where they are going, in other words what the targeted learning outcomes are, their possibility of believing that they are going to accomplish those goals also increases. That increased self-efficacy enables them to put more efforts into the tasks (McMillan, 2020). In addition, students have a clearer picture of the future steps that should be taken to develop learning thanks to assessments, feedback, and self-assessment. This information enables them to perceive the development under their control and to get motivated to take the necessary steps to develop learning (Brookhart & Moss, 2015).

The FA increases students' self-awareness of their thinking and enables them to use this awareness to adjust their own thinking processes. During the FA process, students evaluate their works by using the success criteria. They can define the inadequacies in their works when they have a clear picture of what a superior work looks like. This gives the responsibility back to the students and enables them to self-regulate their learning. In addition, teacher's formative feedback develops students' self-assessment skills by guiding them on how to evaluate their performances by using the success criteria (Moss & Brookhart, 2009; Panedero et al., 2018; Shepard & Penuel, 2018).

1.6. How is the Formative Assessment Implemented in the Classroom?

The implementation of the FA in the classroom is guided by the three questions mentioned in the previous sections. The question that should be addressed at the beginning of the FA process is "Where am I going?". At the beginning of the process, teachers must decide which skills and sub-skills they are going to teach, which learning progression students may follow while they are acquiring those skills, when and how they are going to attain the information and evidence showing students' current levels of learning and with which success criteria they are going to compare students' performance. According to researchers, these plans are necessary to define the process called learning progression (Popham, 2011). For example, if the targeted learning outcome is writing a compare-contrast essay, then, teachers should define the related skills to this outcome and plan in what order and how they are going to teach these skills. To do this, teachers must consider and note down the features and qualities of a superior compare-contrast essay. These qualities make clear both the related skills that should be developed in students and the success criteria that will be used to determine students' current levels in terms of those skills. The teacher may specify the qualities such as comparing and contrasting the given topic, situation or entities, supporting the thinking with appropriate samples and evidence, organizing an essay including an introduction, body and conclusion parts.

After planning the FA process, teachers must share this plan with students by using work examples embodying the learning outcomes and success criteria. Teachers should use good or weak examples of compare-contrast essays to provide students with a clearer picture of what a good compare-contrast essay looks like. Teachers and students should have a discussion on what features make the essay a good or a weak example of compare-contrast essay, and teachers should help students to discover those features by themselves by using strategical questioning. Thus, teachers can model their students how to evaluate their essays based on the defined success criteria. After modelling, students can examine and evaluate a good and a weak compare-contrast essay example by using the success criteria. When students complete their evaluations, teachers can start a discussion in which they attract students' attention to the features of the sample essays mentioned by the students when they are sharing their evaluations with the teacher. This step is crucial both for students and teachers to give a clear and concrete answer to the question of "where am I going?". Using good and weak examples of compare-contrast essay and evaluating them by comparing them with the features of a good essay (success criteria) help students to have a clear picture of what a good compare-contrast essay looks like.

It is important for teachers and students to answer the question of "where am I now?" in the learning process. To answer that question, teachers should assess students' current writing skills. The teacher can use a performance task in which students can write an essay based on their prior knowledge and experiences without making any research. For example, students may be asked to compare and contrast living in an apartment with living in a house with a garden. Students should assess their essays by comparing them with the performance criteria of the task. Teachers must also provide feedback including the information for the weak parts of the essay that need be improved and suggestions how to improve those parts. Students should have a second chance to work on their essays again to be able to use self-assessment and teacher's feedback to develop their essays. If teachers find it necessary, the same process should repeat with a different performance task. When teachers decide that students are competent enough in writing compare-contrast essay, they can implement the main performance task in which they can ask students to write a compare-contrast essay on a topic that will require them to make some research.

The FA process requires teachers and students to answer the question "what should we do to close the gap between the current and targeted learning?". At this point, the information attained from assessments guides teachers and students on what to do to develop the writing skills. Teachers and students should specify the weak parts of their essays by using the performance criteria. For example, if majority of the students mention only differences in their essays or they are not competent in writing the topic sentence yet, then, the teacher should adjust the instruction to address those deficiencies. Students also review and edit their essays based on their self-assessment and teacher's feedback. Similar to that example, The FA is an iterative process including goal setting, attaining information and evidence, self-assessment and formative feedback, and this process proceeds until students become competent in the targeted skills.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Seval Kula-Kartal https://orcid.org/0000-0002-3018-6972

REFERENCES

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change *Psychological Review, 84*(2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117-148. https://doi.org/10.1207/s15326985ep2802_3

Berberoğlu, G. (2006). Sınıf içi ölçme değerlendirme teknikleri. Morpa Kültür Yayınları.

Black, P., & William, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Brookhart, S. M (2020). *Feedback and measurement.* S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 63-78). Taylor & Francis.

Brookhart, S.M. (2008). *How to give effective feedback to your students.* ASCD.

Brookhart, S.M. & Helena, M. T. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5-12. https://doi.org/10.1111/j.1745-3992.2003.tb00139.x

Brookhart, S.M., & McMillan, J.H. (2020). *Classroom assessment and educational measurement.* Taylor & Francis.

Brown, G.T.L., Irving, S.E., Peterson, E.R., & Hirschfeld, G.H.F. (2009). Use of interactive informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction, 19*(2), 97-111. https://doi.org/10.1016/j.learninstruc.2008.02.003

Chappuis, J. (2009). *Seven strategies of assessment for learning.* Pearson Education.

Chappuis, J., Stiggins, R., Chappuis, S., & Arter, J.A. (2013). *Classroom assessment for student learning.* Pearson Education.

Ferrara, S., Maxey-Moore, K., & Brookhart, S.M. (2020). *Guidance in the standards for classroom assessment: useful or irrelevant?* S.M. Brookhart & J.H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 97-119). Taylor & Francis.

Kluger, A.N., & DeNisi, A. (1996). The effect of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284. https://doi.org/10.1037/0033-2909.119.2.254

Kutlu, Ö., & Kula-Kartal, S. (2018). The prominent student competences of the 21st century education and the transformation of classroom assessment. *International Journal of Progressive Education, 14*(6), 70-82. https://doi.org/10.29329/ijpe.2018.179.6

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2014). *Ölçme ve değerlendirme: performansa ve portfolyaya dayalı durum belirleme* [*Measurement and evaluation: Performance andportfolio assessment*]. Pegem Akademi.

Leighton, J.P. (2020). *Cognitive diagnosis is not enough: the challenge of measuring learning with classroom assessments*. S.M. Brookhart & J.H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 170-191). Taylor & Francis.

McMillan, J.H. (2013). *Why we need research on classroom assessment?* J.H. McMillan (Ed.). SAGE handbook of research on classroom assessment. (p. 3-16). SAGE.

McMillan, J.H. (2018). *Classroom assessment principles and practice that enhance student learning and motivation*. Pearson Education, Inc.

McMillan, J.H. (2020). *Assessment information in context.* S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 79-94). Taylor & Francis.

Moss, C.M., & Brookhart, S. (2009). *Advancing formative assessment in every classroom.* ASCD.

Moss, C.M., & Brookhart, S. M. (2015). *Formative classroom walkthroughs.* ASCD.

Nitko, A.J., & Brookhart, S.M. (2014). *Educational assessment of students.* Pearson.

Panadero, E., Andrade, H., & Brookhart, S.M. (2018). Fusing self-regulated learning and formative assessment: a roadmap of where we are, how we got here, and where we are going. *The Australian Educational Researcher, 45*, 13-31. https://doi.org/10.1007/s13384-018-0258-y

Popham, W.J. (2011). *Classroom assessment: What teachers need to know.* Pearson.

Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144. https://doi.org/10.1007/BF00117714

Senemoğlu, N. (2015). *Gelişim, öğrenme ve öğretim: kuramdan uygulamaya* [*Development, learning and teaching: from theory to practice]*. Yargı Yayınevi.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14. https://doi.org/10.3102/0013189X029007004

Shepard, L.A. & Penuel, W.R. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice, 37*(1), 21-34. https://doi.org/10.1111/emip.12189

Shepard, L.A. (2020). *Discussion of part II: Should "measurement" have a role in teacher learning about classroom assessment?* S.M. Brookhart & J.H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 192-206). Taylor & Francis.

Wilson, M. (2016). The importance of classroom assessment. *NCME Newsletter, 24*(2), 2-3.

William, D. (2010). *An integrative summary of the research literature and implications for a new theory of formative assessment*. H.L. Andrade & G.J. Cizek (Eds.). Handbook of formative assessment. (p. 18-40). Taylor & Francis.

Wylie, E.C., & Lyon, C.J. (2020). *The role of technology-enhanced self- and peer assessment in formative assessment.* S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 27-45). Taylor & Francis.

Zimmerman, B.J. (2000). Self-efficacy: an essential motive to learn. Contemporary *Educational Psychology, 25,* 82-91. https://doi.org/10.1006/ceps.1999.1016

Zimmerman, B.J., & Schunk, D.H. (2001). *Self-Regulated learning and academic achievement.* Routledge.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 28–45

https://doi.org/10.21449/ijate.1132687

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of the article.

Are different responses related to the different affective features? CHAID analysis study

Neslihan Tugce Ozyeter[1,*]

[1]Kocaeli University, Faculty of Education, Department of Educational Sciences, Kocaeli Türkiye

ARTICLE HISTORY

Received: June 18, 2022

Revised: Aug. 13, 2022

Accepted: Aug. 17, 2022

Keywords:
CHAID analysis,
Rubric,
Affective features,
Student answer.

Abstract: In education, examining students' learning in detail, determining their strengths and weaknesses and giving effective feedback have gained importance over time. The aim of this study is to determine the distribution of students' answers to the reading comprehension achievement test items which were written at different cognitive levels and to investigate the affective variables that are effective in classifying students based on their incorrect, blank, and unrelated answers identified via rubric. For this purpose, a reading comprehension achievement test, a student information form, the perceived academic self-efficacy scale and the learned helplessness tendency scale were used to collect data. The student information form included perseverance, achievement motivation, exposure to bullying and test anxiety subscales. A rubric was used to determine the students' response categories. According to the findings of the study, the rate of blank and incorrect answers increases as the cognitive level of the items become more complex. While the most correct response rates are decreasing, partially-correct answers are increasing relatively. While students' learned helplessness tendencies were effective in classifying their blank and unrelated answers at the most basic reading comprehension level, as the cognitive process became more complex, the affective characteristics classifying the student responses increased in number. It was concluded that these variables are important in improving the students' answers and in leading them to the partially correct and the most correct answer. It can be suggested to create trainings and classroom environments that will equip and improve students' features about these variables.

1. INTRODUCTION

Many discrete test item structures are used to measure various psychological features in education and psychology. The preferred item structure depends on the psychological feature to be measured. The use of multiple-choice items, frequently administered throughout the 20th century, is quite common in national and international high-stakes tests and in-class measurements. Although multiple-choice items offer various advantages, their limitations have been debated today, and there seems to be a consensus on the existence of more valid and reliable methods to measure certain skills.

*Corresponding Author: Neslihan Tugce OZYETER ✉ simsekneslihantugce@gmail.com Kocaeli University, Faculty of Education, Department of Educational Sciences, Kocaeli, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

One major disadvantage of multiple-choice items is that they do not inform any of the education stakeholders about how students transfer what they have learned and how they structure their answers. With the effect of chance success, simply marking the correct answer in multiple-choice items provides very limited information about students' learning and disregards their partial learning. Additionally, the feedback to students with the same score is quite similar and limited particularly in terms of students' individual learning characteristics and shortcomings.

Many weaknesses of multiple-choice items, changes in the features to be measured, and advances in learning and teaching theories have brought about a more in-depth examination and measurement of student performance. Therefore, educators meet open-ended item use. An open-ended item requires students to construct and write their own answers (Badger & Thomas, 1991). Although it requires expertise to prepare an open-ended item, it is much easier than writing a multiple-choice item, which offers a great advantage. The fact that students create their own answers is very informative about their learning progress. Whether and to what degree the student has achieved the targeted outcome measured by the item, and how much s/he is able to answer correctly can be reliably analyzed thanks to the zero-chance success regarding the item. A rubric must be used to carry out this process and to eliminate errors arising from subjective scoring of the item.

A rubric (Kutlu *et al.,* 2017; Popham, 1999) is a tool for scoring the performance of students in general or by dividing the performance into sub-dimensions, in line with certain indicators. The use of the rubric is very important to explain which performance indicator the student will match with how many points, or which performance indicator the student's score on the item corresponds to. Thus, any doubts about the subjective context of scoring open-ended items are eliminated. Effective feedback is the most important benefit that the rubric adds to the teaching and learning process. Whether the student's performance is measured in general or divided into sub-dimensions, it shows the performance level of the student's response to the item. It also provides information on how to perform to reach higher performance levels, which not only increases the validity of the scores, by providing objective scoring of the answers given to the item but also provides feedback to the student about his/her own performance.

Rubrics are of two types: holistic and analytical (Kutlu *et al.,* 2017). Whereas in analytical rubrics, performance is defined by its sub-dimensions with ratings corresponding to the individual's performance for each sub-dimension, the holistic rubric contains an overall assessment of student performance. A holistic rubric prepared for an achievement test consisting of open-ended items includes possible answers that a student can give. Hence, the most correct answer, partially-correct answers, blank answers, incorrect answers, and unrelated answers are the answer categories for the achievement test. These response categories can be summarized as follows:

The most correct answer: It is the answer that accurately and completely describes the construct and scope measured by the item. When creating the rubric, this response category is written first.

Partially correct answer(s): Responses in this category include answers that accurately but incompletely describe the measured construct and scope.

Incorrect answer: An incorrect answer is the answer that is correct in itself but not true to the scope asked by the item.

Blank Answer: It is the absence of any response regarding the scope and construct of interest in the item.

Unrelated answers: These are the answers that are not related to the scope or construct measured by the item or reflect the cases where the student's writing is illegible.

The response categories described above reveal student performance. The high performance of students in an achievement test is related to their answers being close to the most correct answer. In other words, the first step to be taken to have more successful students is to teach students who give blank, incorrect and unrelated answers in a way that they will give partially correct and the most correct answers. To increase student performance in this way, students should be informed about their performance at the first place. Later, the performance indicators at the target performance level should be examined, and the thinking and learning strategies of the students should be reviewed in light of the feedback.

The answer to the question of why some of the students who receive education in the same class under similar conditions can give the most correct answer, while others leave the item unanswered or give wrong answers is thought to be related to the students' affective features. Reviewing the related research literature, it is clear that self-efficacy comes first among the affective characteristics associated with the academic performance of students (Manzano-Sanchez *et al.,* 2018; Nasir & Iqbal, 2017; Olivier *et al.,* 2019). High self-efficacy, defined as the individual's belief in his or her own capacity (Bandura, 1982), is a factor that increases student success, while low self-efficacy means that students have low self-esteem and low performance. Learned helplessness is a variable that affects both the academic success and emotional wellbeing of the student. Learned helplessness occurs when the individual cannot achieve the expected result despite her/his repeated efforts and weakens the relationship between her/his behavior and the result expected. This situation results in the individual not doing what s/he needs to do to achieve his/her goal. The literature states that students with learned helplessness have low school achievement (Ghasemi, 2021; Walling & Martinek, 1995). Perseverance, on the other hand, is the continuation of the goal-oriented behavior of the individual despite the obstacles (Dweck, 1986). It is closely related to motivation, and both perseverance and motivation are variables that affect student success. Test anxiety is another factor that affects an individual's performance. While low-level anxiety increases the performance of the individual (Parvez & Shakir, 2014), the increase in anxiety makes it difficult for the individual to start work and leads to cognitive and emotional harm (Zahrakar, 2008). Test anxiety also describes the individual's fear of mental failure (Hembree, 1988). This anxiety, which arises when the student takes the test or being evaluated, affects his/her performance. Another variable within the scope of this study that is thought to affect the academic performance of students is exposure to bullying. The child who is bullied by her/his friends at school or in the educational environment suffers from some emotional consequences, and thus her/his academic performance becomes poorer. Studies in the literature confirm the negative relationship between exposure to bullying and academic performance (Roman & Morilla, 2011; van der Werf, 2014).

One of the prerequisites for increasing student performance, getting higher-quality students' answers, and enabling students to perform better in national and international assessments is to identify students who give incorrect, blank and unrelated answers. By doing so, it is thought that necessary measures can be taken to ensure that these students are paid attention to give partially correct and the most correct answers. Identifying the affective characteristics that can be used in classifying student responses is important, especially in determining the characteristics of students who give incorrect, blank, and unrelated answers, and in taking precautions for this student group. Therefore, with this study, it is aimed to determine the affective characteristics that are effective in classifying the distribution of the answers given by the students to the items written at different cognitive levels in the reading comprehension achievement test.

1.2. Research Questions

Questions to be answered within the context of this study are as follows:

1- What is the distribution of the answers, given by the fifth-grade students, to the open-ended items written at different cognitive levels based on the response categories?

2- What role do achievement motivation, perseverance, test anxiety, perceived academic self-efficacy, exposure to bullying, and learned helplessness of fifth-grade students play in classifying their responses to open ended items at different cognitive levels in the reading comprehension test?

3- How do the students' affective characteristics, which play a role in classifying the response categories of the answers given to the reading comprehension test, differ according to the cognitive level of the item?

2. METHOD

2.1. Research Design

This research is designed as a correlational study, which aims to reveal the relationships between students' affective features and their responses to the reading comprehension achievement test. To decide which variables are discriminators, the relationships between independent variables and the dependent variables were examined.

2.2. Study Group

The Study group consisted of 944 fifth grade students from Ankara and Kocaeli provinces in Turkey. The students were chosen from different districts of the cities in order to minimize the effects of socioeconomic variables. The gender and the location distribution of the study group is presented in Table 1.

Table 1. *Gender and the Location Distribution of Study Group.*

	f	%
Gender		
Female	436	46.2
Male	508	53.8
Province		
Ankara	313	33.2
Kocaeli	631	66.8

Table 1 shows that the gender distribution of the study group is quite even. The percentages of female and male students are close to each other. The study group mainly composed of students from Kocaeli province.

2.3. Data Collection Tools

To collect data, a reading comprehension test, a scoring rubric and student information form were used. All data collections tools were constructed by the researcher. To collect data, ethical permission from Ankara University was approved on 30/03/2020 and the decision number is 64. Additionally, data collection permissions were received from provincial directorates of national education of Ankara and Kocaeli.

2.3.1. *Reading comprehension test*

The reading comprehension achievement test was composed of 4 open-ended items. These items were generated based on the reading comprehension processes suggested by Progress in International Reading Literacy Study (PIRLS). PIRLS defines reading comprehension processes with four cognitive processes. These processes are focusing on and retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information and evaluating and critiquing content and textual elements. These processes are hierarchical which means that they are constructed from the simplest to the most complicated. While focusing on and retrieving explicitly stated information process requires students to use the explicitly stated information as it is in the text, evaluating and critiquing content and textual elements, which is the most complex comprehension process, allows students to benefit from their own experiences and learning and present an evaluation or produce a critique (Mullis *et al*., 2016).

In order to receive an expert opinion for the items developed, PIRLS reading comprehension processes document, the text and the items were sent to an expert group of measurement and evaluation in education and a Turkish language teacher with 5-year experience. Experts were asked to provide feedback regarding the validity of the items, technical features of the items and the instructions while Turkish teacher was requested to provide feedback about the suitability of the text and the items with the age of the students. All the feedback was carefully studied and necessary editing and corrections were made in line with the feedback. The reading comprehension achievement test was finalized.

Upon finalizing the reading comprehension achievement test, it was piloted with a small group which is similar to the target group. This small session was used to predict the necessary time for students to read the text and write their answers. Additionally, students' questions during the session were noted down to be used to have more reliable and valid data collection process.

2.3.2. *Rubric*

Rubric was constructed to objectively score the open-ended items in the reading comprehension achievement test, and to identify the students' response categories. To prepare a valid rubric, the answers of the students collected from the pre-test application provided an insight.

The rubric included response categories that can be used to give feedback to the students. The response categories were the most correct answer, partially correct answers, blank answers, incorrect and unrelated answers. For each item, the most correct answer was written first. Partially correct answers were defined according to their distance to the most correct answer. Blank answers were those in which the student did not write anything. Incorrect answer was the correct answer of another item, while unrelated answer referred to the student's answers unrelated from the text.

2.3.3. *Student information form*

The student information form measured the student's affective and demographic characteristics. The affective characteristics of the student measured within the scope of the study were achievement motivation, test anxiety, perceived academic self-efficacy, exposure to bullying, perseverance and learned helplessness tendency.

2.3.3.1. Achievement Motivation. A 5-item subscale used in the PISA 2015 application was used to measure students' achievement motivation (OECD, 2017). The reliability coefficient calculated for the study group was 0.77. The Confirmatory Factor Analysis (CFA) results for the study group showed that the subscale was validated for the study group (RMSEA=0.035; CFI=0.99; TLI=0.99; SRMR=0.014).

2.3.3.2. Test Anxiety. Test anxiety was measured through the 5-item-subscale used in the PISA 2015 application. The CFA results for the study group constituted the validity evidence of the scale (RMSEA=0.053; CFI=0.99; TLI=0.97; SRMR=0.023). The internal consistency coefficient calculated for the group of this study was 0.74.

2.3.3.3. Perceived Academic Self-Efficacy. Students' academic self-efficacy was measured with the Perceived Academic Self-Efficacy scale adapted by Özyeter and Kutlu

(2022). There were 30 items under 3 dimensions in the scale. The CFA results of the scale for the study group showed that the construct was confirmed for the study group (RMSEA=0.066; CFI=0.91; TLI=0.90; SRMR=0.063). The Cronbach Alpha internal consistency coefficient was 0.68.

2.3.3.4. Exposure to Bullying. Exposure to bullying subscale was used in PISA 2018 application (OECD, 2019). According to the CFA results, the scores obtained from the scale were valid for the study group (RMSEA=0.054; CFI=0.98; TLI=0.96; SRMR=0.033). The Cronbach Alpha reliability coefficient was calculated as 0.81.

2.3.3.5. Perseverance. Another subscale used in the study was perseverance subscale (OECD, 2014). When the CFA results for the study group (RMSEA=0.034; CFI=0.99; TLI=0.99; SRMR=0.017) and the internal consistency coefficient (0.77) were examined, it can be concluded that the subscale produced valid and reliable results.

2.3.3.6. Learned Helplessness Tendency Scale. The learned helplessness tendency scale developed by Kutlu and Özyeter (in press) produced valid (RMSEA=0.037; CFI=0.92; TLI=0.90; SRMR=0.036) and reliable (0.68) results for the study group.

The fact that the sub-scales and scales in the student information form generally had lower reliability than the original forms was thought to be related to their application to a single grade level. Data collected from the fifth grade students may have become more homogeneous in terms of the feature of interest. Therefore, the reliability was lower than the original forms. Still, they are above the acceptable lower limit of 0.60.

2.4. Data Analysis

Before proceeding to the analysis of the data, descriptive statistics of the scores obtained from the subscales and scales measuring affective characteristics were presented in order to explain the situation of the study group in terms of the variables measured in the study.

To answer the first research question of the study, frequencies and percentages were used and graphs were created to examine the distribution of the answers given by the fifth-grade students to the items written at different cognitive levels in the reading comprehension achievement test. CHAID analysis was used to answer the second research question, which is "What role does fifth grade students' achievement motivation, perseverance, test anxiety, perceived academic self-efficacy, exposure to bullying, and learned helplessness play in classifying their responses to items written at different cognitive levels in the reading comprehension achievement test?". CHAID analysis is one of the oldest and best-known tree classification methods developed by Kass in 1980 and uses the chi-square test for categorical dependent variables (Nisbet *et al.,* 2009). CHAID classifies the analyzed data set on the condition that the change in the dependent variable is minimum (homogeneous) within groups and maximum (heterogeneous) between groups, and repeats this process until there is no statistically significant differentiation for the subgroups formed after each node (Kass, 1980). Within the scope of the study, the correct answer category was created by combining the most correct answer and partially correct answers of the students together. The dependent variable of the CHAID analysis is the students' response categories (correct answer, incorrect answer, blank answer and unrelated answer). The independent variables are learned helplessness, perceived academic self-efficacy, achievement motivation, perseverance, test anxiety and exposure to bullying. In order to answer the last research question, the similarities or differences of the affective characteristics, which were effective in classifying the answers given to the items in the reading comprehension achievement test which are written at different cognitive levels, were examined. CHAID analysis demands no assumptions regarding the distribution of the relationships of variables. However, defining the correct scale levels of both dependent and independent variables is of the most importance (IBM, 2012). In order to answer the last research question, the similarities or differences of the affective characteristics, which are effective in classifying the answers given to the items written at different cognitive levels in the reading comprehension achievement test, were examined.

3. FINDINGS

Before answering the research questions, the descriptive statistics of the scales and subscales and the factors of the achievement test used in the study were calculated. These statistics are presented in Table 2.

Table 2. *Descriptive Statistics of Data Collection Tools.*

	$\bar{x}$	*Sd*	Minimum Score	Maximum Score
Reading comprehension achievement test total score	21.82	6.78	0.00	37.00
First item (first cognitive process)	8.12	3.13	0.00	10.00
Second item (second cognitive process)	4.78	2.88	0.00	10.00
Third item (third cognitive process)	5.29	2.83	0.00	10.00
Forth item (forth cognitive process)	3.62	2.11	0.00	10.00
Achievement motivation	16.57	3.04	5.00	20.00
Perseverance	16.08	2.78	6.00	20.00
Text anxiety	12.32	3.66	5.00	20.00
Perceived academic self-efficacy	92.35	11.21	51.00	120.00
Exposure to bullying	9.16	4.03	6.00	24.00
Learned helplessness tendency	4.12	2.43	0.00	13.00

When the descriptive statistics presented in Table 2 are examined, the scores corresponding to the answers given by the students to the achievement test items are observed to be the highest at the level of focusing and retrieving explicitly stated information, and the lowest at the level of examination and evaluation and critiquing the context and the textual elements, which is the most complex level of reading comprehension. The scores decrease as the students proceed through the complex reading processes. When the student affective characteristics are examined in general, it can be seen that the students get the highest scores on the achievement motivation and perseverance subscales.

3.1. Findings Regarding the First Research Question

The first research question sought to be answered is how the students' responses to the items at different cognitive levels in the reading comprehension test are distributed in response categories. The distribution of the answers given to the first item of the reading comprehension achievement test, which measures the cognitive process of focusing on and retrieving the clearly stated information, is given in Table 3.

Table 3. *Distribution of the answer to the first item (cognitive process: focusing on and retrieving explicitly stated information).*

Cognitive process	Response Category	*f*	%
Focusing on and retrieving explicitly stated information	The most correct answer	640	68.3
	Partially correct answers	116	12.4
	Blank answers	65	6.9
	Incorrect answers	69	7.4
	Unrelated answers	47	5.0

According to Table 3, most of the students gave the most correct answer in the process of focusing on and retrieving the explicitly stated information, which is the first cognitive level in the measurement of reading comprehension. Including the partially correct answers, more than 80.7% of the group gave the correct answer, while 19.3% failed to do so. Table 4 shows the distribution of the responses given to the second item, which measures the process of making straightforward inferences.

Table 4. *Distribution of the answer to the second item (cognitive process: making straightforward inferences).*

Cognitive process	Response Category	*f*	%
Making straightforwa rd inferences	The most correct answer	92	9.8
	Partially correct answers	323	34.5
	Blank answers	200	21.3
	Incorrect answers	251	26.8
	Unrelated answers	71	7.6

Looking at the given response categories in Table 4, only 9.8% of the group had the most correct answer in the cognitive process of making straightforward inferences. Students who gave correct answers together with those who gave partially correct answers constitute only 44.3% of the whole group. The number of students who gave blank, incorrect and unrelated answers in the process of making simple inferences is remarkable, more than half of the group. Table 5 shows the distribution of the answers given to the third item, which measures the process of interpreting and integrating ideas and information.

Table 5. *Distribution of the answer to the third item (cognitive process: interpreting and integrating ideas and information).*

Cognitive process	Response Category	*f*	%
Interpreting and integrating ideas and information	The most correct answer	103	11.0
	Partially correct answers	392	41.8
	Blank answers	235	25.1
	Incorrect answers	123	13.1
	Unrelated answers	84	9.0

Examining Table 5, it can be observed that only 11% of the answers to the item that measures the cognitive process of interpreting and integrating ideas and information are the most correct answer. Those who answered this item correctly constitute only half of the group. Similar to the case in the cognitive process of making straightforward inferences, students have quite a lot of blank, incorrect, and unrelated answers for this item. The number of students who gave the most correct answers were outweighed by the number of students who left the items blank or made it wrong. Table 6 shows the distribution of the responses given to the fourth item in the reading comprehension achievement test, which measures evaluating and critiquing the content and the textual elements.

Table 6. *Distribution of the answer to the fourth item (cognitive process: evaluating and critiquing content and textual elements)*

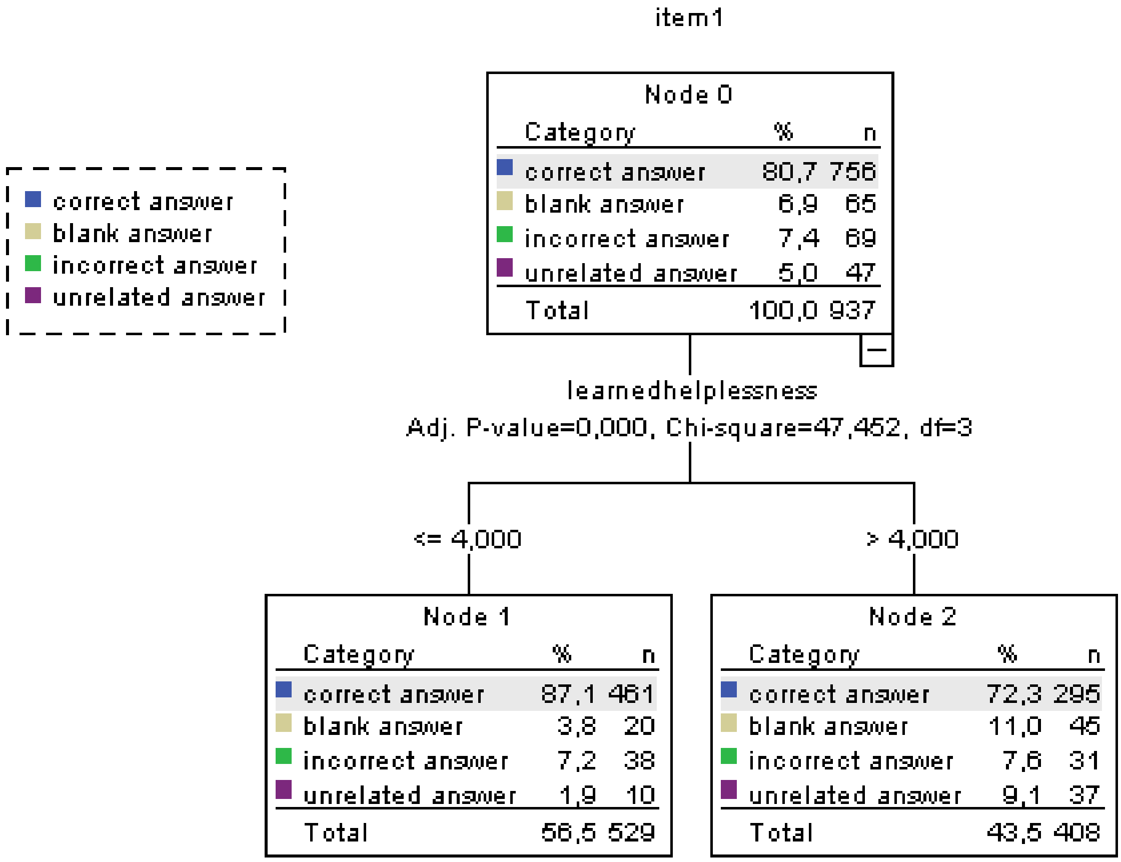
Cognitive process	Response Category	*f*	%
evaluating and critiquing content and textual elements	The most correct answer	22	2.3
	Partially correct answers	135	14.4
	Blank answers	393	41.9
	Incorrect answers	326	34.8
	Unrelated answers	61	6.5

The distribution of the responses given to the item focusing on evaluating and critiquing the content and elements of the text, which is the most complex level of reading comprehension, is given in Table 6. Accordingly, the rate of students who gave the most correct answer is only 2.3% of the group. Notably, almost half of the group (41.9%) left this item unanswered. The rate of those who gave incorrect answer to the item is one third of the group (34.8%).

3.2. Findings Regarding the Second Research Question

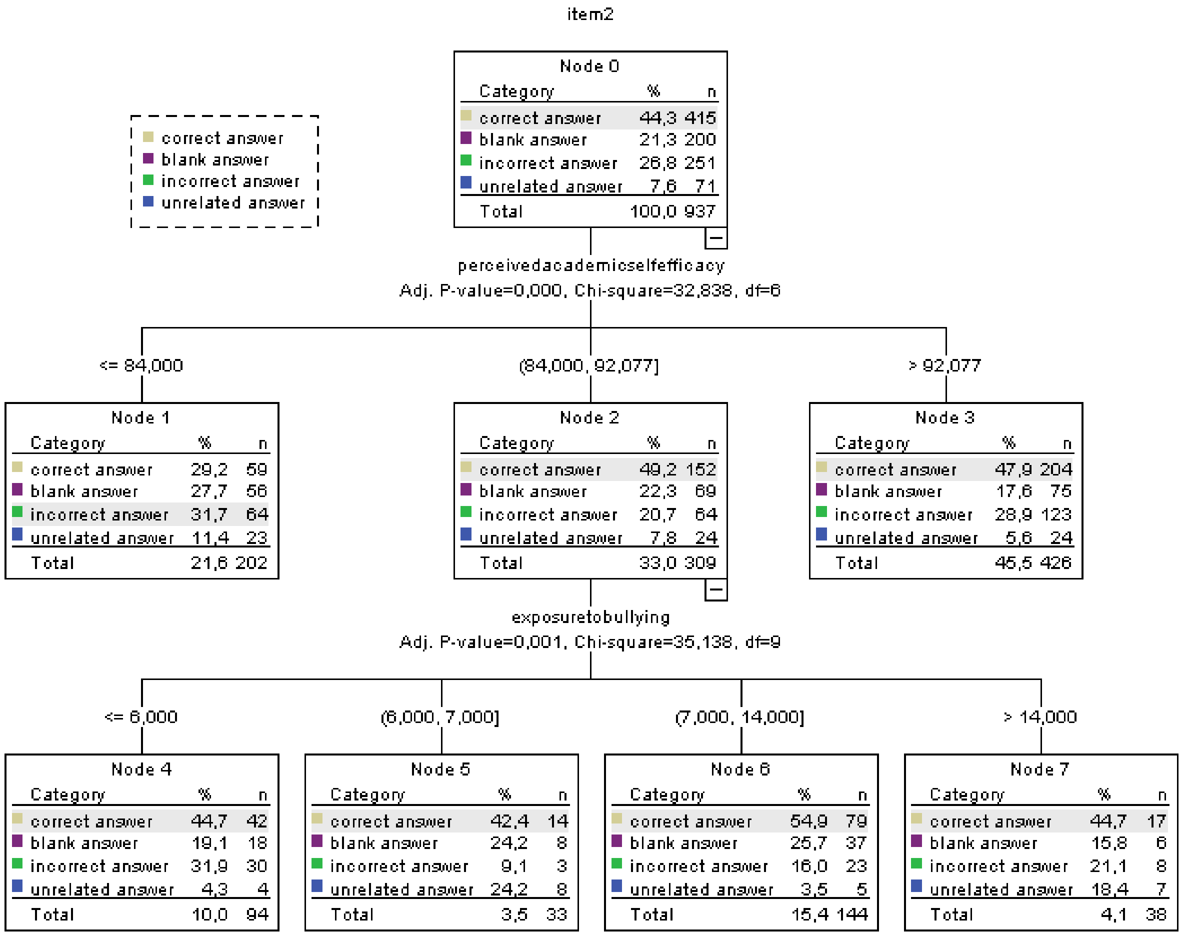
Figure 1 shows the tree graph created through the CHAID analysis to examine the role of achievement motivation, perseverance, test anxiety, perceived academic self-efficacy, exposure to bullying, and learned helplessness tendency on classification of the students based on their response categories.

Figure 1. *Decision Tree for the first item.*



Looking closely at the decision tree given in Figure 1, one variable is noted as affecting the students' incorrect, blank, and unrelated answers at the cognitive level of focusing and retrieving explicitly stated information. This variable is the learned helplessness tendency. Accordingly, 7.4% of the students in the first branch answered incorrectly; 6.9% gave blank, and 5.0% gave unrelated answers. Learned helplessness tendency has a strong impact on students' incorrect, blank and unrelated answers ($\chi 2=47.452$; df=3; $p<0.01$). While 12.9% of the students with a learned helplessness tendency score of 4 and below gave incorrect, blank or unrelated answers, 27.7% of the group with a learned helplessness tendency score above 4 gave incorrect, blank or unrelated answers. The decision tree was created to identify the variables that play a role in the incorrect, blank and unrelated answers given at the cognitive level of making straightforward inferences was presented in Figure 2.

Figure 2. *Decision Tree for the second item.*



Node 0

Category	%	n
correct answer	44,3	415
blank answer	21,3	200
incorrect answer	26,8	251
unrelated answer	7,6	71
Total	100,0	937

Node 1

Category	%	n
correct answer	29,2	59
blank answer	27,7	56
incorrect answer	31,7	64
unrelated answer	11,4	23
Total	21,6	202

Node 2

Category	%	n
correct answer	49,2	152
blank answer	22,3	69
incorrect answer	20,7	64
unrelated answer	7,8	24
Total	33,0	309

Node 3

Category	%	n
correct answer	47,9	204
blank answer	17,6	75
incorrect answer	28,9	123
unrelated answer	5,6	24
Total	45,5	426

Node 4

Category	%	n
correct answer	44,7	42
blank answer	19,1	18
incorrect answer	31,9	30
unrelated answer	4,3	4
Total	10,0	94

Node 5

Category	%	n
correct answer	42,4	14
blank answer	24,2	8
incorrect answer	9,1	3
unrelated answer	24,2	8
Total	3,5	33

Node 6

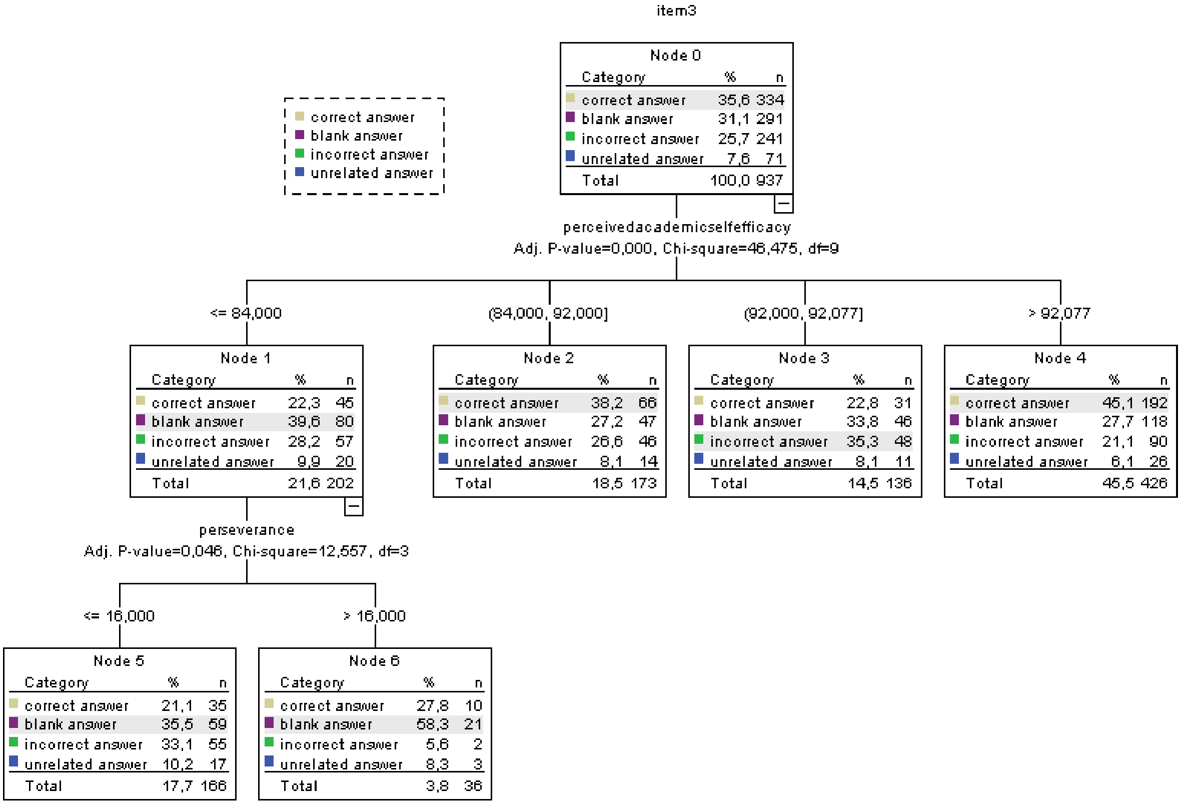
Category	%	n
correct answer	54,9	79
blank answer	25,7	37
incorrect answer	16,0	23
unrelated answer	3,5	5
Total	15,4	144

Node 7

Category	%	n
correct answer	44,7	17
blank answer	15,8	6
incorrect answer	21,1	8
unrelated answer	18,4	7
Total	4,1	38

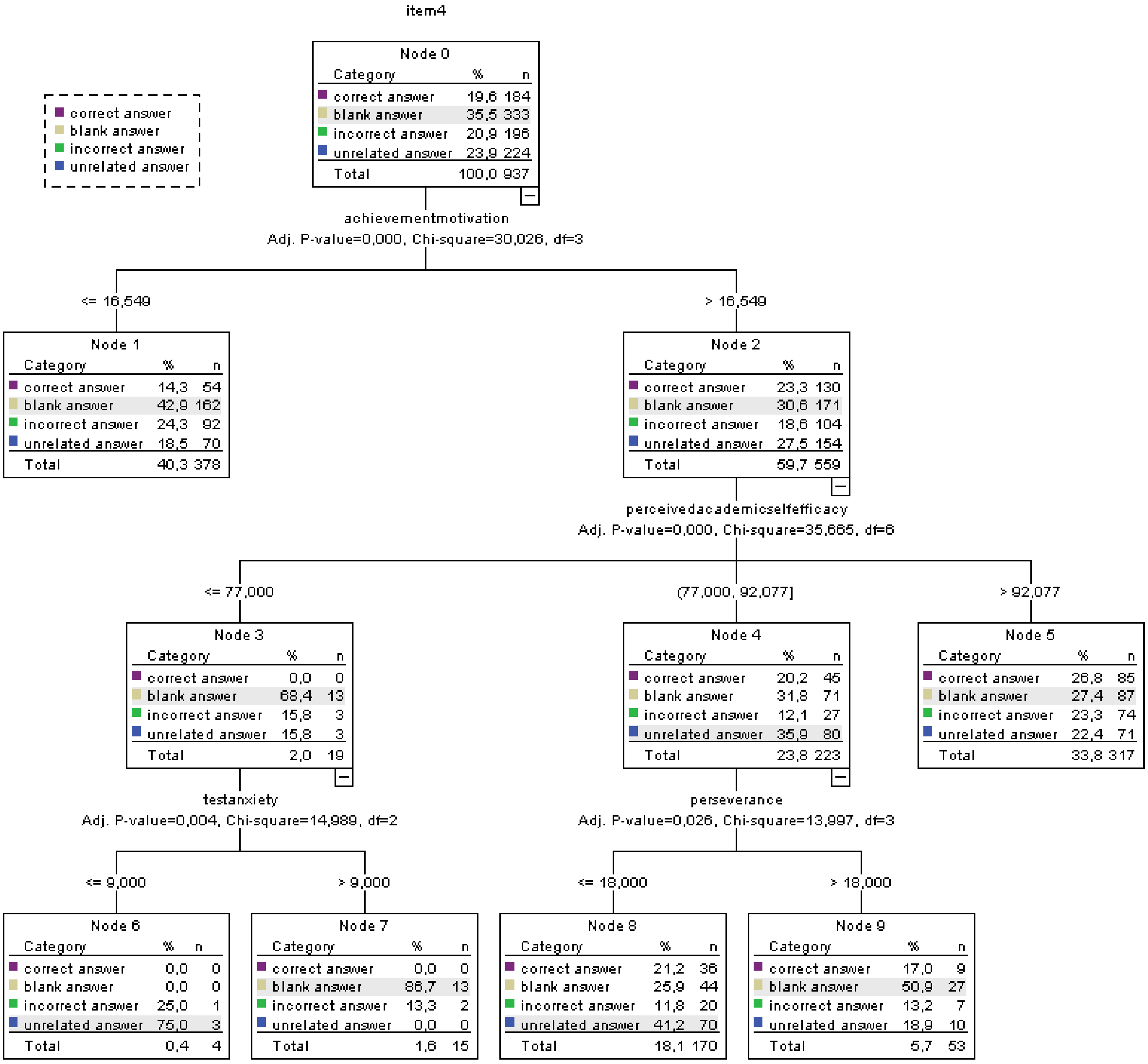
As shown in Figure 2, the strongest variable in classifying the students' responses at the cognitive level of making straightforward inferences was perceived academic self-efficacy (χ^2=32.838; *df*=6; *p*<0.01). Accordingly, students with a perceived academic self-efficacy of 84 points or less constituted 21.6% of the group; those between 84 and 92 constituted 33.0% of the group, and those with more than 92 points constituted 45.5% of the group. When the response distributions of the students were examined according to their perceived academic self-efficacy, 70.8% of the group with the lowest perceived academic self-efficacy gave incorrect, blank or unrelated answers. 50.8% of the students in the middle group and 52.1% in the top group gave incorrect, blank or unrelated answers. In terms of perceived academic self-efficacy, the branch in the middle group, where half of the group answered incorrectly or gave blank or unrelated answer, formed a knot again. In other words, the variable that classifies the answers given by students whose perceived academic self-efficacy score was between 84 and 92 points was the variable of being bullied (χ^2=35.138; *df*=9; *p*<0.01). The percentages of incorrect, blank and unrelated answers in the leaves prepared according to the scores obtained from the bullying scale were very similar (between 45.2% and 57.5%). When the leaves were examined in more detail, the differences in the response categories of the students according to the scores of being exposed to bullying are striking. Accordingly, as the scores obtained from the students' exposure to bullying subscale increase, the unrelated response rates increase as well. The decrease in the scores of being exposed to bullying can be seen in the leaves with a higher number of wrong answers. Figure 3 shows the tree created to examine the variables that have an effect on the answers at the level of interpreting and integrating ideas and information, which is the third cognitive level of reading comprehension.

Figure 3. *Decision Tree for the third item.*



When the decision tree given in Figure 3 is examined, it can be pointed out that the most important affective factor in the students' responses at the level of interpreting and integrating ideas and information was their perceived academic self-efficacy (χ^2=46.475; *df*=9; *p*<0.001). Accordingly, the rate of incorrect, blank and unrelated answers in the group with the highest perceived academic self-efficacy was 54.9%, while in the other levels of perceived academic self-efficacy, this rate ranged between 61.8% and 77.7%. The perseverance variable affected the answers of the students with the lowest perceived academic self-efficacy, who constituted 21.5% of the whole group and had the highest rate in terms of incorrect, blank and unrelated answers (χ^2=12.557; *df*=3; *p*<0.05). Accordingly, while 78.9% of the students with a perseverance score of 16 or less gave incorrect, blank or unrelated answers, the rate of incorrect, blank and answers was 72.2% for the students who scored higher than 19 points. What is noteworthy here is the distribution of these percentages to the answers. While students with low perceived academic self-efficacy and low perseverance had the same rate of blank and incorrect answers, 35.5% and 33.1% respectively, the rate of incorrect answers was only 5.6% for those with low perceived academic self-efficacy and high perseverance. The blank answers were 58.3%. The decision tree formed for the answers given for the evaluating and critiquing the content and textual elements, which was the final level of reading comprehension was presented in Figure 4.

Figure 4. *Decision Tree for the fourth item.*



When the decision tree presented in Figure 4 is examined, the first thing that stands out regarding the process of evaluating and critiquing the content and the textual elements is that the variables that classifies the answers of the students are achievement motivation (χ^2=30.026; *df*=3; *p*<0.01), perceived academic self-efficacy (χ^2=35.665; *df*=6; *p*<0.01), test anxiety (χ^2=14.989; *df*=2; *p*<0.01) and perseverance (χ^2=13.997; *df*=3; *p*<0.05). Accordingly, while 85.7% of students with achievement motivation scores below 16.5 points gave incorrect, blank or unrelated answers, this rate was 76.7% for the students scoring above 16.5 points. While the blank answers given by students with low achievement motivation (42.9%) were quite high, the same rate was 30.6% for the highly-motivated students (>16.5). In the incorrect and unrelated response categories, the percentage in the group with low motivation (42.8%) was lower than that in the high-motivation group (46.0%). The variable that affected the responses of the group with a high score on the achievement motivation scale was observed as perceived academic self-efficacy. Students who scored more than 16.5 points from the achievement motivation scale and had low perceived academic self-efficacy (<77) mostly gave blank answer, and they had no correct answer at all. While the rate of blank and unrelated answers was high for students with perceived academic self-efficacy scores between 77 and 92, a more balanced distribution was observed in the incorrect answers of students who scored more than 92 points. The test anxiety was the one of the variables that affected classifying the answers given in the cognitive process of examining and evaluating content and textual elements. As such, among the students with low perceived academic self-efficacy, incorrect and unrelated answers were observed by the students with low test anxiety (<9), while no blank answers were observed. Unrelated response behavior was never observed in students with high-test anxiety (>9). The students in this group mostly gave blank answers. The variable that classifies the distribution of students with moderate perceived academic self-efficacy (<77-92<) into response categories is the variable of perseverance. Accordingly, the answers of the students with a perseverance score of 18 points and below were mostly unrelated answers, while the students with a perseverance score of more than 18 points generally left the questions unanswered.

3.3. Findings Regarding the Third Research Question

The third problem of the study focuses on how the students' affective characteristics that has a role on classifying the responses differ according to the cognitive level of the item. Accordingly, the learned helplessness tendency variable was found to play a significant role in classifying student responses in the cognitive process of focusing on and retrieving explicitly stated information. Perceived academic self-efficacy and being exposed to bullying were found to have a significant role in classifying responses in the cognitive process of making straightforward inferences. Perceived academic self-efficacy and perseverance were found to play a significant part in classifying the responses in the cognitive process of interpreting and integrating ideas and information. Finally, perceived academic self-efficacy, achievement motivation, test anxiety, and perseverance variables were found to be significant in classifying the responses given to the cognitive level of evaluating and critiquing content and textual elements.

4. DISCUSSION and CONCLUSION

In this study, the answers of the fifth-grade students to the open-ended reading comprehension items were examined. The main point in this examination was to determine the affective characteristics that play a key role in classifying the incorrect, blank and unrelated answers indicating student failure. Thus, the aim was to outline a profile based on the affective characteristics of the students in these response categories, which is an indication of failure to understand what they read. In addition, based on the processes of measuring reading comprehension it was expected that the affective profiles of students may change. In other words, it was thought that the affective variables classifying the blank, incorrect and unrelated answers given to the most complex reading comprehension level, which is evaluating and critiquing content and textual elements, would not be the same with the variables classifying the blank, incorrect and unrelated answers given at the most basic reading comprehension level, namely focusing on and retrieving explicitly stated information. The change in affective characteristics that play a role in the classification of students' responses to items at different cognitive levels was also examined. The first conclusion can be drawn from the finding of the study is that as the reading comprehension cognitive processes became more complex, the correct response rates of the students decreased, and the rates of incorrect and blank answers increased. The unrelated response rates were close across all cognitive processes.

The reason why incorrect and blank response rates increase as cognitive processes become more complex is that students need to perform better and make deeper connections in the complex reading comprehension process (Mullis *et al.,* 2016). PISA 2018 assessment results also support this finding (OECD, 2019). Accordingly, students who perform at the highest proficiency levels (5 and 6) in reading literacy constitute only 3% of the whole group. Based on this finding, it can be suggested that in order to raise students with adequate and improved reading comprehension performance, students' blank and incorrect answers should be reduced

Another finding of the study is related to the affective characteristics of students who gave blank, incorrect or unrelated answers. As such, the most influential variable in classifying students in terms of blank, incorrect and unrelated answers at the level of focusing on and retrieving explicitly stated information is the learned helplessness tendency. When the answers at the level of making straightforward inferences were examined, it was seen that perceived academic self-efficacy was the most effective variable in classifying the answers given in this process. Students, who are at the group of lowest academic self-efficacy score, had the lowest rate of correct answers, and the highest rate in giving blank or unrelated answers. The students who are in the group of the higher academic self-efficacy group had the highest correct response rate, and the lowest rate of blank and unrelated responses. The variable that classifies the responses of students who are in the middle group on the perceived academic self-efficacy scale is exposure to bullying. The strongest variable in classifying student responses for the third cognitive process is perceived academic self-efficacy. For students in the group of lowest perceived academic self-efficacy, the strongest variable was perseverance. Thus, the number of blank and incorrect answers are high among students with low perceived academic self-efficacy and low perseverance whereas the group with low perceived academic self-efficacy and high perseverance had high blank response rates. The blank and incorrect response rates of students with moderate perceived academic self-efficacy were observed to be high while those with high perceived academic self-efficacy had the highest correct response rates. The distribution of blank and incorrect answers was similar. In the process of evaluating and critiquing content and textual elements, which is the most complex reading comprehension process, achievement motivation is the most effective variable in classifying student responses. It can be observed that the group with high achievement motivation, low perceived academic self-efficacy and low test anxiety gave mostly unrelated answers, while those with high test anxiety gave generally blank answers; the group with high achievement motivation, middle perceived academic self-efficacy and low perseverance were observed to give unrelated answers while those with higher perseverance were mainly watched to leave the items unanswered. The response distribution of students with high achievement motivation scores and high perceived academic self-efficacy were similar. In the group with low achievement motivation, the most common response was the blank response.

Considering the findings, the variables that shape the variations in thinking processes (and thus the categories) at different cognitive levels and the responses resulting from the thinking processes are diverse. Overall, the strongest variable in shaping the classification of the non-correct answers (incorrect, blank and unrelated answers) given in the simplest reading comprehension process is the learned helplessness tendency. In their longitudinal study, Fincham *et al.* (1989) concluded that learned helplessness plays a role in students' current reading comprehension success and their success two years later. Valås (2001) proved that helplessness is associated with academic performance. According to the result of this study, students with less learned helplessness tendency have low unrelated and blank answer rates, while the percentage of correct answers is high. Perceived academic self-efficacy is a meaningful classifier in all reading comprehension processes except from the first one. In general, students with low perceived academic self-efficacy gave blank and unrelated answers mostly, while the students with higher perceived academic self-efficacy scores mostly gave incorrect answers. Academic self-efficacy is often associated with student performance in the literature (Honicke & Broadbent, 2016; Nasir & Iqbal, 2019; Zysberg & Schwabsky, 2021). Komarraju and Nadler (2013) state that individuals with high self-efficacy have higher belief in what they can do. This finding explains why individuals with low academic self-efficacy have mostly blank and unrelated answers, while those with high academic self-efficacy have a high percentage of incorrect answers. The student with low academic self-efficacy may show the behavior of not responding to the item due to low belief in his/her own actions or writes unrelated things because he/she thinks the answer will be incorrect no matter what. The reason why students with high academic self-efficacy gave the highest number of incorrect answers may be because of their belief in the answer they will write. Their belief that the answer they give would be correct may have led them to respond to the item and to think while creating the answer. The variable of exposure to bullying is a significant classifier only for the process of making straightforward inferences. The literature reports that the social, personal and academic lives of students who are exposed to bullying are affected from those experiences (Strøm *et al.,* 2013). Within the scope of the current study, the rate of giving non-correct answers by the children who were bullied was quite high. This finding is in parallel with the literature. Considering why being bullied is only effective on cognitively basic reading comprehension processes, the very first thing to come into one's mind is its possible relation to students' backgrounds. Though bullying can take place regardless of the schools' and students' socio economic and cultural background, it is a fact that it is much more common in economically disadvantaged schools (Bowes, 2009; Lumeng *et al.,* 2010). Students attending those schools are poor performers (are able to write answers to cognitively simple items such as making straightforward inferences; however, fail to properly perform for complex cognitive processes). This situation is thought to be the reason behind this finding.

The affective features that affect students' answers in the second or third degrees in classification are perseverance and test anxiety. The relationship between perseverance, test anxiety and academic performance has been given a substantial focus in the literature (Cassady & Johnson, 2002; Chapel *et al.,* 2005; Culler & Holahan, 1980; Kutlu *et al.,* 2017). It is very particular to note that students with low anxiety gave largely unrelated answers and those with high anxiety left the items mostly unanswered. High anxiety has both psychological and physical consequences that prevent students' cognition from working properly, due to which the student may avoid answering items. Low anxiety, on the other hand, shows that the student does not care about the academic task at all. The behavior of not caring can also lead the student to write the answers s/he wants and writes meaningless words instead of the answers required by the item. Regarding perseverance, there are mostly incorrect answers when low perseverance is coupled with low self-efficacy, and many blank answers when while when perseverance is high with low self-efficacy. There are unrelated answers when moderate self-efficacy is coupled with low perseverance, and blank answers in case of high perseverance with moderate self-efficacy.

The final variable that is influential in classification is achievement motivation. While students with low achievement motivation have more blank and incorrect answers, students with high achievement motivation have more unrelated answers. This finding is inconsistent with the researcher's expectations. What was expected that students with high achievement motivation would have more incorrect answers than unrelated ones. This finding can be interpreted as the fact that the Turkish education system may not adequately prepare students for higher order thinking processes. The fact that achievement motivation, which is the strongest classifier at the most complex level of reading comprehension, plays a role in students' unrelated response can be interpreted in two ways. First, students were so focused on answering the item and being successful that they answered the item even if it was not meaningful. The second interpretation may be related to the failure of students to show the expected performance in the process of evaluating and critiquing the content and textual elements. This may have affected the algorithm of the analysis method used.

When the classifiers of the answers examined, it becomes clear that learned helplessness tendency is important at the level of focusing on and retrieving explicitly stated information; perceived academic self-efficacy and exposure to bullying in making straightforward inferences; perceived academic self-efficacy and perseverance in interpreting and integrating ideas and information, and finally, achievement motivation, perceived academic self-efficacy, test anxiety, and perseverance in evaluating and critiquing the content and elements of the text. As can be seen, as cognitive levels become more complex, more variables are involved in classifying students' cognitive performance.

5. LIMITATIONS OF THE STUDY and SUGGESTIONS

The most important limitation of the study is that the answers given to the items written at different cognitive levels were measured with a single item at each level. The reason for this limitation is the poor performance of students' reading and reading comprehension skills probably due to the pandemic process. In 40 minutes (one class hour), the students had difficulty in reading the text and answering the test items. For that reason, the number of items was limited to four, and each cognitive level could only be measured through one item. Thus, further research may include more items for each comprehension level.

Based on the findings of this study, it can be suggested that efforts must be made to help students who give incorrect, blank or unrelated answers in classroom activities. They should be encouraged to overcome their past failures, increase their self-confidence and self-efficacy. Teachers are advised to organize their classroom settings in a way that does not allow peer bullying, guides students to continue their goal-oriented behavior despite the difficulties that may arise, and increases their motivation for success.

As for in-classroom practices, it can be suggested that teachers should include open-ended items in the classroom assessment and evaluation processes and use rubrics to show students the content and category of their answers and the correct answer performance expected from them. In this way, students can understand where they are at and how they can improve themselves. By doing so, the number of students who give incorrect or unrelated answers can be reduced. Another suggestion that can be made based on the findings is the planning of curriculum and taking precautions that will activate the appropriate affective processes of the students and prepare them to learn better. By observing the helplessness experiences of socially disadvantaged students more closely, the teacher can implement proper psychoeducational practices that can prevent this experience of the student. Finally, the teacher, who discusses sample response categories with the help of rubrics in the classroom, can increase his/her students' motivation for success by raising their perception of what they are doing, and the student, who knows about the expected performance, can take a more objective stance regarding his/her own self-efficacy, and have a chance to improve it.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number**: Ankara University, Social Sciences Sub-Ethics Committee, 30/03/2020-3-54.

Orcid

Neslihan Tugce OZYETER https://orcid.org/0000-0003-1558-1293

REFERENCES

Badger, E. & Thomas, B. (1991) Open-Ended questions in reading. *Practical Assessment, Research, and Evaluation, 3*(4). https://doi.org/10.7275/fryf-z044

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American psychologist, 37*(2), 122. https://doi.org/10.1037/0003-066X.37.2.122

Bowes, L., Arseneault, L., Maughan, B., Taylor, A., Caspi, A., & Moffitt, T.E. (2009). School, neighborhood, and family factors are associated with children's bullying involvement: A nationally representative longitudinal study. *Journal of the American Academy of Child*

& Adolescent Psychiatry, 48(5), 545-553. https://doi.org/10.1097/CHI.0b013e31819cb017

Cassady, J.C., & Johnson, R.E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270-295. https://doi.org/10.1006/ceps.2001.1094

Chapell, M.S., Blanding, Z.B., Silverstein, M.E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268–274. https://doi.org/10.1037/0022-0663.97.2.268

Culler, R.E., & Holahan, C.J. (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology, 72*(1), 16–20. https://doi.org/10.1037/0022-0663.72.1.16

Dweck, C.S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040–1048. https://doi.org/10.1037/0003-066X.41.10.1040

Fincham, F.D., Hokoda, A., & Sanders Jr, R. (1989). Learned helplessness, test anxiety, and academic achievement: A longitudinal analysis. *Child development, 60*(1), 138-145. https://www.jstor.org/stable/1131079

Ghasemi, F. (2021). A motivational response to the inefficiency of teachers' practices towards students with learned helplessness. *Learning and Motivation, 73*(1), 101705. https://doi.org/10.1016/j.lmot.2020.101705

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of educational research, 58*(1), 47-77. https://doi.org/10.3102/00346543058001047

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review, 17*(1), 63-84. https://doi.org/10.1016/j.edurev.2015.11.002

IBM SPSS. Decision trees 21. IBM Cooperation.

Kass, V.G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119-127. https://www.jstor.org/stable/2986296

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences, 25*(1), 67-72. https://doi.org/10.1016/j.lindif.2013.01.005

Kutlu, Ö., & Özyeter, N.T. (in press). Development of the learned helplessness tendency scale for children: Validity and reliability studies. Studies in Psychology.

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2017). *Ölçme ve değerlendirme performansa ve portfolyoya dayalı durum belirleme [Measurement and evaluation, assessment based on performance and portfolio]*. Pegem Akademi.

Lumeng, J.C., Forrest, P., Appugliese, D.P., Kaciroti, N., Corwyn, R.F., & Bradley, R.H. (2010). Weight status as a predictor of being bullied in third through sixth grades. *Pediatrics, 125*(6),1301-1307. https://doi.org/10.1542/peds.2009-0774

Manzano-Sanchez, H., Outley, C., Gonzalez, J.E., & Matarrita-Cascante, D. (2018). The influence of self-efficacy beliefs in the academic performance of Latina Students in the United States: A systematic literature review. *Hispanic Journal of Behavioral Sciences, 40*(2), 176–209. https://doi.org/10.1177/0739986318761323

Mullis, I.V., Martin, M.O., & Sainsbury, M. (2016). *PIRLS 2016 reading framework*. PIRLS, Chapter-1, 11-29. https://timss.bc.edu/pirls2016/downloads/P16_FW_Chap1.pdf

Nasir, M., & Iqbal, S. (2019). Academic self-efficacy as a predictor of academic achievement of students in pre-service teacher training programs. *Bulletin of Education and Research, 41*(1), 33-42. https://files.eric.ed.gov/fulltext/EJ1217900.pdf

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Elsevier.

Olivier, E., Archambault, I., De Clercq, M., & Galand, B. (2019). Student self-efficacy, classroom engagement, and academic achievement: Comparing three theoretical frameworks. *Journal of Youth and Adolescence, 48*(2), 326-340. https://doi.org/10.1007/s10964-018-0952-0

Organisation for Economic Co-Operation and Development. (2014). *PISA 2012 technical report. OECD Publishing*. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organisation for Economic Co-Operation and Development. (2017). *PISA 2015 technical report. OECD Publishing*. https://www.oecd.org/pisa/data/2015-technical-report/

Organisation for Economic Co-Operation and Development (2019). *PISA 2018 results (Volume II): Where all students can succeed*, PISA, OECD Publishing, https://doi.org/10.1787/b5fd1b8f-en

Özyeter, N.T., & Kutlu, Ö. (2022). Adaptation of the children's perceived academic self-efficacy scale: Validity and reliability study. *International Journal of Assessment Tools in Education, 9*(2), 430-450. https://doi.org/10.21449/ijate.958871

Parvez, M., & Shakir, M. (2014). Academic achievement of adolescents in relation to academic anxiety, gender, and choice of academic stream. *Research on Humanities and Social Sciences, 4*(1), 107-115. ISSN (Paper) 2224-5766, ISSN (Online) 2225-0484 (Online)

Popham, W.J. (1999). *Classroom assessment: What teachers need to know*. Allyn & Bacon. http://www.abacon.com

Román, M., & Murillo, J. (2011). *Latin America: School bullying and academic achievement. Cepal* Review. https://repositorio.cepal.org/bitstream/handle/11362/11502/104037053I_en.pdf?sequence=1&isAllowed=y

Strøm, I.F., Thoresen, S., Wentzel-Larsen, T., & Dyb, G. (2013). Violence, bullying and academic achievement: A study of 15-year-old adolescents and their school environment. *Child Abuse & Neglect, 37*(4), 243-251. https://doi.org/10.1016/j.chiabu.2012.10.010

Valås, H. (2001). Learned helplessness and psychological adjustment II: Effects of learning disabilities and low achievement. *Scandinavian Journal of Educational Research, 45*(2), 101-114. https://doi.org/10.1080/00313830120052705

van der Werf, C. (2014). The effects of bullying on academic achievement. *Revista Desarrollo y Sociedad, 74*, 275-308. https://doi.org/10.13043/dys.74.6

Walling, M.D., & Martinek, T.J. (1995). Learned helplessness: A case study of a middle school student. *Journal of Teaching in Physical Education, 14*(1), 454-466. https://core.ac.uk/reader/213401840

Zahrakar, K. (2008). The relationship between parents' child rearing practice and young adults' mental health in Islamshahr. *Innovation in Management Education (Journal of Modern Thoughts in Education), 3*(2), 71-90. https://www.sid.ir/en/Journal/ViewPaper.aspx?ID=181803

Zysberg, L., & Schwabsky, N. (2021). School climate, academic self-efficacy and student achievement. *Educational Psychology, 41*(4), 467-482. https://doi.org/10.1080/01443410.2020.1813690

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 46–71

https://doi.org/10.21449/ijate.1132923

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of this article.

Inside the black box: do teachers practice assessment as learning?

Ozen Yildirim[1], Safiye Bilican Demir[2*]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye
[2]Kırıkkale University, Faculty of Education, Department of Educational Sciences, Kırıkkale, Türkiye

ARTICLE HISTORY

Received: June. 19, 2022
Revised: Aug. 18, 2022
Accepted: Sep. 1, 2022

Keywords:
Formative assessment,
Assessment for learning,
Assessment as learning,
Self-regulation,
Self-assessment.

Abstract: The conceptual development of assessment literature in recent years has been remarkable. One of the latest concepts to have emerged in parallel with this development is Assessment as Learning (AsL). This study investigated how AsL pertains to classroom practices within its conceptual framework by examining teacher reports. Case study design, a qualitative research method, was used to collect detailed information about in-class teacher practices. The teachers were interviewed with semi-structured interview forms and the data obtained were then analyzed using content analysis. The results revealed that in-class teacher practices were incapable of supporting AsL and promoting self-regulated behaviors and that many of the activities conducted in class were teacher-centered. Teachers did not apply self-assessment or peer-assessment practices, and the feedback they gave to students was mainly based on measurement scores. The researchers discussed the results in relation to the relevant literature and offered some suggestions for applying AsL in practice.

1. INTRODUCTION

Assessment greatly impacts student learning. Given the relationship between assessment and learning, it is no surprise that many studies have been made that examine this relationship. The terms formative assessment and summative assessment have been widely used in assessment literature, particularly since the 1990s. Formative assessment is used to support and improve student learning, whereas summative assessment is used for certification, ranking, or accountability purposes concerning student achievement.

The literature on formative assessment has continued to develop with different concepts for more than 30 years: mastery learning programs in the 1970s and 80s (Bloom, 1974; Popham, 1978), feedback-based assessment approaches (Sadler, 1989), and issues related to measuring, reporting, and profiling success in the 1990s (Torrance, 1991). In 1999, the Assessment Reform Group (ARG), an influential group of educational researchers within the United Kindom (UK),

*Corresponding Author: Safiye Bilican Demir ✉ safiye.demir@kku.edu.tr Kırıkkale Universitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Kırıkkale, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

used the concepts of "Assessment of Learning" (AoL) and "Assessment for Learning" (AfL) for summative and formative assessment, respectively, by increasing the emphasis on learning in the assessment process (ARG, 1999). While AoL is generally used to judge measurement results and performance after a formal learning activity, AfL serves the purpose of improving the process of learning and teaching (ARG, 1999; Earl, 2003). In addition, based on Black and William's (1998) review of the literature and this study, it can be seen that debates about formative assessment are influenced by the studies made by the British ARG (2002) and Black et al. (e.g., Black et al., 2003, 2006) and focus on "assessment for learning."

The review study conducted by Bennett (2011) stated that the most frequently repeated definition of formative assessment, more specifically AfL, is that it is an assessment method that provides both students and teachers with feedback on student development and what more can be done to facilitate this development. Bennett (2011) also mentioned two different goals that stand out when conceptualizing AfL. The first is to develop diagnostic measurement tools within the scope of full learning tradition, and the second is informal ways to understand student outcomes and steer their learning. When the current definitions are examined, the second goal is more prominent. For example, the definition of AfL made by the UK Assessment Reform Group (2002) is as follows: "Assessment for Learning is part of everyday practice by students, teachers, and peers that seeks, reflects upon, and responds to information from dialogue, demonstration, and observation in ways that enhance ongoing learning." However, this definition has been criticized, particularly in classroom practices, because it focuses too much on achieving narrow learning program goals through tests (Swaffield, 2011; Torrance, 2012). In parallel with these criticisms, Klenowski (2009) made a definition of AfL that is "more pedagogical" and focuses on learning: "AfL is a part of everyday life activities in which the individual obtains information from conversations and observations, reflects this in his thoughts and actions, and reacts to it."

Conceptual discussions about AfL are also closely related to learning theories. When compared to the social constructivist approach, the role of assessment is completely different from the behavioralist traditional approach. In the behavioralist approach, we define learning goals, teach them specifically to students, and ensure that teachers know what "counts" for students to achieve that goal; that is, they know what behaviors are needed to complete the task at hand. This indicates a very well-structured and hierarchical approach in terms of organizing the syllabus and assessment processes, just like "building blocks." The social constructivist approach influenced by Vygotsky's (1978, 1986) arguments is treated as an interaction rather than a "transference" of knowledge and understanding. This interaction takes place between student-teacher, student-task, and student-student. Consistent with Vygotsky's (1978, 1986) arguments, what matters is determining what students have learned (what they have achieved or failed to achieve), as well as what they can achieve or are ready to achieve with teacher support or, in some cases, peer collaboration.

Although these two theoretical approaches assign different roles to assessment, in the accountability system, which focuses on test results, the development of AfL appears stuck in the "past" to a great extent (Torrance, 2012). More specifically, many studies addressing the relationships between AfL and learning (e.g., Graham et al., 2015; Klute et al., 2017; Lee et al., 2020; Hattie & Timperley, 2007), took increased test scores into account and highlighted "raising standards" (Torrance, 2012). This situation is seen as an important problem since an increase in test scores is not always an indicator of real improvement in academic achievement (Wyse & Torrance, 2009). When test scores are the only way to assess improvement, teachers have to increase test scores or exam results instead of focusing on the students' learning experiences or the diversity in their learning outcomes. This problem can be called "inside the black box." In other words, education policies in many countries treat the classroom as a "black box" (Black & Williams, 1998). To increase the quality of education, the focus is on changing the inputs (e.g., teacher quality, standards for student achievement, technical and educational resources) and mainly using standard achievement tests to assess the outputs. This means that little or no consideration is given to what is happening in the classroom.

Conceptual discussions about AfL, particularly the criticisms of addressing AfL in relation to test scores, prompted us to consider the relationship between assessment and learning from a different perspective. In 2003, Earl (2003, 2013) added a new concept to assessment literature: Assessment as Learning (AsL). AsL is a key concept that facilitates learner independence and flexibility to improve learning. AsL refers to the development of learning by incorporating environments that support self-assessment, self-efficacy, and other self-regulated behaviors into the teaching and assessment processes (Dann, 2014; Earl 2013; Torrance 2007). With this suggestion by Earl, we can see that the scope of AfL has expanded in terms of assessing the role of the learner (student) in the link between the assessment and learning processes (Dann, 2014). From an AsL perspective, the student becomes involved in the learning process when his metacognitive and self-regulated skills are supported, and this in turn directly supports the learning process (Black et al., 2003; Lam, 2014). Students in a classroom organized according to AsL have more of a say in steering the learning process. At the same time, they understand the learning objectives and evaluation criteria and can their metacognitive skills to provide quantitative and qualitative feedback to steer their future learning (Davies & LeMahieu, 2003; Ferris & Hedgcock, 2014).

Conceptual definitions of AsL show that feedback, self-regulation, and self-assessment are key components of AsL. Up until the mid-20th century, the behaviorist learning approach was applied to feedback, and it was seen as a reward or a punishment that either increased or decreased learning (Kruger & Denisi, 1996; William, 2018). The definition of feedback evolved over the 20th century in line with changes in learning theory with the behaviorist approach dominant until the mid-20th century before being superseded by the cognitive and constructivist approaches (Brookhart, 2018). Black and William (2006) stated that since feedback for student studies reflects knowledge and understanding of student performance, it is accepted as an integral part of the learning process. Many empirical studies have confirmed the positive impact of effective feedback on learning outcomes (Butler & Winne, 1995; Clark, 2012; Manuel, 2015).

In addition to academic performance, effective feedback is also discussed in terms of its relationship with the other component of AsL, i.e., self-regulated features. Andrade and Brookhart (2016) and Clark (2012) suggested that feedback can support students' self-regulated learning and that it complements self-assessment in improving learning outcomes. These discussions reveal that effective feedback improves learning outcomes both directly and indirectly through self-regulated features. As a separate component, self-regulated behaviors are considered a form of self-regulation, a more general concept, that has been adapted to educational settings (Dinsmore et al., 2008). Zimmerman (2000) defined self-regulated learning as cyclically adapted self-generated action, emotion, and thought planned so as to achieve personal goals. According to these explanations, self-regulated learning refers to the mental, metacognitive, emotional, and motivational processes that learners go through while striving toward a goal. Weinstein et al. (2011) considered all cognitive, metacognitive, emotional, and motivational self-regulated learning processes used by students as "learning strategies" applied to generate meaningful learning content. The results of a meta-analysis of studies conducted at different grade levels confirmed that self-regulated learning has a positive effect on academic achievement (Dignath & Büttner, 2008; Theobald, 2021).

When considering self-regulated learning in the context of AsL, another structure that affects this relationship stands out: self-assessment. Self-assessment is an important component of formative assessment (Assessment Reform Group 1999) and is defined as students evaluating their own work according to well-defined and understandable criteria and standards to improve their learning or performance (Brown & Harris, 2013). Panadero et al. (2018) stated that self-assessment is a critical self-regulated behavior. Recent meta-analysis studies have shown that self-assessment positively affects self-regulated learning (Andrade, 2019; Panadero et al., 2017) and success (Andrade, 2019).

Another important issue when discussing self-regulated learning is the presence of "others" (Panadero et al., 2018). "Others" interact with the learner and assist him in completing the task and regulating his actions (McCaslin & Hickey, 2001). In classroom settings, this interaction can take place with peers as well as an expert (teacher) (Andrade & Brookhart, 2016). Peer assessment is defined as the arrangements in which the success, quality, or value of the individual's product or learning outcomes are evaluated by their peers of equal status (Topping, 1998). The benefits of peer assessment for learning outcomes seem to be closely related to Vygotsky's (1978) theory of social development, which says that a child's development occurs through interaction with peers, teachers, and/or parents within a community and that a rich social environment supports learning and development by strengthening this interaction. Peer assessment can support students' cognitive development (Topping & Ehly, 2001), metacognitive awareness (Kim & Ryu, 2013), and social-affective development (van Gennip et al., 2009). Furthermore, if students are actively involved in peer assessment, they can be more autonomous learners (Bloxham & West 2004).

Discussions about formative assessment as a whole emphasize the active role of the student in the learning process. Studies made in the past 10 years show that the focus has shifted from the teacher to the learner (Lee et al., 2020). Assessment approaches including AoL as well as AsL need to become widespread to help students cope with the challenges they will face with their future learning and to support the lifelong learning process (Boud & Falchikov, 2006).

Despite the developing literature on AfL in the past two decades, it is noteworthy that studies on AsL as a sub-concept of AfL are particularly concentrated at the conceptual level. Given the breadth of definitions and the diversity in educational contexts, it is not easy to understand AsL, AfL, and AoL completely and accurately (Baird et al., 2017). In their comprehensive review study, Black and William (1998) pointed out that AfL remained a "weak" teaching practice. Remarkably, even though more than 20 years have passed, this situation is still true today. Recent studies have revealed that AfL and AsL are concepts not well understood by teachers (Dann, 2014; Lam 2013). Marshall and Drummond (2006) emphasized that formative assessment practices of teachers are often "convergent," that is, they focus on whether students have achieved the goals set in the syllabus. In other words, empirical evidence confirms that the correct and effective use of formative assessment in practice is still incomplete. This reveals how AsL, a somewhat new concept, pertains to teacher practices and shows that more studies are needed to determine its place in classroom practices.

Study results based on actual practice will facilitate our understanding of the interrelationships between assessment, learning, and teaching in a school context. It appears we need to examine what happens inside the classroom as a "black box," particularly in the context of preparing programs and content in support of AsL and to support teachers' skills in these practices. Although the conceptual framework of AsL has been defined, this study differs in that it reclarifies this concept to make it easier to understand AsL's complex conceptual framework and facilitate its integration into the learning process. In addition, this study methodologically focuses on what teachers do in practice beyond external test scores. Lam (2020) argues that this concept should be investigated with the best qualitative methods due to the process-oriented, content-sensitive, and reflexive nature of AsL. In light of all this, the researchers decided it was best to use qualitative methods to determine how AsL pertains to teacher practices.

This study aimed to collect information about the in-class teaching and assessment activities of teachers working in primary schools and to assess them in terms of their congruence with the conceptual framework of AsL. To this end, the researchers interviewed the teachers and asked them about the approaches they adopted with respect to the learning process, what they used measurement tools for, and feedback.

2. METHOD

2.1. Research Design

The study was conducted as a single case study based on qualitative research methodology. Case studies involve the researcher using qualitative data collection methods to collect in-depth information about cases in real life or those bounded by time (Creswell, 2018). This study focused on whether the teacher practices supported the critical components of AsL and attempted to reveal existing practices in detail using the information obtained from the interviews.

2.2. Participants

The participants of the study were 16 teachers teaching language, mathematics, and science in different secondary schools. In order to select most participants most beneficial to the study, maximum variation, volunteer (convenience), and criterion sampling were used together. Accordingly, taking into account ease of accessibility, the participants were selected from three Turkish provinces (Denizli, Istanbul and Kocaeli). The researchers were careful to select teachers who taught different subjects in different middle schools to ensure maximum variation.

Data saturation was considered when deciding on the number of samples (Hennink et al., 2017). According to Francis et al. (2010), saturation is a key indicator that the study's sample size is sufficient, and states that the collected data should the diversity, depth, and nuances of the topics being examined, thus ensuring content validity. Data collection was terminated when no new information was obtained from interviews and when the data began to repeat. It was, therefore, assumed that it was no longer necessary to collect more data and that adequate sample size had been reached. Table 1 describes the participants.

Table 1. *Demographic characteristics of the participants*

Participants	Gender	Seniority	Branch
P1	Female	16	Language Teacher
P2	Male	11	Mathematics
P3	Female	15	Science
P4	Male	14	Language Teacher
P5	Female	12	Language Teacher
P6	Female	8	Language Teacher
P7	Female	12	Science
P8	Female	12	Mathematics
P9	Male	22	Mathematics
P10	Male	12	Mathematics
P11	Female	12	Mathematics
P12	Female	20	Science
P13	Female	20	Mathematics
P14	Male	21	Language Teacher
P15	Male	18	Science
P16	Male	19	Mathematics

Table 1 shows that nine of the participants were female and seven were male. Seven of the participants were mathematics teachers, four were science teachers and five were language teachers, and their seniority varied between eight and 22 years.

2.3. Instrumentation and Procedures

The research data were collected by online meeting using a semi-structured interview form. An interview form was prepared based on the related literature of ASL and presented to expert opinions. The steps suggested by Cresswell (2018) were taken into account when planning the interview steps. First of all, open-ended and general questions were created in line with the research problem. In addition to these questions, sub-questions were asked to obtain more detailed information depending on how the interview progressed. The interview form contained three general questions that represent the critical concepts of AsL relevant to planning and teaching the lesson, student autonomy, and measurement-feedback. These questions were: "How do you start your teaching process and how do you proceed, what do you pay attention to?" "What are the roles of the teacher and the student in the learning process?" and also "What do you do to measure and assess your students' progress? How do you use feedback?" In addition, subquestions questions were asked to elaborate the questions and make them better understood. Experts (three in measurement and assessment, two native tongue experts, and two teachers) examined the interview form to check for clarity, comprehensibility, and suitability for the study. In accordance with the feedback from the experts, additional explanations about some concepts (for example, learning strategies) were included in order to enable the participants to understand and easily answer the questions in the interview form. In addition, probe (sub) questions were added to provide more detailed information about the measurement tools and the use of feedback. All of the experts reported that the questions could be answered by the teachers.

In line with the participants' preferences, the researchers decided to conduct the interviews in online meetings so they could take place in a relaxed and practical setting. Before the participants were interviewed, the researchers conducted pilot interviews with two teachers. All of the interviews were scheduled by making prior appointments and agreeing on a time. The researchers recorded each interview on video while paying attention to the quality of the sound and image. To gain the participants' trust, the researchers briefed them about the study and told them the video and audio recordings would be kept confidential, and that their identities would not be shared. They also asked the participants to participate voluntarily in the interviews. In addition, they told the participants to truthfully explain the actual situation, not the ideal situation, when answering the questions. Both researchers conducted the interviews, which lasted 30 to 40 minutes each. Ethics Committee Permission (Document No: E-93803232-622.02-193607) was obtained from Pamukkale University Institute of Social Sciences before the study began.

2.4. Data Analysis

The researchers followed the content analysis steps suggested by Berg and Lune (2016) and Cresswell (2018) when analyzing the data. They first collected data using interviews and then watched the videos to check whether there were any problems with the data recording. Once it was determined that the recordings were fine, one researcher transcribed the audio recordings verbatim. Analytical codes were then developed. This involved one of the researchers creating a code list and the other researcher re-coding the interviews using this code list. The codes that were not in the coding list or that the other coder could not determine were revised and this process continued until the two coders were in accord. The commonly used codes were grouped to determine categories and sub-themes. The sub-themes were then grouped to find the main themes. Two academics with experience in qualitative research were consulted to determine the logical fit of the main themes, sub-themes, and categories and to see if they were appropriate for the study. Once consistency between themes and categories was assured, the researchers presented the findings obtained in the relevant theme and sub-theme together with examples taken from the teachers' comments. Furthermore, two other experts in the field coded two videos selected at random using these themes and the table containing the codes to check for consistency between the researchers and the other coders.

2.4.1. *Validity and reliability of the study*

The researchers pursued four strategies to ensure the validity of the study: clarifying researcher bias, member checking, rich thick description, and external audit (Creswell & Miller, 2000). Based on their previous interview experiences and because they are experts in measurement and assessment in education, the researchers briefly explained the purpose of the study to the participants at the start of the interview and told them how important it was that they answer based on their own in-class practices. Member checking involved creating a focus group consisting of four out of the 16 interviewees and asking them to evaluate the results. The participants were asked what they thought of the analyses and to offer any additional opinions they might have. All of the participants in the focus group said the findings reflected their views. Rich and thick descriptions were added to increase validity. This involved writing a detailed report about the participants, data collection, and the findings obtained from the interviews. Finally, for the purposes of external audit, the opinion of the two external researchers were asked to examine and evaluate the research process and the results. It can be said that the steps taken for validity also support the reliability of the study. The codes generated by the two researchers were checked for consistency with one another, then the first researcher created the code list. The other researcher used this code list to see if their codes were consistent with those of the outside coders. To do this, they checked for consistency between the codes by applying them to three randomly selected videos. Furthermore, to ensure external audit, the two coders, who had nothing to do with one another and the study, were asked to compare the relevant coding list with their own coding lists. Themes and codes that did not match were revised again.

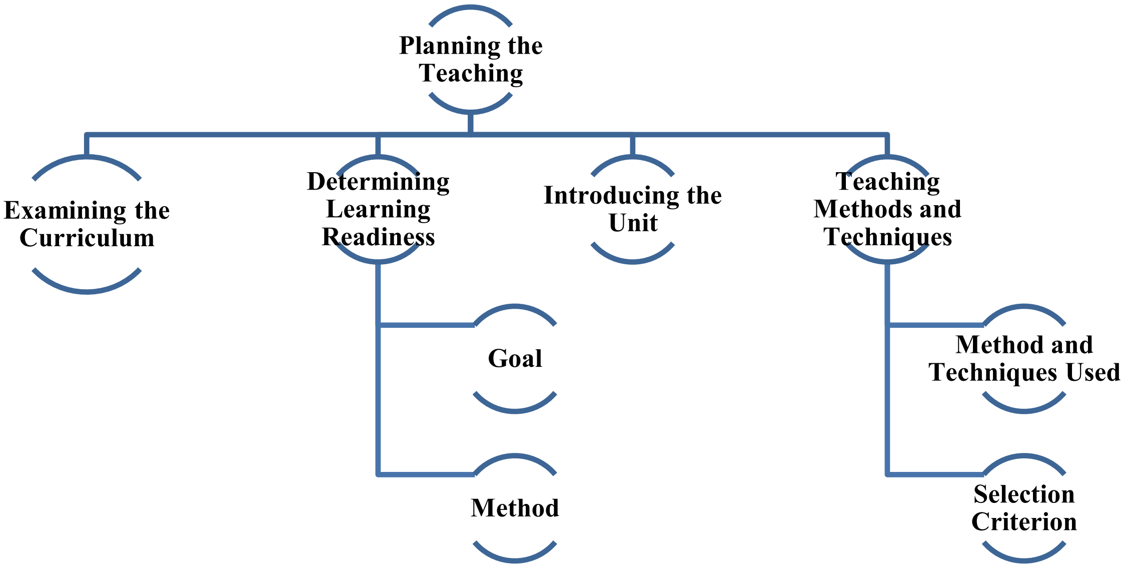
3. FINDINGS

The findings of the research were grouped under three themes (planning the teaching, teaching, assessing the learning outcomes) and 10 sub-themes; also, common categories were seen to form under the sub-themes.

3.1. Planning the Teaching

This theme includes the teacher's plans for the entire process before learning begins. Teach-er responses resulted in the creation of four sub-themes. Figure 1 shows sub-themes and categories in the theme of planning the teaching.

Figure 1. *Sub-themes and categories in the theme of planning the teaching*



3.1.1. *Reviewing the curriculum*

The first sub-theme gave information about how teachers prepare for the lesson before they start to teach it. The teachers stated in the interviews that they reviewed the syllabus and learning outcomes before starting the lesson or the learning process. The teachers' stated goals in doing this are "remembering the student's performance, planning activities around classroom facilities, scheduling time, remembering the topic content, planning activities suitable for the student level, and preparing a daily lesson plan."

The participants explained how they started would begin the lesson by examining the curriculum and said their basic aim was to plan in-class activities. The teachers also stated that they reviewed the curriculum and learning outcomes to determine the topic to be covered in the lesson. The participants express their views as follows:

> The curriculum changes all the time depending on the grade level. For 6th Graders, there are a lot of changes between the subjects and outcomes I taught them last year and what I'm teaching them this year. One topic can have three learning outcomes in one grade, then none at all in the next. (P16)
>
> I review the curriculum at the start of each week and decide what to do in each lesson and what activities to do. Some weeks, I do special activities and use materials, so I look at the outcomes every week. (P6)
>
> I review the curriculum and its outcomes to see which outcomes are different in which unit, to plan how I'm going to teach, plan our activities, and make initial preparations by realizing what is different in that unit. (P3)

The teachers said they do not do this for every grade level or at the start of each unit/learning process; nor do they feel the need to review the outcomes, given their experience. They did say that they review the curriculum "to remind themselves" of the outcomes if they are going to teach in a class that is new to them. One participant said:

> Some themes have specific outcomes. I need to see which theme has different outcomes and plan my teaching accordingly. Beginner teachers always have their lesson plans with them, but because we are a bit more experienced, we know what lessons have what outcomes. We don't need to keep a constant track of them. (P4)

3.1.2. *Determining learning readiness*

The teachers said they use what they know about the students from previous terms or at the beginning of the learning process to determine learning readiness before starting the lesson. In addition, teachers who said they know the class and the student well stated that they do not do any activities to determine readiness. Two categories emerged in line with the answers of the teachers who do activities related to determining readiness: The reason why the teacher determines readiness and the method he uses to do this. The teachers' main objectives in determining readiness are planning the teaching process and motivating the student about the lesson. Here are some teachers' opinions:

> At the beginning of each semester, I give my students an achievement test. This lets me get to know the student. If I already know the students, I don't really need to do this. I decide what kind of activities to do. (P9)
>
> I check the students' readiness before each lesson. We plan what to give and how much to give in the lesson. If the student's achievement level is low, I start the lesson at a lower level. If the child is successful, I start the lesson with an outcome that measures higher-level processes. (P16)

One teacher who thought that determining readiness played an important role in motivating the student before the lesson begins made the following remarks:

> I usually check learning readiness at the beginning of the year. I see what level the kids are on. For example, kids come not knowing much about multiplication tables. If this skill is not learned enough, we are going to have trouble solving other problems, and this can be demotivating. (P2)

The teachers used different ways to determine learning readiness levels. They assessed it by using readiness tests, asking short questions based on previous learning, assigning homework, or examining test scores from previous terms. Some teachers said that they also create small spaces to discuss the concepts in the unit to be taught. Here are the comments made by two teachers on this topic:

> I hold a readiness test at the start of the year. For 5th-graders, I do a test on what they learned in the 4th grade. I started doing this as I became more aware of the students. (P7)
>
> In the first five minutes of the lesson, I try to find out whether the student is competent enough to learn the topic. By asking questions. This is mostly a question/answer session. If the answers are incomplete, I try to complete them. I check whether they are lacking anything with respect to the previous topics. (P10)

3.1.3. *Introducing the unit*

The participants said they brief the students on the unit or the outcome before the learning begins. For example, telling them the topic titles, mentioning the content of the topic, associating the unit with daily life, explaining the relationship with the previous unit, explaining the activities to be done during the lesson, and having the student look over the content of the unit. Here are three teachers' comments:

> Just before starting the unit, I draw their attention to the topic headings in the list of contents. First, I introduce the general outline. This is important because if I don't, they can drop out of class, I tell them what they are going to learn, and I give them examples from daily life. (P12)
>
> We need to do some work every time we switch to a new unit. We may have to do some groundwork or some preliminary research. Apart from that, let's say he will acquire different skills, for example, writing skills, so I first explain what we are going to do, what our goal is, and step-by-step how we are going to do this in the lesson. (P1)
>
> I associate it with something from our daily lives so the student can picture it in his mind. I do this using a question/answer technique. I try to introduce the topic by using the images in

> the book. I also tell them what information they're going to get. I don't explain the outcome in detail. (P15)

3.1.4. *Methods and techniques of teaching*

The last sub-theme determined under the theme of Planning the Teaching was teaching methods and techniques. It was also determined that when planning the learning process, teachers made preparations for choosing the teaching methods and techniques they are going to use. This process is divided into two categories, namely, "the method they use" and "the criteria for choosing this method." The teachers said the preferred to use the "direct instruction, group work, discovery, experiment-observation, creative drama, and case study" methods in the classroom.

> Unfortunately, we utilize straightforward instruction. But, what I want him to do is learn by doing and living. Alas, there are many methods we can't use in the classroom. Sometimes, I teach using a smart board. We can also do group work. (P5)
>
> If I am going to teach grammar, I usually teach the lesson, meaning I give them the lesson. Sometimes, if we are going to do writing activities, they can be active. (P6)
>
> If necessary, I use direct instruction and learning by discovery, depending on the diversity of the topic. Sometimes, we use methods that allow them to learn interactively with each other. We use the weighted expression technique and the question/answer technique. (P10)

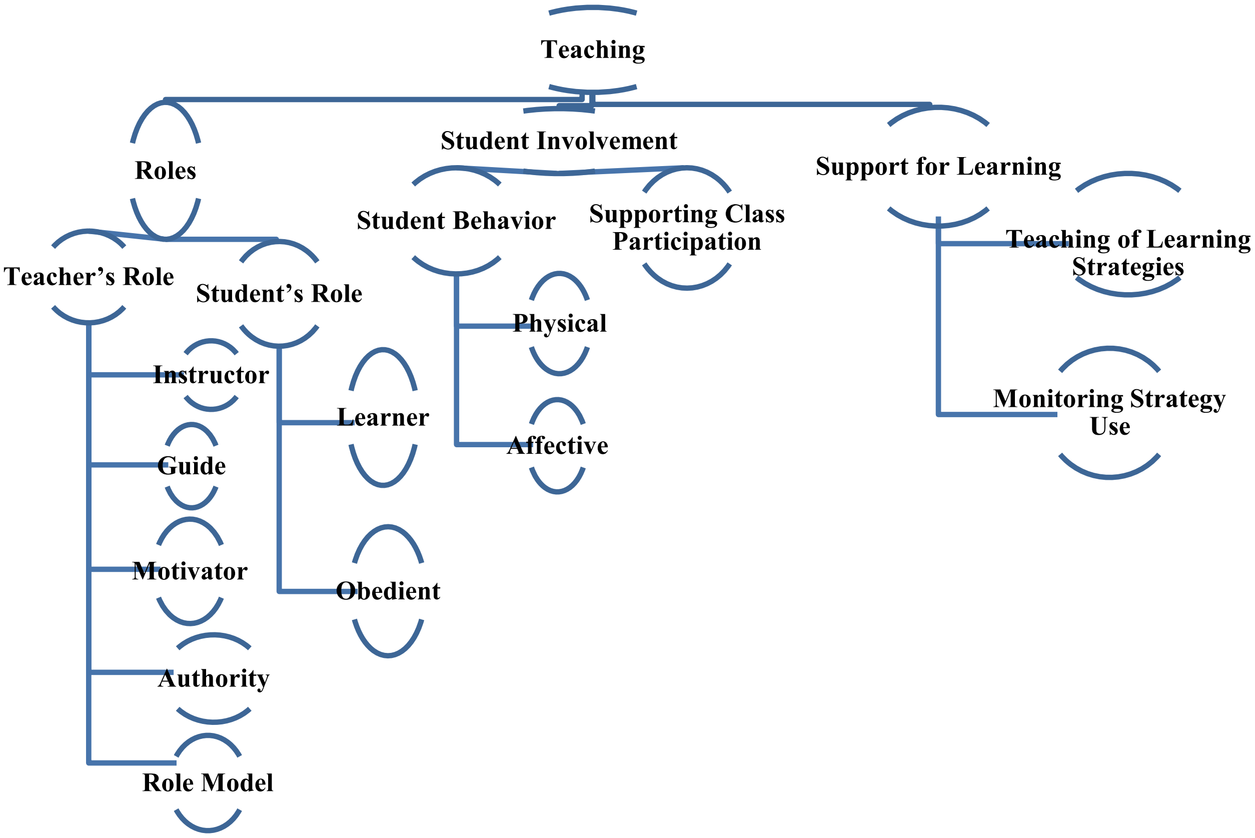
As can be understood from the comments above, although the teachers tended to use different teaching methods and techniques, they preferred the direct instruction method in which the teacher is active. They paid particular attention to "the content of the topic, the learning outcomes, the class's physical characteristics, the time allocated for the relevant unit in the syllabus, and student readiness" when choosing which teaching method to use. Here are three teachers' comments:

> I decide which method of instruction I'm going to use depending on the outcome or based on a daily plan. Sometimes it may not be appropriate for the level of the class, but I usually decide which teaching method and technique I'm going to use based on the learning outcome. (P13)
>
> I determine the method I will choose depending on the topic. If the topic appeals to more than one sense, the method I choose, such as writing, reading, and grammar also changes. (P14)
>
> The means offered by the classroom and the school influence how I determine student readiness. (P11)

3.2. Teaching

The second theme obtained from the participants' responses was related to the teaching. This theme had three sub-themes, namely, roles in the class, student involvement in the class, and support for learning. The sub-themes were further subdivided into six categories. Figure 2 shows the themes, sub-themes, and categories obtained from the opinions of the participants.

Figure 2. *Sub-themes and categories in the teaching theme*



3.2.1. *Roles*

The participants were asked about the roles necessary for successfully concluding the teaching. The answers were grouped under two categories: teacher's role and student's role. The teacher defined himself as "instructor, guide, authority, motivator, and role model" in the teaching process. Participants defined their *instructor* role as "transferring information, giving feedback, showing learning paths, and solving questions." Some participant comments:

> Usually, I'm the center of the class, and I'm the one who talks. This situation varies according to the topic; for example, I teach the lesson about mixtures first and then I let them work. They experiment with the materials they bring, and I stay in the background. (P12)
>
> I specifically tell them what the learning outcomes are when the lesson begins, and then we share roles. I ask them questions and let them talk more, and I observe them. (P6)
>
> I'm not supposed to direct; I'm more of a narrator. (P1)

Another role that the participants talked about was being a *guide*. They described this role as "guiding or steering the students, observing them, and intervening when necessary." This guiding role is explicitly mentioned in one participant's comments.

> As a teacher, I don't want to give students ready-made information only. I want to guide them. Of course, this is a difficult path; it's hard for me and the students. In the sense of learning, I want to teach them how to learn, how to study; I want to steer them. I don't want to give them a fish. I want them to learn how to fish. (P5)

The teacher set himself the role of the person who makes students like the lesson and who arouses curiosity in them. One teacher described the role of motivator as follows:

> Our first goal in education and instruction is to arouse the child's sense of curiosity in the classroom. Children do not always come to school with the same enthusiasm and excitement. I always try to light that initial spark of curiosity by asking questions or using activities. (P10)

Teachers emphasized the importance of being a leader in the lesson and exercising control so that learning could continue at a regular pace and defined themselves as the person who is in charge in the classroom. They defined the role of *authority* by using such concepts as "control center, class manager, directive giver." Teachers' remarks:

> You have to be the class's control center. Sometimes, when we show flexibility in managing the class, learning proceeds differently. This is why I try to be in control and have complete dominance over the class. I take charge at the start of the lesson. (P9)
>
> I want to be a guide, but I also like being a leader. Being in charge. I want to be an authority that the students respect and like. (P8)

From the teachers' comments, we can see that they emphasized the need to be a role model and used descriptions that highlighted the "role model" role.

> We must set an example both in society and in class. We try to manage our behavior from entering the classroom to leaving it. We try to set an example by writing something on the board or in our discourses. Right down to the clothes we wear. (P16)

The teachers also had views on what the student's role should be in the teaching. In their responses, the teachers grouped the role of the student into two categories, namely, "*learner* and *obeyer*." As a *learner*, the student should come to the lesson prepared, repeat what he has learned, participate in the lesson (follow the lesson, ask questions, join in activities, etc.), listen to the lesson, strive to learn, and be willing. The teachers gave the following examples in their comments:

> The student should come to the lesson prepared and curious. If he has a goal, he should follow the lesson as much as possible. he should have expectations at the beginning of the lesson. (P14)
>
> When I talk about the topic, the students are listeners only. But, when solving a problem, the responsibility lies entirely with them. They should ask when they get stuck. That's when I step in. (P12)
>
> As far as they are concerned, I am the one who possesses the knowledge and I try to present this knowledge to them. I want them to consult me, but it is debatable just how successful I am here. When explaining grammar, I am the only one talking. It can vary depending on the lesson, but I am usually the narrator. (P1)

The roles of *learner* and *obeyer* were usually mentioned together in the teachers' comments. The student in the role of *obeyer* should fulfill the assigned tasks, follow the rules and instructions, obey the teacher, and act in line with society's expectations. The following remarks support this role:

> The class has specific rules. What I pay attention to is the student doing what I want. As long as the student follows those rules, he can have freedom in the classroom. For example, he should do his homework, respect his friend, and bring the materials I want. (P12)
>
> I can be aggressive when they don't do the activities and homework that I set. The student should both obey the teacher and better himself in some way. (P10)
>
> My goal is that the student should be a good person first and then good at mathematics, a person who isn't unfair to others, who is honest, and who loves his homeland. I want him to like the lesson first. Success comes later. (P2)

3.2.2. *Student involvement*

Another sub-theme related to the teaching was student participation. In their explanations, teachers emphasized that student participation in class is a crucial part of learning and teaching. This sub-theme covered student behavior with respect to class participation and teacher behavior in class to increase student participation.

It was seen from the teachers' comments that they differed in what they considered to be class participation. While some teachers treated class participation in terms of the students' *physical* *(obvious) behavior*, some emphasized *affective behavior*. Those teachers who considered student involvement in terms of their obvious behaviors stated that any student who spoke in the lesson, asked questions, did homework, took notes, participated in class activities, came prepared, listened to the lesson, and had high exam scores actively involved in the lesson. Here are comments by two teachers:

> Any student who applies what he has learned when asking a question or solving a problem on the board is a student who is actively participating in the lesson. Students who join in activities willingly in group work are those who participate in the learning process. (P8)
>
> The student comes into play both when the topic is being introduced and when it is being reinforced. I direct them to problem-solving and ask them to prepare some materials. Any student who stands up and solves problems on the board is participating in the lesson. (P3)

The teachers who associated lesson participation with affective behaviors thought that students who care about the lesson and show interest are actively participating in the lesson. One teacher explained:

> For example, we have been doing distance learning and holding live lessons online, but student turnout is low. I mean, there are supposed to be 27-28 students in the class, but only six or seven are online. I'm talking about their interest in the class and their anxiety. Active participation does not necessarily mean raising your hand or speaking up. I do need to see some commitment on the part of the student. he should care about the lesson. (P10)

The teachers said that they have a motivating role to play to increase the level of student participation. However, teacher behaviors also differed according to the degree of student involvement in class. For example, if there was a student who never participated in the lesson and insisted on this, the teacher would meet with the student one-on-one or direct the student to the school's counselor. Another way might be to contact the student's parents. One teacher explained:

> Some students never participate in class. I invite these students to join me in turn and we talk during recess. I talk to them once, then once again, but I won't push it if their behavior doesn't change. (P10)

The teachers stated that in classes where the level of class participation is moderate to high, when the motivation of the class decreases, or the students become distracted, they do activities that will attract students' interest to increase their class participation (giving examples based on daily life, giving awards, asking interesting questions, playing games, role-play, group work, grade threats) or they give them tasks in the classroom that they can be active in.

> I give my students reinforcements to keep them interested in the lesson. Well done, I say. I do mock exams once in a while, and I buy gifts for the top five in these exams. I sometimes give them stickers that say well done. They work hard to win one. They motivate the students. (P14)
>
> I try to increase student involvement by putting additional questions to the students. However, this situation changes for the 8th graders; those who answer the additional questions are the ones who already participate in the lesson, and I may have to threaten others with grades. (P12)

The teachers also stated that participation varied according to grade level. Participation in the 5th, 6th, and 7th grades is high, but participation in the 8th grade decreases. They explained the main reasons for this as the pressure and anxiety created by the exam to start secondary education and the distance learning during the COVID-19 pandemic. They highlighted student fatigue.

> Participation varies from grade to grade and the pandemic has created a gap between the 7th and 8th grades. The 8th-graders are very tired. They think they should go when the lesson ends, while the 7th graders compete among themselves and motivate each other. (P6)

3.2.3. *Supporting learning*

The last sub-theme in the teaching theme was supporting learning. The researchers tried to determine what the teacher did to ensure that the student was an independent learner. Under this sub-theme, the teachers' explanations were grouped into two categories: *teaching of learning strategies, and monitoring/checking the use of these strategies*.

The teachers in this study emphasized the teaching strategies they used in the classroom more than learning strategies. The teachers' remarks showed that they did not teach their students the learning strategies mentioned in the literature and did not create an opportunity to use them. Teachers intuitively assumed that students could determine the most appropriate strategy for themselves from among different learning strategies. However, they did not know which learning strategies students are aware of or use. The teachers reported this as:

> It is very difficult to determine which strategies individual students use; some students are auditory learners and others are visual learners. Actually, we do this without realizing it. Take explaining grammar, for example. We emphasize some important points. We write these points on the board and ask them to take notes, so what we want is for them to see it on the board. In addition, we use smartboards to get them to do topic-related activities. (P5)

> I use coding a lot when I teach the lesson, I make analogies and give real-life examples. For example, when I ask what is observed when light passes from a very dense environment to a less dense environment, they cannot answer the question, but when I ask what a vehicle does when it moves from dense traffic to light traffic, they answer, and I tell them that light does the same. (P12)

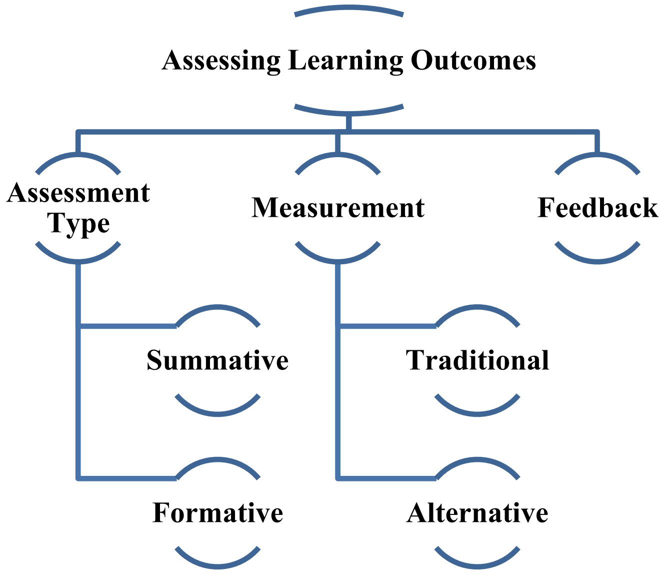
The teachers stated that although they did not do any activities to teach learning strategies, they did monitor whether students used appropriate strategies for themselves through individual observations. One teacher said that if the student gives correct answers to the questions asked, succeeds in the test exams, can do his homework, participates in classroom activities, and can self-evaluate, this means that he can choose and use the appropriate learning strategy.

> We diversify learning strategies to help students learn by using different strategies. We are increasing participation, as well. For example, if the student can apply what he has learned in the lesson alone; what we mean by "apply" is can he make similar examples or solve a question correctly? I use my personal observations to assess this. (P13)

3.3. Assessing Learning Outcomes

The last theme obtained from the teacher interviews was assessment of learning outcomes. This theme covered information relating to the process and assessment of the outcomes. The purpose of the questions put to the teachers was to determine what route they followed to support student self-learning. Three sub-themes were found under this theme: what route the teacher follows to assess learning outcomes (assessment type), the measurement tools and methods used to make this assessment, and feedback. Figure 3 shows the sub-themes and categories under the theme of assessing learning outcomes.

Figure 3. *Sub-themes and categories under the theme of assessing learning outcomes*



3.3.1. *Supporting learning*

The teachers stated that they assessed student performance at the beginning of the academic year, during the semester, or at the end of the semester. The responses were thus grouped into two categories, namely, formative and summative assessment. For formative assessment, they carried out activities aimed at "identifying and repeating what is missing in learning outcomes, receiving feedback from students about the teaching process, following students' individual development, identifying misconceptions, doing activities to reinforce what has been learned, assessing the difficulty of questions, and planning the lesson around student outcomes." In addition, the teachers added that formative evaluations also provide information about their own teaching practices. Here are examples of what the teachers said:

> I test the students at the end of each unit. I want to see where the student is lacking. I do this to repeat the topic where they are missing something. I see what they do and don't understand. This isn't for grades! (P5)
>
> There can be plenty of misconceptions in my class. I want to identify them and find out about the student's performance. Has the student's performance improved? Can I move on to the next unit? I tell him how he's doing. (P10)
>
> I hold a quiz at the end of each unit. I'm doing this for my benefit. I need to see how much the students have learned. If they don't achieve the learning outcomes, I don't move on to the next topic. I don't tell the students if they're doing very badly. (P4)

In formative assessment, they tended to "grade the student, assess student performance based on exam results, and check to see if the learning outcomes have been achieved."

> I quiz them every week, give them a test, and tell them what they are doing right and what they are doing wrong. This kind of assessment does not show the teacher what they have and havent achieved in terms of learning outcomes. It only tells the teacher if the students have learned the topic or not. (P14)
>
> I assess and grade the students based on their exam results at the end of the semester, their participation in the class, and the materials they made. (P8)

3.3.2. *Measurement*

The teachers utilized both traditional and alternative measurement methods. Pen and paper achievement tests, assignments, oral examinations, and opinion scores (observation of classroom student behaviors) are examples of traditional measurement methods.

> I mostly use multiple-choice tests and short-answer questions in the classroom. Observation is the most useful resource I have when it comes to students. I examine the students' notebooks and I look at their in-class speaking skills. I sometimes ask open-ended questions. P(8)
>
> We do tests to determine student performance. I use open-ended questions in classrooms for 5th grades, but I use multiple-choice questions to prepare 7th- and 8th-graders for the high school entrance exam. (12)
>
> I give oral exams to the students to assess what we do in the lesson. Sometimes I get a piece of paper out and ask questions, and I can tell from their answers what they can and can't do. P(15)

From the general comments of the participants, it was understood that many of them frequently use achievement tests consisting of multiple-choice items. They used sometimes alternative measurement methods such as peer assessment, performance assessment, and performance tasks/projects less. These were usually carried out in conjunction with group work.

> When I talk about percentages, I form groups of four. It may not be a fully detailed peer review, but one student evaluates the other in terms of his activities. (P2)
>
> I give the students questions, which they then solve. I then give them the solutions, and they or their peers check the answers. They mark them up or down. They help me with the scoring. (P13)
>
> Sometimes I give performance tasks, although not in every unit. I give performance tasks in the middle of the unit and collect student products at the end of the unit. (P7)

3.3.3. *Feedback*

The practices teachers used included telling students how many of the questions were answered correctly and incorrectly, explaining exam scores, telling them what was missing in their homework, and congratulating them when they succeed (well done, very good, applause, etc.). Here are some teacher comments:

> I call the student over after the test and tell him you made a mistake here. I tell him which type of question he makes the most mistakes in. I don't make a different activity for this. (P13)
>
> I tell him that he can do better if he wants to and is more careful. I congratulate students who excel in the test and tell them well done. (P10)
>
> I give students the answer keys for their homework so they can check for themselves what they got right and wrong. Sometimes, when they give answers, we weren't expecting, I get their classmates to applaud them. I give them plus and minus scores, but it's not that effective. (P12)

Another remarkable finding in the teachers' responses about feedback was that not every student receives feedback relating to access to learning outcomes and that those who ask for such feedback (e.g., students who reject the exam results and want their answer paper rechecked) or successful students are given feedback by telling them the number of right and wrong answers in the exam.

> We have weekly multiple-choice tests. Tests with 10 questions about the gains learned that week. After students answer, I tell them how many right or wrong they did. (P15)
>
> I administer tests to students in the middle and end of the semester. Some students object to their test results and want to see where they went wrong. I show these students their exam papers. (P9)

4. DISCUSSION and CONCLUSION

This study examined teacher reports to determine how AsL pertains to classroom practices taking into account its conceptual framework. Case study design, a qualitative research method, was used to collect detailed information about classroom teacher practices. The teachers were interviewed using semi-structured interview forms and the data obtained were then analyzed using content analysis. Teacher responses were discussed under the themes of planning the teaching, activities, and measurement-feedback. The results obtained from these themes and the discussions on them are given under the relevant headings.

4.1. Planning the Teaching

The teachers made some preparations before teaching the lesson. These preparations included examining the outcomes in the syllabus, determining the level of student readiness, introducing the unit to the students, and deciding on the teaching techniques to be used. They used achievement tests or short question-answer activities to determine the students' degree of readiness before they start learning. Teachers should make initial assessments to determine what their learning needs are and it is clear that this practice contributes to AsL. Remarkably, some teachers were seen not to make this planning and associated not doing so with their experience.

What stood out in the teacher responses was that the mechanical way in which they informed the students about what was to be learned. The teachers mainly told them about the topics in the unit in question or its scope. The teachers did not provide sufficient information about what learning outcomes they expected their students to achieve by the end of the lesson. In other words, the teachers provided their students with content-oriented information about the subject/unit but did not tell them about the thought processes involved or the outcomes. As a result, the students began the lesson not knowing what was expected of them or what standards/criteria they were expected to meet if they were to pass. When considered in the context of AsL, students need to know the answer to the question, "Where are we going?" to be able to regulate their learning. Simply giving students a mechanical overview is not going to be enough to activate students' self-regulation behaviors within the scope of AsL.

Another aspect of planning the learning process is determining which teaching methods and techniques to use. Most of the participating teachers stated they preferred direct instruction. Teachers can find themselves with students having different levels of knowledge and can show them how to build on their current achievement levels. In this respect, teacher assessment practices should include innovative and efficient teaching, monitoring, and scaffolding activities and should take into account differences between students (Schellekens et al., 2021). The teacher responses did not reflect this point of view, however. The reason for this was seen to be closely related to the scheduling set out in the syllabus. It was understood from the teacher responses that they felt under pressure to complete the units/topics on time. This finding is similar to the findings of Akıncı et al., (2015) and Balbağ and Karaer (2017) studies, which found that the lack of time related to the implementation of the curricula is a problem. Furthermore, most of the teachers stated that the content of the topic was a key factor when deciding what teaching techniques to use. The teachers' answers did not reveal their thought processes or reasoning for the methods and techniques that would allow the students to play an active role in the learning process. Fenwick (2017) emphasized the incompatibility between the planned curriculum and classroom-level active assessment practices.

4.2. Teaching

The researchers obtained the participants' answers concerning the roles of teachers and students in the learning process. Most of the teachers defined their roles in keeping with the behaviorist approach. In other words, the roles mentioned the most were "the authority figure who manages the class" and "the one who teaches." Correspondingly, the role of the student was confined to "learner and obeyer." This finding seems to be smilar with other research findings (Thompson et al., 2017; Schellekens et al., 2021) revealing that learning in practice still depends on the teacher.

Although the teachers stated that the student should be at the center in the learning process, they also said that they adopted roles in which the student was less active and the teacher was the instructor at the center of the class because of overcrowded classrooms, the packed syllabus, and the lack of resources and amenities at the school. The teachers' remarks do not seem consistent with AsL's approach to creating opportunities and environments that support learner autonomy. This is because the "self" is a key point of focus in AsL-related activities and learning experiences are structured on the "self."

The researchers asked the teachers what they thought about active participation, considering their responses to the role of the student in the learning process, and the majority of them defined active participation as observable student behavior. In other words, teachers thought that students who take the floor in the lesson, perform the tasks given by the teacher, and listen to the lesson were actively participating in class. The teachers' answers here seem to be consistent with the role of the student. AsL requires the student to be active in the learning process. In classroom practices, active participation occurs when there are activities that enable students to work on their self-assessment skills and use them (such as self-peer assessment) (Schellekens et al., 2021). In this case, the student is expected to take responsibility for directing their learning. Activities where the student can plan, monitor, and assess their own learning will support active participation. However, the teachers' responses to active participation seem to be a long way from activating "self" structures and true active participation. Some studies said most students reported that they participated very little in such assessment activities or not at all (DeLuca et al., 2018; Leirhaug & Annerstedt, 2016). Yet, a study has reported students having positive attitudes toward activities involving active participation (Thompson, 2017).

The teachers stated they carried out activities to ensure and maintain student participation in class. This is especially important in the context of AsL because participatory behaviors and motivation are necessary if students are to be self-regulated (Pintrich, 1999). In this case, it becomes difficult for students who are not academically ready and motivated for the lesson to manage their own learning processes. Remarkably, the teacher responses showed that teachers resort to in-class, context-independent methods such as silence, making jokes, or talking about extracurricular topics to ensure or maintain student motivation. Many strategies can be used to keep student motivation alive (for example, self-consequences, self-verbalization, game learning). These strategies make it easier for the student to manage his learning process, and result in the student developing a sense of being important or useful with respect to content or materials (Wolters, 2003).

The teachers' responses regarding the use of learning strategies and how accurately and effectively they are used showed that practices concerning in-class learning strategies are incomplete or wrong. The teachers said they do not do any activities relating to teaching and monitoring learning strategies or giving feedback to the student throughout the learning process. The reason for this is again understood from the teacher's responses. Remarkably, most of the participant teachers showed conceptual deficiencies or errors in their responses about learning strategies. From their answers, it was clear that the practices they adopted thinking they were learning strategies were teaching techniques. Some teachers stated that they adopted problem-solving (mostly multiple-choice) or repetition of the topic as a learning strategy. In this case, it is naturally difficult for teachers who do not have theoretical knowledge about learning strategies to teach these strategies to students and enable students to use them in different contexts. Other findings have shown that as a consequence of teachers' shortcomings here, students in various grades use basic strategies such as summarizing and making outlines more often than regulatory strategies (Garcia-Perez et al., 2021; Rovers et al., 2018).

4.3. Measurement and Feedback

The teachers stated that they most often used achievement tests to assess learning outcomes. These measurement tools were mostly used for summative purposes at the end of the learning process, and particularly to manage teaching when the process was under way. Performance tasks and longer-term tasks such as projects were used less frequently than achievement tests.

The multiple-choice item format is widely used in both in-class and high-stakes testing. The participating teachers' responses saying that they frequently used this item format in their classroom exams support this. Similar findings are also found in other studies (e.g. Gelbal & Kelecioğlu, 2007; Karatay & Dilekçi, 2019). As other researchers have pointed out, multiple-choice items can strengthen students' short-term memory, but not foster critical thinking skills (Credé & Phillips 2011; Rovers et al., 2018). This shows why assessment activities matter. Assessment activities are known to have a strong bearing on learning approaches (Panadero et al., 2019). Other research findings showed that innovative assessment practices that support student learning are not regularly applied in the classroom (Hawe & Parr, 2014; Marshall & Drummond, 2006) and teachers are more committed to traditional approaches by focusing on their test scores (Hawe & Parr, 2014). In addition, Tan (2013) suggested that practical assessment applications are for the improving of short-term learning.

The teachers' responses revealed that classroom practices made little use of the feedback mechanism. While the learning process is under way, teachers who use measurement tools for formative purposes primarily use their results to check the effectiveness of their teaching. Their students received very little feedback regarding learning outcomes or student studies, and the feedback that was given was very superficial. Teacher feedback at the end of the learning process was largely limited to the number of right and wrong answers in the exams. Yet, the formative feedback given by the teacher is vital if students are to carry out and manage the learning process correctly. The purpose of formative feedback is to provide the person with the power to supervise and direct their own learning so that the person can be a more determined, responsible, and effective learner (Black & Jones, 2006). This explanation reveals the relationship between formative feedback and self-regulated learning. Butler and Winne (1995) stated that feedback is a natural catalyst for all self-regulated activities to support this. In this case, feedback such as informing the student about the number of right and wrong answers in the exam, telling him to "work harder" or "revise and recalculate your answer" will not help them become self-regulated learners because this does not strategically guide the student on how or why they should do this. The results of many meta-analysis studies revealed that formative feedback is effective for supporting students' high-level skills and deep learning (Hattie & Timperley, 2007; Swart et al., 2019). The responses of the teachers in this study showed that the feedback process takes place from teacher to student and there was no interaction between teacher and student. In the study conducted by Hargreaves (2014), interviews were conducted with teachers and similar answers were obtained emphasizing that teachers are active regarding the functioning of the feedback mechanism.

The teachers' responses showed that the students did not carry out activities to evaluate their own performance or the performance of their peers. The answers that stand out here reveal that the teachers did not trust the students when it came to assessment. In other words, the teachers did not believe that students could assess their own performance or that of their peers accurately and fairly, which is why they chose not to use self- and peer-assessment in class. In addition, this finding was not surprising considering the responses of the participating teachers that they mostly adopt approaches focused on test scores in their classroom practices.

Yet, a series of studies demonstrated good reliability and validity of peer assessments on average (Li et al., 2016; Liu & Ji, 2018). On the other hand, other studies support the teachers' concerns about self- and peer-assessment (e.g. Kovach et al., 2009; Ward et al., 2002). However, the reliability of self- and peer-assessment can be improved by increasing the assessors' understanding of content, quality and standards, assessment criteria, training, and means of self-

and peer-assessment (Sung et al., 2005). Nevertheless, despite these concerns, other meta-analysis studies have demonstrated the positive effect of formative self-assessment on self-regulated learning (Andrade, 2019; Panadero et al., 2017). Similarly, peer assessment is known to support autonomous learner characteristics (Bloxham & West 2004).

4.4. Limitations and Suggestions

This study does have some limitations. First, we met the teachers once only. It would be interesting to conduct follow-up interviews and observations, especially with teachers who practice ASL-based activities in the classroom, to collect more reflexive data regarding the process. Second, we collected online data based on solely teacher reports in the context of AsL. Future qualitative studies can collect and analyze data that reflect a more detailed process, such as in-class observation and interviews with students. Third, we did not limit our interviews in this study to any particular task. To provide a better perspective for AsL, teacher behaviors can be examined in learning tasks that require high-level skills.

This study presents some theoretical and practical implications for teachers, policy makers, and researchers concerning in-class AsL. As understood from our discussions, the teachers' AsL activities were highly superficial and seemed far from supporting learner autonomy. In addition, national high-stake testing, in particular, closely influenced what assessment activities teachers choose to conduct. AsL should be reflected in national-level curricula and activities rather than simply on a classroom scale and adopted as policy because teachers cannot be expected to adopt AsL conceptually and apply it in the classroom without knowing what it is. By adopting a political approach at the national level, teachers' professional development or the content of teacher training could be organized to accommodate AsL.

If teachers are to design a learner-centered learning process, they should acquire skills that will allow them to teach learning strategies. In particular, teachers should be helped academically in teaching deep learning strategies and designing assessment activities in support of this.

For AsL, teachers must use feedback effectively throughout the learning process. Examples of formative feedback and practices can be made available to the teacher through teacher education and digital content.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Pamukkale University/Institute of Educational Sciences, E-93803232-622.02-193607.

Authorship Contribution Statement

Ozen Yildirim and **Safiye Bilican Demir** performed the same contribution for all the processes of the research from the beginning to the end.

Orcid

Ozen Yildirim https://orcid.org/0000-0003-2098-285X
Safiye Bilican Demir https://orcid.org/0000-0001-9564-9029

REFERENCES

Akıncı, B., Uzun, N., & Kışoğlu, M. (2015). Fen bilimleri öğretmenlerinin meslekte karşılaştı kları problemler ve fen öğretiminde yaşadıkları zorluklar [The problems experienced by science teachers in their profession and difficulties they are confronted with in science teaching]. *International Journal of Human Sciences*, *12* (1), 1189-1215. https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/3188

Andrade, H.L. (2019). A Critical review of research on student self-assessment. *Frontiers in Education*, *4*(87). https://doi.org/10.3389/feduc.2019.00087

Andrade, H., & Brookhart, S.M. (2016). The role of classroom assessment in supporting self-regulated learning. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 293–309). Springer.

Assessment Reform Group (1999). *Assessment for learning: Beyond the black box.* University of Cambridge School of Education.

Assessment Reform Group (2002). Assessment for Learning: 10 principles. http://www.assessment-reform-group.org/CIE3.PDF.

Baird, J.A., Andrich, D., Hopfenbeck, T.N., & Stobart, G. (2017). Assessment and learning: Fields apart?. *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317-350. https://doi.org/10.1080/0969594X.2017.1319337

Balbağ, M.Z., & Karaer, G. (2017). Sınıf öğretmenlerinin fen öğretiminde karşılaştıkları sorunlar [The problems of primary school teacher faced in the science teaching process]. *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, *8*(1), 28-46. https://doi.org/10.24315/trkefd.364015

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W H Freeman/Times Books/ Henry Holt & Co.

Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25. https://doi.org/10.1080/0969594X.2010.513678

Berg, B.L., & Howard, L. (2016). *Qualitative research methods for the social sciences* (8th ed.). Pearson.

Black, P., & Jones, J. (2006). Formative assessment and the learning and teaching of MFL: sharing the language learning road map with the learners. *Language Learning Journal*, *34*(1), 4-9. https://doi.org/10.1080/09571730685200171

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Open University Press.

Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning how to learn and assess ment for learning. *Research Papers in Education*, *21*(2), 119-132. https://doi.org/10.1080/02671520600615612

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81–100). SAGE Publications, Inc.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Bloom, B. (1974). An introduction to mastery learning theory. In J. Block (Ed.) *Schools, society and mastery learning.* Holt, Rinehart & Winston, Inc.

Bloxham, S., & West, A. (2004). Understanding the rules of the game: making peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education*, *29* (6),721-733. https://doi.org/10.1080/0260293042000227254

Brookhart, S.M. (2018). Summative and formative feedback. In A. Lipnevich & J. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 52–78). Cambridge University Press.

Brown, G.T.L., & Harris, L.R. (2013). Student self-assessment. In J. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367-393). SAGE Publications, Inc.

Boud, D., & N. Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31* (4),399-413. https://doi.org/10.1080/02602930600679050

Butler, D.L., & Winne, P.H. (1995). Feedback and self-regulated learning: A theoretical synth esis. *Review of Educational Research*, *65,* 245-281. https://doi.org/10.3102/00346543065003245

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205–249. https://doi.org/10.1007/s10648-011-9191-6

Credé, M., & Phillips, L.A. (2011). A meta-analytic review of the motivated strategies for Learning Questionnaire. *Learning and Individual Differences, 21*(4), 337–346. https://doi.org/10.1016/j.lindif.2011.03.002

Creswell, J.W. (2018) *Research design: Qualitative, quantitative, and mixed methods aproach es* (4th Edition). SAGE Publications, Inc.

Creswell, J. W., & Miller, D.L. (2000). Determining validity in qualitative inquiry. *Theory into practice, 39*(3), 124-130. https://doi.org/10.1207/s15430421tip3903_2

Dann, R. (2014). Assessment as learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 149-166. https://doi.org/10.1080/0969594X.2014.898128

Davies, A., & LeMahieu, P. (2003). Assessment for learning: Reconsidering portfolios and research evidence. In M. Segers, F. Dochy, and E. Cascallar (Eds.), *In optimising new modes of assessment: In search of qualities and standards*, (pp. 141–69). Kluwer Academic.

DeLuca, C., Chapman-Chin, A., LaPointe-McEwan, D., & Klinger, D.A. (2018). Student perspectives on assessment for learning. *The Curriculum Journal, 29*(1), 77-94. https://doi.org/10.1080/09585176.2017.1401550

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231-264. https://doi.org/10.1007/s11409-008-9029-x

Dinsmore, D.L., Alexander, P.A., & Loughlin, S.M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, *20*(4), 391–409. https://doi.org/10.1007/s10648-008-9083-6

Earl, L.M. (2003). *Assessment as learning using classroom assessment to maximise student learning*. Corwin Press.

Earl, L.M. (2013). *Assessment as learning: Using classroom assessment to maximize student l earning* (2nd Edition). Corwin Press.

Fenwick, L. (2017). Promoting assessment for learning through curriculum-based performance standards: Teacher responses in the northern territory of Australia. *Curriculum Journal, 28*(1), 41–58. https://doi.org/10.1080/09585176.2016.1260486

Ferris, D., & Hedgcock, J. (2014). *Teaching L2 composition: Purpose, process, and practice* (3rd Edition). Routledge.

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grim shaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health*, *25*(10), 1229-1245. https://doi.org/10.1080/08870440903194015

García-Pérez, D., Fraile, J., & Panadero, E. (2021). Learning strategies and self-regulation in context: How higher education students approach different courses, assessments, and challenges. *European Journal of Psychology of Education*, *36*(2), 533-550. https://doi.org/10.1007/s10212-020-00488-z

Gelbal, S., & Kelecioğlu, H. (2007). Öğretmenlerin ölçme-değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar [Teachers' proficiency perceptions of about the measurement and evaluation techniques and the problems they confront]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 33, 135-145. http://efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1017-published.pdf

Graham, S., Hebert, M., & Harris, K.R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523-547. https://doi.org/10.1086/681947

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hattie, J., & Jaeger, R. (1998). Assessment and classroom learning: A deductive approach. *Assessment in Education: Principles, Policy & Practice*, *5*, 111-122. https://doi.org/10.1080/0969595980050107

Hawe, E., & Parr, J. (2014). Assessment for learning in the writing classroom: An incomplete realization. *Curriculum Journal*, *25*(2), 210-237. https://doi.org/10.1080/09585176.2013.862172

Hennink, M.M., Kaiser, B.N., & Marconi, V.C. (2017). Code saturation versus meaning saturation: how many interviews are enough?. *Qualitative Health Research*, *27*(4), 591-608. https://doi.org/10.1177/1049732316665344

Karatay, H., & Dilekçi, A. (2019). Türkçe öğretmenlerinin dil becerilerini ölçme ve değerlendirme yeterlikleri [Competencies of turkish teachers in measuring and evaluating language skills]. *Milli Eğitim Dergisi*, *48*(1), 685-716. https://dergipark.org.tr/tr/pub/milliegitim/issue/51765/674598

Kim, M., & Ryu, J. (2013). The development and implementation of a web-based formative peer assessment system for enhancing students' metacognitive awareness and performance in ill-structured tasks. *Educational Technology Research and Development. 61*(4), 549–561. https://doi.org/10.1007/s11423-012-9266-1

Kovach, R.A., Resch, D.S., & Verhulst, S.J. (2009). Peer assessment of professionalism: A five-year experience in medical clerkship. *Journal of General Internal Medicine. 24*(6), 742–746. https://doi.org/10.1007/s11606-009-0961-5

Klenowski, V. (2009) Assessment for learning revisited: An Asia-Pacific perspective. *Assessment in Education*, *16*(3), 263–268. https://doi.org/10.1080/09695940903319646

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A Review of the Evidence*. REL 2017-259. Regional Educational Laboratory Central.

Kruger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: An historical review, meta-analysis and preliminary feedback theory. *Psychological Bulletin*, *119*, 254-285. https://doi.org/10.1037/0033-2909.119.2.254

Lam, R. (2013). Formative use of summative tests: Using test preparation to promote performance and self-regulation. *The Asia-Pacific Education Researcher, 22*(1), 69–78. https://doi.org/10.1007/s40299-012-0026-0

Lam, R. (2014). Promoting self-regulated learning through portfolio assessment: Testimony and recommendations. *Assessment & Evaluation in Higher Education, 39*(6), 699–714. https://doi.org/10.1080/02602938.2013.862211

Lam, R. (2020). Investigating assessment as learning in second language writing: A qualitative research perspective. *International Journal of Qualitative Methods*, *19*, 1-10. https://doi.org/10.1177/1609406920938572

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: A systematic review. *Applied Measurement in Education*, *33*(2), 124-140. https://doi.org/10.1080/08957347.2020.1732383

Leirhaug, P.E., & Annerstedt, C. (2016). Assessing with new eyes? Assessment for learning in Norwegian physical education. *Physical Education and Sport Pedagogy, 21*(6), 616-631. https://doi.org/10.1080/17408989.2015.1095871

Li, H., Xiong, Y., Zang, X., Kornhaber, M., Lyu, Y., Chung, K., & Suen, H.K. (2016). Peer assessment in a digital age: A Meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245-264. https://doi.org/10.1080/02602938.2014.999746

Liu, L., & Ji, X. (2018). A Study on the acceptability and validity of peer scoring in Chinese university EFL writing classrooms. *Foreign Language World, 5*, 63-70. https://doi.org/10.1016/j.jslw.2006.09.004

Manuel, A.K. (2015). *The effects of immediate feedback using a student response system on math achievement of eleventh grade students* (Unpublished doctoral dissertation). Mercer University, Macon, GA.

Marshall, B., & Jane Drummond, M. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education*, *21*(02), 133-149. https://doi.org/10.1080/02671520600615638

McCaslin, M., & Hickey, D.T. (2001). Educational psychology, social constructivism, and educational practice: A case of emergent identity. *Educational Psychologist*, *36*(2), 133-140. https://doi.org/10.1207/S15326985EP3602_8

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Panadero, E., Andrade, H., & Brookhart, S.M. (2018). Fusing self-regulated learning and formative assessment: A roadmap of where we are, how we got here, and where we are going. *The Australian Educational Researcher, 45*(1), 13-31. https://doi.org/10.1007/s13384-018-0258-y

Panadero, E., Broadbent, J., Boud, D., & Lodge, J.M. (2019). Using formative assessment to influence self-and co-regulated learning: The role of evaluative judgement. *European Journal of Psychology of Education, 34*(3), 535-557. https://doi.org/10.1007/s10212-018-0407-8

Pintrich, P. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research, 31*(6), 459-470. https://doi.org/10.1016/S0883-0355(99)00015-4

Pintrich, P.R. (2000). The role of goal orientation in self-regulated learning. In M. Boekhaerts, P.R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451-502). Academic Press.

Popham, J. (1978). *Criterion-referenced measurement*. Prentice-Hall.

Rovers, S.F.E., Stalmeijer, R.E., van Merriënboer, J.J.G., Savelberg, H.H.C.M., & de Bruin, A.B.H. (2018). How and why do students use learning strategies? A mixed methods study on learning strategies and desirable difficulties with effective strategy users. *Frontiers in Psychology*, 9, 2501. https://doi.org/10.3389/fpsyg.2018.02501

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. (1998). Formative assessment: Revisiting the territory. *Assessment in Education, 5*(1), 77–84. https://doi.org/10.1080/0969595980050104

Schellekens, L.H., Bok, H.G., de Jong, L.H., van der Schaaf, M.F., Kremer, W.D., & van der Vleuten, C.P. (2021). A scoping review on the notions of assessment as learning (AaL), assessment for learning (AfL), and assessment of learning (AoL). *Studies in Educational Evaluation*, *71*. https://doi.org/10.1016/j.stueduc.2021.101094

Sung, Y.T., Lin, C.S., Lee, C.L., & Chang, K.E. (2003). Evaluating proposals for experiments: An application of web-based self-assessment and peer-assessment. *Teaching of Psychology, 30*(4), 331-334. https://doi.org/10.1207/S15328023TOP3004_06

Swaffield, S. (2011). Getting to the heart of authentic Assessment for Learning. *Assessment in Education*, *18*(4), 433–449. https://doi.org/10.1080/0969594X.2011.582838

Swart, E.K., Nielen, T.M., & Sikkema-de Jong,M.T. (2019). Supporting learning from text: A meta-analysis on the timing and content of effective feedback. *Educational Research Review*, 28, 100296. https://doi.org/10.1016/j.edurev.2019.100296

Tan, K. (2013). A framework for assessment for learning: Implications for feedback practices within and beyond the gap. *ISRN Education*, *2013*, 1-6. https://doi.org/10.1155/2013/640609

Theobald, M. (2021). Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis. *Contemporary Educational Psychology*, *66*, 1-19. https://doi.org/10.1016/j.cedpsych.2021.101976

Thompson, C.S. (2017). An Exploration of faculty involvement in and attitudes toward strategic planning in their institutions. *Educational Planning, 24*(1), 7-21. https://files.eric.ed.gov/fulltext/EJ1208234.pdf

Thompson, J., Houston, D., Dansie, K., Rayner, T., Pointon, T., Pope, S., … Grantham, H. (2017). Student & tutor consensus: A partnership in assessment for learning. *Assessment and Evaluation in Higher Education, 42*(6), 942-952. https://doi.org/10.1080/02602938.2016.1211988

Topping, K.J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research 68*(3), 249-276. https://doi.org/10.3102/00346543068003249

Topping, K.J., & Ehly, S.W. (2001). Peer assisted learning: A framework for consultation. *Journal of Educational and Psychological Consultation, 12*(2),113-132. https://doi.org/10.1207/S1532768XJEPC1202_03

Torrance, H. (1991). Records of achievement and formative assessment: some complexities of practice, in: R. Stake (Ed.) *Advances in program evaluation: Using assessment policy to reform education* (pp. 231-245). JAI Press.

Torrance, H. (2012). Formative assessment at the crossroads: Conformative, deformative and transformative assessment. *Oxford Review of Education*, *38*, 323-342. https://doi.org/10.1080/03054985.2012.689693

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice, 14* (3), 281–94. https://doi.org/10.1080/09695940701591867

van Gennip, N.A.E., Segers, M.S.R., & Tillema, H.H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review 4*(1), 41–54. https://doi.org/10.1016/j.edurev.2008.11.002

Vygotsky, L.S. (1978). Interaction between learning and development. In M. Cole, V. JohnSteiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes* (pp. 79–91). Harvard University Press.

Vygotsky, L. (1986). *Thought and language*. MIT Press.

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: current state of the art. *Advances in Health Sciences Education*, *7*(1), 63-80. https://doi.org/10.1023/A:1014585522084

Wiliam, D. (2018). Feedback: At the heart of – But definitely not all of – Formative assessment. In A. Lipnevich & J. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 3–28). Cambridge University Press. https://doi.org/10.1017/9781316832134.003

Weinstein, C.E., Acee, T.W., & Jung, J. (2011). Self-regulation and learning strategies. *New Directions for Teaching and Learning, 126*, 45–53. https://doi.org/10.1002/tl.443

Wolters, C.A. (2003). Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning. *Educational Psychologist, 38*(4), 189-205. https://doi.org/10.1207/S15326985EP3804_1

Wyse, D., & Torrance, H. (2009). The development and consequences of national curriculum assessment for primary education in England. *Educational Research, 51*(2), 213–238. https://doi.org/10.1080/00131880902891479

Zimmerman, B.J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P.R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–40). Academic Press.

Zimmerman, B.J., & Pons, M.M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, *23*(4), 614–628. https://doi.org/10.3102/00028312023004614

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 72–87

https://doi.org/10.21449/ijate.1124382

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Automatic story and item generation for reading comprehension assessments with transformers

Okan Bulut[1,*], Seyma Nur Yildirim-Erbasli[2]

[1]University of Alberta, Centre for Research in Applied Measurement and Evaluation, Edmonton, AB Canada
[2]Concordia University of Edmonton, Faculty of Arts, Department of Psychology, Edmonton, AB Canada

ARTICLE HISTORY

Received: June 1, 2022
Revised: Sep. 15, 2022
Accepted: Sep. 21, 2022

Keywords:
Reading comprehension,
Natural language processing,
Automatic item generation,
Language modeling,
Text generation.

Abstract: Reading comprehension is one of the essential skills for students as they make a transition from learning to read to reading to learn. Over the last decade, the increased use of digital learning materials for promoting literacy skills (e.g., oral fluency and reading comprehension) in K-12 classrooms has been a boon for teachers. However, instant access to reading materials, as well as relevant assessment tools for evaluating students' comprehension skills, remains to be a problem. Teachers must spend many hours looking for suitable materials for their students because high-quality reading materials and assessments are primarily available through commercial literacy programs and websites. This study proposes a promising solution to this problem by employing an artificial intelligence (AI) approach. We demonstrate how to use advanced language models (e.g., OpenAI's GPT-2 and Google's T5) to automatically generate reading passages and items. Our preliminary findings suggest that with additional training and fine-tuning, open-source language models could be used to support the instruction and assessment of reading comprehension skills in the classroom. For both automatic story and item generation, the language models performed reasonably; however, the outcomes of these language models still require a human evaluation and further adjustments before sharing them with students. Practical implications of the findings and future research directions are discussed.

1. INTRODUCTION

Reading comprehension is one of the essential skills that all students need to foster in K-12 education because their learning and success in other subjects (e.g., math, social studies, and history) are strongly associated with their proficiency in reading comprehension (Bigozzi et al., 2017). Reading comprehension is also the key ability that students need to master to make the transition from "learning to read" to "reading to learn" by understanding, analyzing, and applying information gathered through reading different materials (e.g., books, articles, and newspapers). Students without adequate reading comprehension skills may not be able to understand what they read and fail to make this transition.

[*]Corresponding Author: Okan Bulut ✉ bulut@ualberta.ca University of Alberta, Centre for Research in Applied Measurement and Evaluation, Edmonton, AB, Canada

e-ISSN: 2148-7456 /© **IJATE** 2022

Developing reading proficiency requires students to read more texts with varying volumes, genres, and difficulties (Allington et al., 2010; Duke et al., 2011; Kim & White, 2008). To help students develop reading comprehension skills, teachers give students various texts (e.g., fables, fairy tales, and stories) and ask them to read these texts repeatedly. Students who struggle with reading comprehension might have to practice their skills by reading more texts until they become fluent readers. Once students can read the text fluently, teachers also provide a set of items related to the text to measure students' understanding of the text. This suggests that teachers may need new reading materials and items to continuously monitor students' growth in reading. Teachers attempt to find a suitable text from the literature to meet this need efficiently. If they try to find a text from the literature, they need to go through many pieces of literature to find a suitable text, but this is a very time-consuming process. Also, it is not easy to find free reading materials because most of the materials on the Internet are commercially available.

Alternatively, the teachers may attempt to develop their own text and items associated with each text. However, writing original texts with different volumes, genres, or complexities is a highly complex task, even for a professional writer. In addition to finding a suitable text or creating an authentic text, developing high-quality items related to the text is another tedious task. Teachers must formulate high-quality items related to the text by targeting different difficulty levels and ensuring that each item is strongly associated with the text. Therefore, a more practical and sustainable solution is necessary to help teachers find suitable reading materials for their students.

1.1. Story and Item Generation

Writing and telling stories have been central to the human experience in every culture. As humans attempt to make sense of the world surrounding them, they make discoveries and learn new information. Storytelling is one of the most popular communication tools for gathering and sharing the knowledge gained through such valuable experiences. However, writing stories or narratives is not necessarily an easy task for humans. Even good writers struggle with creating a story that is not only syntactically and semantically sound but also describes the chain of events in a meaningful way. Also, finding the correct language elements leading to the generation of a good story is challenging. For example, the type of text (e.g., narrative vs. expository text) and readability (e.g., sentence and passage length) may affect how accurately individuals with differential reading abilities can comprehend a story (Begeny & Greene, 2014; Sáenz & Fuchs, 2002).

In schools, storytelling has always been a part of children's language and literacy development, especially in terms of oral fluency and reading comprehension (Agosto, 2016; Miller & Pennycuff, 2008; Peck, 1989). Both fluency and comprehension are highly essential skills for learning other subjects because students' ability to understand what they read in these subject areas is strongly associated with their reading fluency and comprehension (Bigozzi et al., 2017). Teachers typically use a variety of literature selections to improve children's oral fluency and comprehension skills and help them make the transition from learning to read to reading to learn. With the emergence of online or digital reading materials, teachers have also begun to use learning and assessment tools focusing on online reading comprehension (Bulut et al., 2022). Therefore, teachers always need new learning resources (i.e., online reading materials) and assessment tools to gauge children's academic growth in online reading comprehension.

Researchers found that the development of reading comprehension skills depends highly on the quality of reading materials teachers select for their students (Taylor et al., 2003; Tivnan & Hemphill, 2005). Teachers must look for digital reading materials suitable for their students to support children's literacy development. However, this is costly because most digital literacy materials are commercial and thus require a paid subscription. Also, teachers need to develop items based on each reading material that could help them evaluate students' reading comprehension skills. Traditional procedures for creating items for reading comprehension assessments (e.g., manually developing items starting with where, when, when, who, and so on) are laborious, challenging, and costly. Emerging technologies can facilitate the search for appropriate reading materials and items for teachers, such as text generation using language models and automatic item generation (see Das et al. [2021] for a detailed summary of the state-of-the-art techniques used to generate items automatically).

1.2. Current Study

Previous studies indicated that students could improve their reading comprehension skills when they practiced reading frequently (Allington et al., 2010; Duke & Pearson, 2009; Duke et al., 2011; Guthrie, 2004; Kim & White, 2008; Rasinski, 2012; Taylor et al., 2000). In K-12 education, teachers use different kinds of grade-appropriate texts (e.g., fables, fairy tales, and short stories) to help students develop reading comprehension skills. This approach is essential for students who struggle with reading comprehension because they need to practice their reading skills more often by reading more texts. Because intensive reading is necessary for students with or without adequate reading comprehension skills, teachers need new reading materials constantly. Finding a relevant text from the literature is time-consuming because teachers must go through many pieces of printed or digital literature, and most materials are commercially available. In addition, the digital learning environment in the 21st century requires digital tools, including the availability of digital reading materials that can support teaching and learning activities. Therefore, there is a need in K-12 education to leverage the potential of digital instructional materials to foster students' reading comprehension skills. To address this need and provide a practical and sustainable solution, we aimed to build a story generation system to help teachers find suitable reading materials for their students. The primary objective of our study was to create an artificial intelligence (AI) system that can analyze existing reading materials to develop new stories and related items to improve students' reading comprehension skills.

2. METHOD

Emerging technologies, such as digital learning platforms and intelligent tutoring systems, have reshaped education during the past decade. These tools are frequently used in the classroom by K-12 teachers, and it is vital to design more digital tools to suit the learning needs of students. One of these learning needs is to provide reading resources and items to help students improve their reading comprehension skills. However, there is only a limited number of open-access digital reading resources available, and thus, teachers would have to spend a significant amount of time searching for appropriate materials for their students. This study aims to create an AI-based system that can analyze existing reading materials to create new, authentic texts and related items that can be used to improve and assess elementary students' reading comprehension skills. To achieve our goals, we fine-tune a pre-trained transformer model to generate new texts (i.e., reading passages) based on existing reading materials and create related items for the texts generated by the transformer model. The following sections will describe the story and item generation sections in detail.

2.1. Story Generation

We fine-tuned a pre-trained transformer model using classic children's books to perform story generation through a decoding approach. We searched reading materials (i.e., fairy tales and fables) that were freely available on the Internet and saved the grade-appropriate examples. In total, the dataset consisted of 3,700 human-written stories. During the training process, the

Adamax optimizer was applied with a learning rate of 5e-5, the batch size was 32, and the total number of training epochs was 3.

2.1.1. *Transformer Model: GPT-2*

Large-scale neural language models such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019), Generative Pre-trained Transformer (GPT; Radford et al., 2018) and GPT-2 (Radford et al., 2019) have been extensively trained on massive amounts of text to be used for complex language tasks. Pre-trained transformer language models demonstrate state-of-the-art performance across different natural language tasks such as text generation, summarization, and translation. These models can be expected to generate fluent and diverse texts due to the large amounts of data they were trained on (See et al., 2019). The reason behind the success of the transformer-based models for different natural language tasks is the diversity of the training dataset. They generate texts representative of the corpora on which they were trained. A common approach is to fine-tune these language models to a specific domain of interest by providing different corpora of exemplars. These transformer-based models can effectively learn from training data and generate high-quality texts by fine-tuning pre-trained models. This study uses GPT-2 model—a neural language model that achieves state-of-the-art performance across different tasks. The GPT-2 language model was trained with 1.5 billion parameters on a dataset of 8 million web pages to predict the next word for the previous words within a text (Radford et al., 2019).

2.1.2. *Decoding Algorithms*

Neural text decoding algorithms highly influence the quality of text generated (Holtzman et al., 2019; Kulikov et al., 2018). During decoding, a vector is applied to the softmax function to convert it into a probability for each word:

$$P(x|x_{1:i-1}) = \frac{\exp\,(u_i)}{\sum_j \exp\,(u_i)}, \tag{1}$$

where x is a token (e.g., words, characters, or subwords) at timestep i and u is a vector that contains the numerical value of every token in the vocabulary V. Considering the critical role of decoding algorithms in improving the performance of language models, we experimented with different decoding algorithms (beam search, random sampling with and without temperature, top-k sampling, and top-p sampling) with different parameters for each method (e.g., $p = 0.90$, $p = 0.92$, or $p = 0.95$ for top-p sampling) because the correct decoding algorithm is needed to generate high-quality and meaningful texts.

2.1.2.1. Beam Search. Beam search generates all possible tokens in a vocabulary list and then chooses the top B number of candidates with the highest probability at each timestep (Holtzman et al., 2019). However, the search may fail to choose between the two words or phrases and yield a text that repeats the same word or phrase. Therefore, it tends to produce low-quality texts with short sentences and excessive repetitions (Fan et al., 2018; Basu et al., 2020).

2.1.2.2. Random Sampling. This method uses the probability of each token from the softmax function to generate the next token (Holtzman et al., 2019). Thus, it samples directly from probabilities estimated by the model and can generate incoherent texts (Holtzman et al., 2019).

2.1.2.3. Sampling with Temperature. A probability distribution can be shaped through temperature (Holtzman et al., 2019). Temperature increases the probability of probable tokens while decreasing the likelihood of less probable tokens. It has been widely applied to text generation (Fan et al., 2018). Higher temperature values result in higher randomness in the generated text. Temperature is used to scale the value of each token before going into a softmax function. Thus, given the temperature *t*, the softmax is re-estimated as follows:

$$P(x|x_{1:i-1}) = \frac{\exp{(u_i/t)}}{\sum_j \exp{(u_i/t)}}. \tag{2}$$

2.1.2.4. Top-*k* Sampling. Top-*k* sampling samples the next word from the *k* most likely words (Fan et al., 2018; Holtzman et al., 2018). Thus, top-*k* sampling involves a fixed number of most likely words and ensures that less probable words are not sampled. Because the top-*k* sampling restricts selection to the *k*-most likely words, the *k* subset of vocabulary, *V*, maximizes the probability of selected words:

$$\sum_{x \in V^{(k)}} P(x|x_{1:i-1}). \tag{3}$$

2.1.2.5. Top-*p* Sampling. Top-*p* or nucleus sampling restricts the sampling process to the smallest possible set of words whose cumulative probability exceeds the probability threshold (Holtzman et al., 2019). Top-*p* sampling distributes the probability among this set of words, and thus, the number of words in that set can dynamically increase or decrease based on the subsequent probability distribution, indicating that it involves a dynamic number of words based on a fixed *p* value:

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p, \tag{4}$$

where $V^{(p)}$ is the smallest possible set of words, $P(x|x_{1:(i-1)})$ is the probability of generating word *x* given the previously generated words *x* from 1 to $(i-1)$. This shows that the model selects the highest probability set of words whose cumulative probability exceeds the pre-chosen threshold *p*. Similar to the beam search, top-*k* and top-*p* sampling methods sometimes repeat words in a generated text for small values of *k* and *p*, while similar to random sampling, they generate incoherent text for large values of *k* and *p* (Basu et al., 2020).

2.1.2.6. Hybrid Sampling. We also tested a hybrid sampling approach (i.e., the combinations of top-*k* and top-*p* sampling).

2.1.3. *Model Evaluation*

To evaluate each story generation model, we performed human evaluation by rating the quality of generated stories based on five criteria: fluency, coherence, grammar, logical ordering of events, and human-sounding. We used a 5-point scale with the following score categories: 1 = Fundamental errors and no meaning; 2 = Fundamental errors and difficult to understand the meaning; 3 = Moderate errors but reasonably easy to understand the meaning; 4 = Minor errors and reasonably easy to understand the meaning; and 5 = Minor errors and easy to understand the meaning. To facilitate human evaluation, we generated stories with 100 words and selected a subsample of 15 texts for each prompt (prompt 1: “It was a beautiful day.” and prompt 2: “Once upon a time”), resulting in 30 texts from each model (i.e., beam search, random sampling with and without temperature, top-*k* sampling, top-*p* sampling, and hybrid sampling) and a total of 180 texts. We selected the parameters of the fine-tuned model and decoding algorithms based on human evaluations.

In addition to human evaluation, we used the perplexity (PPX) index as a data-driven metric for evaluating automatic story generation models. The PPX index is widely used in natural language processing (NLP) for evaluating language models. It measures how well a language model predicts text (i.e., probabilities of selecting the right words for an unseen test set). The

PPX index is typically calculated as the inverse probability of a test set (i.e., a sequence of tokens produced by the language model), normalized by the number of words in the test set:

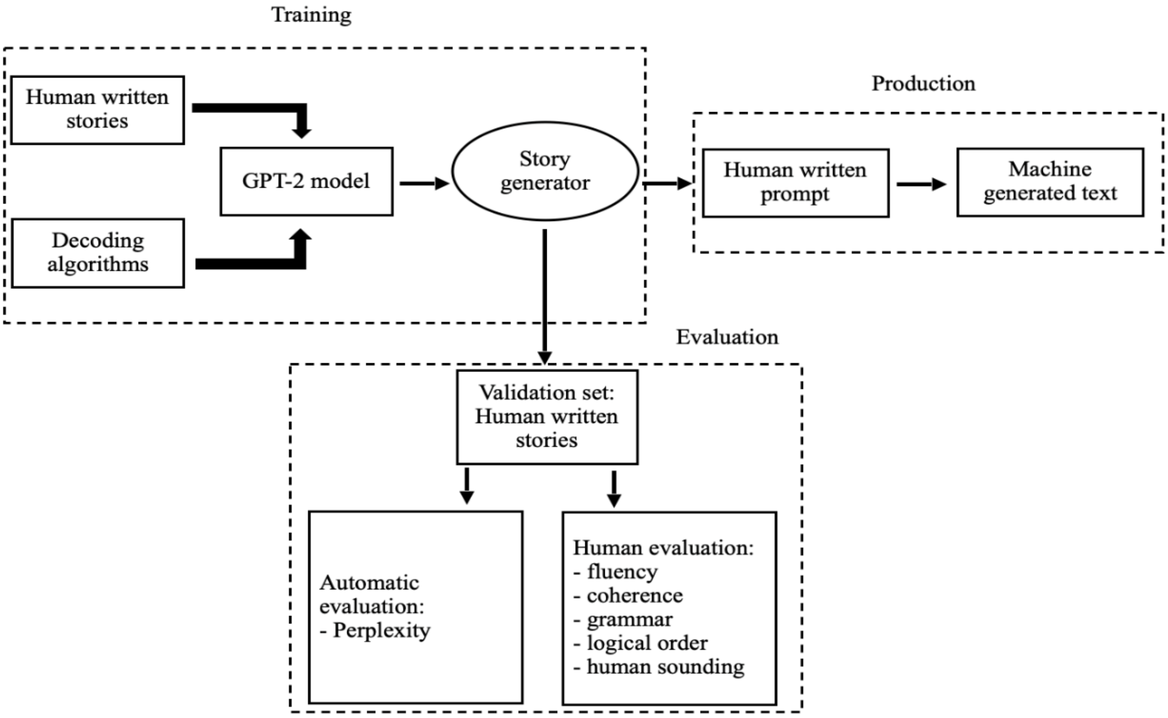
$$PPX(W) = P(w_1, w_2, \dots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}}, \tag{5}$$

where W is a tokenized sequence with N tokens, $W = (w_1, w_2, \dots, w_N)$ and $P(w_1, w_2, \dots, w_N)$ is the probability of observing a particular sequence of tokens. The lower the average value of PPX, the more accurate a language model. The PPX index can also be expressed as the exponential of the cross-entropy:

$$PPX(p, q) = -\sum_{X} p(x) log q(x), \tag{6}$$

where X refers to the language model's vocabulary of possible tokens (e.g., words or phrases), $p(x)$ is the target distribution for tokens, and $q(x)$ is the estimated distribution for tokens. PPX gets smaller as the predicted distribution becomes closer to the target distribution. In this study, the lower the perplexity of a story generation model, the better the model's accuracy when creating a new story. Figure 1 depicts the proposed framework for automatic story generation and model evaluation.

Figure 1. *The proposed framework for automatic story generation.*



2.2. Story Generation

We studied answer-aware item generation by jointly training item generation and answering and answer-agnostic item generation and compared their performance in terms of the quality of items generated. With answer-aware item generation, we aimed to design an algorithm that generates items and answers simultaneously and improves the performance of each other. We used a pre-trained transformer architecture to develop answer-aware and answer-agnostic item generation models. In terms of input, we used the texts generated from the story generation model and did not perform pre-processing (e.g., convert complex sentences into more straightforward sentences).

2.2.1. *Model Evaluation*

Item generation models aim to automatically generate a set of items that can be answered based on a particular content (Rus et al., 2012). This content can be a single sentence, paragraph, document, or database. Some researchers studied item generation and answer generation as dual tasks (e.g., Tang et al., 2017), while others generated items from texts without answers (Du & Cardie, 2017).

2.2.1.1. Answer-Aware Item Generation. Answer-aware item generation systems function with the content and generate items for target answers. However, generated items can be limited to certain types of items and focused on name entities (Dong et al., 2018) or arbitrary entities (Duan et al., 2017; Wang et al., 2020). Thus, answer-aware item generation approaches have the drawback of generating answers focusing on entities, and most items are easy to answer.

2.2.1.2. Answer- Agnostic Item Generation. Answer-agnostic item generation eliminates the requirement of the target answer before the items are generated. Answer agnostic item generation approaches reduce the bias toward entities while expanding the model flexibility (Wang et al., 2020). Although this approach is likely to generate more diverse items, it may also generate unanswered items (Sun et al., 2018; Wang et al., 2020).

2.2.2. *T5: Text-To-Text Transfer Transformer*

There are three approaches to item generation: rule-based, neural-based, and transformer-based. Item generation with a rule-based approach involves manually written rules for item generation based on heuristic rules and linguistic knowledge. The rule-based item generation systems can transform declarative sentences into interrogative sentences (e.g., Heilman & Smith, 2010; overgenerate and rank approach). However, these models are brittle and heavily depend on human effort. Therefore, rule-based models cannot be easily adapted to other domains (Zhou *et al*., 2018). Although rule-based models were more prevalent in generating items until the mid-2010s, there has been an increase in using neural networks since then (Pan *et al*., 2019).

Item generation with a neural-based approach trains a neural network based on a sequence-to-sequence framework from scratch. For example, Du et al. (2017) used a neural language model with an encoder-decoder architecture of the sequence-to-sequence model to generate items without relying on hand-crafted rules. An input sentence and its containing paragraph are encoded, and an item is generated by the decoder. Their proposed model outperformed the rule-based models (e.g., Heilman & Smith, 2010). However, the inherent sequential nature of these models makes it difficult to process long sequences. The sequence-to-sequence models cannot capture paragraph-level content, which is necessary to generate high-quality items. A generated item does not explicitly connect with the context of the target answer, and thus, includes a substantial portion of the target answer (Liu, 2020). Existing item generation models (e.g., Du et al., 2017) mostly use sentence-level content to generate items because models show significant performance degradation when applied to paragraph-level or long content. The transformer-based models address these problems.

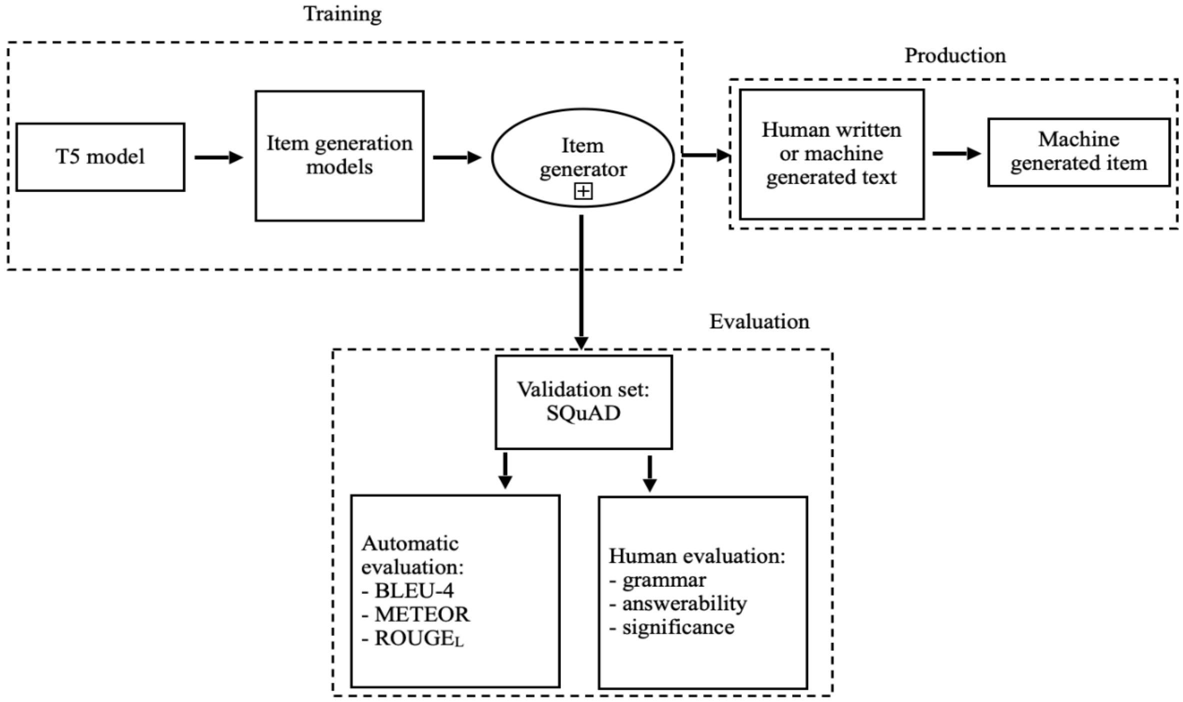
Transformers train and provide pre-trained models that show significant performance improvements in the NLP tasks (Radford et al., 2018). With transformer-based models, it is possible to improve the importance of item generation and to process paragraph-level content for item generation. We used T5: Text-to-Text Transfer Transformer that uses a text-to-text framework (i.e., takes text as input and generates new next as output) (Raffel et al., 2019). The T5 model is pre-trained on Colossal Clean Crawled Corpus (C4) and can be fine-tuned to achieve state-of-the-art results on different NLP tasks (Raffel et al., 2019). We trained the T5-small model for answer-aware and answer-agnostic item generation models and compared their performances.

2.2.3. *Model Evaluation*

We performed data-driven and human evaluations to analyze the performance of the item generation models. In terms of data-driven evaluation, we computed and reported BLEU, METEOR, and ROUGE scores using the SQuAD dataset (Rajpurkar et al., 2016). These metrics assign a score by measuring n-grams (i.e., sequence of words) and their frequency by comparing generated text with reference text. BLEU score is a more precision-based metric that provides an overall assessment of model quality by measuring the similarity of the generated text to the reference texts without considering semantic similarity (Papineni et al., 2002). BLEU-n (e.g., BLEU-4) counts co-occurrences by using up to n-grams. METEOR is a more recall-based metric that provides the similarity between generated texts and reference texts by considering synonyms, stemming, and paraphrases (Denkowski & Lavie, 2014). ROUGE is a more recall-oriented metric that compares generated text against reference text (Lin, 2004). ROUGE$_L$ measures the longest co-occurrence in n-grams by considering sentence-level structure similarities. For all three indices, larger values indicate better results.

In addition to data-driven evaluation based on the BLEU, METEOR, and ROUGE$_L$ scores, the generated items from the answer-aware and answer-agnostic models were also subject to human evaluation. We randomly selected 20 sets of items from each model using the inputs generated by the story generation model with the hybrid sampling approach. Two human evaluators rated the quality of the items based on the following criteria: grammar, answerability (i.e., the item can be answered based on the paragraph), and significance (i.e., the item relies on an essential piece of information from the paragraph). We used a 5-point scale ranging from 1 (very poor) to 5 (very strong) in the human evaluation of generated items. Figure 2 depicts the proposed framework for automatically generating items based on reading passages.

Figure 2. *The proposed framework for automatic item generation.*



3. RESULTS

3.1. Results for Automatic Story Generation

Table 1 shows two example stories generated using each decoding algorithm for two prompts: "It was a beautiful day." and "Once upon a time." Texts generated by the model with beam search showed extreme repetitions. The first example shows that the same sentence was repeated throughout the text, and the second example includes the repetition of words and sentences. Other researchers also reported a similar finding regarding beam search (e.g., Fan et al., 2018). In terms of the texts generated by the model with random sampling, although the texts may initially seem acceptable, when taking a closer look, they are not coherent and human-sounding. The reason might be that random sampling generates the next token by randomly sampling word sequences. When we tried random sampling with temperature to increase the likelihood of high probability words and decrease the likelihood of low probability words, the generated texts became coherent. However, the generated text still seemed problematic based on other criteria, particularly the logical ordering of the events.

Table 1. *Samples of generated texts for different decoding methods*.

Decoding Methods	Generated Texts
Beam search	*It was a beautiful day.* It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day … *Once upon a time*, there was a man in the middle of the street. He was dressed in a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt …
Random sampling	*It was a beautiful day.* The only real question is, will it be an issue when the Jets play at their pre-season home for the first time since 2011? The Jets have one more chance Sunday, when the San Francisco 49ers play host to the Indianapolis Colts. If they can't land a first-round draft pick from Chicago in Philadelphia as they did with the New York pick, this is just an opportunity to build something interesting with an actual head coach … *Once upon a time*, when one felt certain of his or her own powerlessness and self-hatred for others, he or she acted from a position of obedience to the law, but in another way acted from a position of obedience to his or her own and for the common good. Thus it is in the course of the history of the United States that it is possible to find and articulate an attitude of gratitude and admiration for others for acts of kindness …
Sampling with temperature	*It was a beautiful day.* We had fun and we enjoyed it as much as anyone else. We were just very fortunate that we had a great time. The team's goal was to win the game and then make it 5-0. But it didn't work out that way. The Red Bulls, who won their first three games in Seattle, were 2-0 down at the break for the last time, and the Red Bulls were 0-2 down on their way to … *Once upon a time*, when we went to a place like this, it felt like a paradise. It was so cool to go to. It felt like heaven. It felt like a place where you could live. The only thing that I did that was drive past and watch the sunset. I can't remember anything about it. I just remember being in a place where the sun is shining. And I never saw the sunset before …

Table 1. *Continued***.**

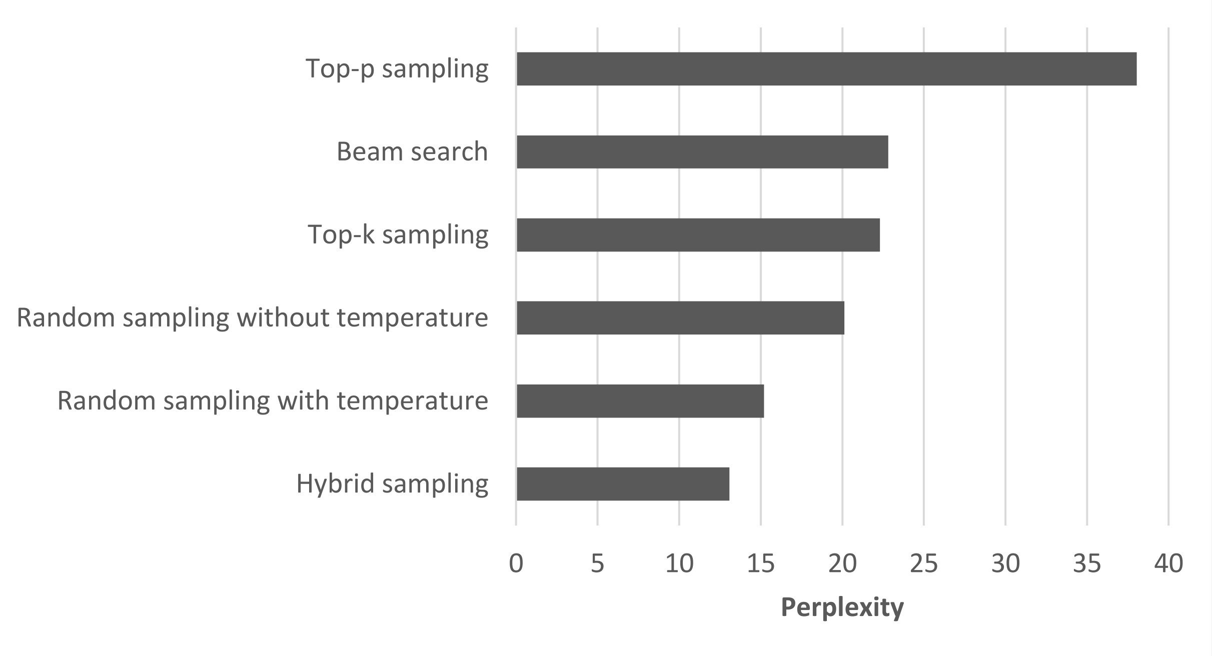
Decoding Methods	Generated Texts
Top-*k* sampling	*It was a beautiful day*. It reminded me that even though it was beautiful to me it was also a little bit hard to understand. I wasn't really ready to learn how to speak or speak Spanish so naturally I was getting into the habit of saying this to friends and family and sometimes I would say something like, "This isn't my first time speaking Spanish." But not anymore. The realization that my accent is more than just a language barrier made me realize I probably could do better … *Once upon a time* the same man had the power to force the enemy to surrender. He had no power to restrain them when he was in charge, so as to not give an enemy the chance of destroying him. The enemy, therefore, was allowed to escape without being put to any use. On this account the men of the village were ordered also to put to death their prisoners, and to eat of the corpses …
Top-*p* sampling	*It was a beautiful day*. I remember once in college the walls of the MIT Building were covered with graffiti and we all went down the hallway. So we all had fun around the MIT Building—people going to get their snowman blankets and they were breaking windows. In the morning I would go back to my office to read the fan pages that I would roll out in the middle of the night and go to class … *Once upon a time*, when thinking was a bold idea and no longer the tool of choice, power processes had a wider arc than experience and knowledge. Now we are beginning to recognize that power processes (as opposed to wishful thinking) are just a subset of working memory—each of them operates on the same information field, but with different energies. The architect of higher-order concepts has his or her own practices, decisions, and combinations of interests that could …
Hybrid sampling	*It was a beautiful day*. I was just amazed and happy to see this amazing team in action. I'm so thankful for the support and the training that they have given me over the years. I'm really happy that I have received so many support from the whole team and I'm so thankful for the support that they have given me … *Once upon a time*, the people of this country are working to create a world where the public is comfortable and the private sector is able to manage the economy. And that includes giving our youth a voice. And that includes educating them about the importance of social responsibility and the role of government in managing our economy. And that includes making sure that they understand the important role of private sector employees in our economy …

Compared to the other decoding algorithms, the model with top-*k* sampling generated higher-quality texts. This approach is more powerful in text generation as it filters only *k* most likely words and distributes probability among those *k* following words. The model generated texts with higher fluency, coherence, grammar, logical ordering of events, and human-sounding. However, it still shows some problems in terms of human-sounding (e.g., "how to speak or speak Spanish so naturally"). The reason can be that top-*k* sampling does not involve a dynamic number of words as it uses a fixed *k* number of words, limiting creativity in the model.

Using top-*p* sampling to sample from the smallest possible set of words instead of sampling only from the most likely k words produced texts with a wide range of words. Although both top-*k* and top-*p* produced high-quality texts, top-*p* seems to be a better decoding algorithm than top-*k* in theory (i.e., dynamic number of words). Finally, we had better results when we tried a hybrid of top-k and top-p sampling. After human evaluation of models by two raters, we selected the hybrid sampling—a combination of top-*p* and top-*k* sampling. The hybrid sampling was substantially more effective than other approaches because it generated texts with better fluency, coherence, grammar, logical ordering of events, and human-sounding.

In addition to human evaluation, we also used the PPX index to make a data-driven comparison among the story generation approaches. Figure 3 shows the perplexity results for each decoding method. The hybrid sampling approach yielded the smallest PPX value, suggesting that the text generated by this approach had the least amount of randomness based on the underlying language model. Surprisingly, top-*p* sampling yielded the largest perplexity value, followed by beam search. This finding indicates that the text generated by the top-*p* decoding method did not necessarily follow the underlying language model accurately. In other words, the text generated by top-*p* sampling included a high amount of randomness.

Figure 3. *Perplexity values for generated texts by different decoding algorithms.*



3.2. Results for Automatic Item Generation

Table 2 shows a sample of answer-agnostic and answer-aware items produced by the T5 model. The results showed that although the answer-agnostic item generation method allows for more flexibility in the language model (Wang et al., 2020), the answer-aware item generation method produced a higher number of quality items. Both methods (i.e., answer-agnostic and answer-aware item generation) yielded reading comprehension items with "what" and "how" questions. We identified two problems with the answer-agnostic items. First, some items already included the answer within the item stem. Second, some items involved a question statement that cannot be answered based on the available information from the reading passage. The items generated by the answer-aware item generation method were also problematic in other ways. For example, some items (e.g., "What kind of day was it?") required very simple or trivial answers. Also, some items were unnecessarily long, such as "What do the people of this country do to create a world where the public is comfortable and the private sector is able to manage the economy?". Further revisions are necessary to make these items helpful in evaluating students' reading comprehension.

Table 2. *Sample items generated by the answer-agnostic and answer-aware models.*

Model Input	Answer-Agnostic Items	Answer-Aware Items
It was a beautiful day. I was just amazed and happy to see this amazing team in action. I'm so thankful for the support and the training that they have given me over the years. I'm really happy that I have received so many supports from the whole team and I'm so thankful for the support that they have given me.	What was a beautiful day for me? What team has given me the support they have given me over the years? How many support did I receive from the whole team?	Q: What kind of day was it? A: Beautiful Q: What team was in action? A: Amazing Q: What is the reason I'm so thankful for? A: The support and the training that they have given me Q: Who has given me so much support? A: The whole team
Once upon a time, the people of this country are working to create a world where the public is comfortable, and the private sector is able to manage the economy. And that includes giving our youth a voice. And that includes educating them about the importance of social responsibility and the role of government in managing our economy. And that includes making sure that they understand the important role of private sector employees in our economy.	What are the people of this country working to create a world where the public is comfortable and the private sector is able to manage the economy? What does that include giving our youth a voice? How do the youth learn about the importance of social responsibility and the role of government?	Q: What are the people of this country working to create a world? A: The public is comfortable and the private sector is able to manage the economy Q: What do the people of this country do to create a world where the public is comfortable and the private sector is able to manage the economy? A: Giving our youth a voice Q: What is the key to educating youth about the importance of social responsibility and the role of government in managing our economy? A: Making sure that they understand the important role of private sector employees in our economy

Q: Question; A: Answer.

Table 3 shows the model evaluation indices for the items generated by the answer-agnostic and answer-aware methods. The results show that the answer-aware item generation performed slightly better than the answer-agnostic item generation; however, the difference between the two methods was negligible. Overall, the findings of our study appear to broadly support the work of other studies in automatic item generation. In our study, the answer-agnostic method yielded unanswerable items and failed to generate diverse items (Sun et al., 2018; Wang et al., 2020). Also, the answer-aware method yielded simple items that do not necessarily require higher levels of reading comprehension to find the correct answer.

Table 3. *Evaluation indices for the items generated by the answer-agnostic and answer-aware methods.*

Item Generation Model	BLEU-4	METEOR	ROGUE$_L$
Answer-Agnostic	18.3	24.7	39.9
Answer-Aware	18.6	24.9	40.2

4. DISCUSSION and CONCLUSION

Pre-trained transformer models can generate high-quality texts and items due to the large amounts of corpus they are trained on (See et al., 2019). In this study, we fine-tuned pre-trained transformer models to generate new stories and related items to enhance and assess students' reading comprehension skills. The proposed story and item generation models attain a fine-tuned understanding to produce human-like stories and items. However, it should be noted that the models might generate stories with repetitive words or unnatural changes in the topic. These weaknesses of language models remain a common challenge for the NLP community (Radford et al., 2019).

Our story generation model with hybrid sampling showed promising results in producing fluent, coherent, grammatically correct, logical, and human-sounding stories that students could use to practice and enhance their reading comprehension skills. Also, our answer-aware item generation model showed promising results in producing grammatically correct, answerable, and significant items. These language models for automatic story and item generation could enable teachers to generate authentic stories and items on the fly and share them with their students easily, without having to look for freely available printed or digital materials for hours. However, it should be noted that the generated items may still require a human evaluation and further adjustments before sharing them with students as they are likely to involve semantic errors (i.e., grammatically correct but nonsensical text). Also, the generated items may not be suitable for measuring complex reading skills such as inferencing, analyzing, and critiquing. Overall, the proposed models provide a feasible solution to the problem of finding new texts from the limited printed or digital materials and related items to the texts.

There are several limitations of this study. First, we used GPT-2 small and T5-small (i.e., the smallest versions of GPT-2 and T5) to generate stories and items due to their relatively less demand for computing power. It is possible that more advanced versions of the GPT-2 (e.g., GPT-2 large) and T5 (e.g., T5-base and T5-large) could generate higher-quality stories and items. Second, this study used a training dataset that involved freely available reading materials (i.e., fairy tales and fables) available on the Internet. A larger-size training dataset including more diverse reading materials (e.g., short stories, articles, or novels) could help fine-tune a transformer model more effectively and yield more consistent results in story and item generation stages. Finally, the sample stories and items generated in this study were not shared with students. Future studies on automatic story and item generation could involve students who can provide feedback on the readability and clarity of the generated stories and items. The feedback from students could facilitate the fine-tuning of pre-trained language models.

Acknowledgments

The research leading to these results received funding from the University of Alberta's Endowment Fund for the Support for the Advancement of Scholarship (SAS) Program.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Okan Bulut: Investigation, Resources, Methodology, Software, Formal Analysis, and Writing-original draft. **Seyma Nur Yildirim-Erbasli**: Methodology, Software, Formal Analysis, and Writing-original draft.

Orcid

Okan Bulut https://orcid.org/0000-0001-5853-1267
Seyma Nur Yildirim-Erbasli https://orcid.org/0000-0002-8010-9414

REFERENCES

Agosto, D.E. (2016). Why storytelling matters: Unveiling the literacy benefits of storytelling. *Children and Libraries*, *14*(2), 21-26. https://doi.org/10.5860/cal.14n2.21

Allington, R.L., McGill-Franzen, A., Camilli, G., Williams, L., Graff, J., Zeig, J., Zmach, C., & Nowak, R. (2010). Addressing summer reading setback among economically disadva ntaged elementary students. *Reading Psychology, 31*(5), 411-427. https://doi.org/10.1080/02702711.2010.505165

Basu, S., Ramachandran, G.S., Keskar, N.S., & Varshney, L.R. (2020). Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint*. https://doi.org/10.48550/arXiv.2007.14966

Begeny, J.C., & Greene, D.J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools*, *51*(2), 198-215. https://doi.org/10.1002/pits.21740

Bigozzi, L., Tarchi, C., Vagnoli, L., Valente, E., & Pinto, G. (2017). Reading fluency as a predictor of school outcomes across grades 4-9. *Frontiers in Psychology, 8*(200), 1-9. https://doi.org/10.3389/fpsyg.2017.00200

Bulut, H.C., Bulut, O., & Arikan, S. (2022). Evaluating group differences in online reading comprehension: The impact of item properties. *International Journal of Testing*. Advance online publication. https://doi.org/10.1080/15305058.2022.2044821

Das, B., Majumder, M., Phadikar, S., & Sekh, A.A. (2021). Automatic question generation and answer assessment: A survey. *Research and Practice in Technology Enhanced Learning* , *16*(1), 1-15. https://doi.org/10.1186/s41039-021-00151-1

Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376-380).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirec tional transformers for language understanding. *arXiv preprint*. https://doi.org/10.48550/arXiv.1810.04805

Dong, X., Hong, Y., Chen, X., Li, W., Zhang, M., & Zhu, Q. (2018, August). Neural question generation with semantics of question type. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 213-223). Springer, Cham.

Du, X., & Cardie, C. (2017, September). Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2067-2073). https://doi.org/10.18653/v1/D17-1219

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *arXiv preprint*. https://doi.org/10.48550/arXiv.1705.00106

Duan, N., Tang, D., Chen, P., & Zhou, M. (2017, September). Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 866-874). https://doi.org/10.18653/v1/D17-1090

Duke, N.K., & Pearson, P.D. (2009). Effective practices for developing reading comprehension. *Journal of Education, 189*(1/2), 107-122. https://doi.org/10.1177/0022057409189001-208

Duke, N.K., Pearson, P.D., Strachan, S.L., & Billman, A.K. (2011). Essential elements of fostering and teaching reading comprehension. *What research has to say about reading instruction*, *4*, 286-314.

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. *arXiv preprint*. https://doi.org/10.48550/arXiv.1805.04833

Guthrie, J.T. (2004). Teaching for literacy engagement. *Journal of Literacy Research, 36*(1), 1-30. https://doi.org/10.1207/s15548430jlr3601_2

Heilman, M., & Smith, N.A. (2010, June). Good question! Statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 609-617).

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.09751

Holtzman, A., Buys, J., Forbes, M., Bosselut, A., Golub, D., & Choi, Y. (2018) Learning to write with cooperative discriminators. *arXiv preprint*. https://doi.org/10.48550/arXiv.1805.06087

Kim, J.S., & White, T.G. (2008). Scaffolding voluntary summer reading for children in grades 3 to 5: An experimental study. *Scientific Studies of Reading, 12*(1), 1-23. https://doi.org/10.1080/10888430701746849

Kulikov, I., Miller, A.H., Cho, K., & Weston, J. (2018). Importance of search and evaluation strategies in neural dialogue modelling. *arXiv preprint*. https://doi.org/10.48550/arXiv.1811.00907

Liu, B. (2020, April). Neural question generation based on Seq2Seq. In Proceedings of *the 2020 5th International Conference on Mathematics and Artificial Intelligence* (pp. 119-123).

Lin, C.Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

Miller, S., & Pennycuff, L. (2008). The power of story: Using storytelling to improve literacy learning. *Journal of Cross-Disciplinary Perspectives in Education*, *1*(1), 36-43.

Pan, L., Lei, W., Chua, T.S., & Kan, M.Y. (2019). Recent advances in neural question generation. *arXiv preprint arXiv:* https://doi.org/10.48550/arXiv.1905.08949

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002, July). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Peck, J. (1989). Using storytelling to promote language and literacy development. *The Reading Teacher*, *43*(2), 138-141. https://www.jstor.org/stable/20200308

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI tech report*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI tech report*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P.J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint*. https://doi.org/10.48550/arXiv.1910.10683

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392).

Rasinski, T.V. (2012). Why reading fluency should be hot! *The Reading Teacher, 65*(8), 516-522. https://doi.org/10.1002/TRTR.01077

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., & Moldovan. C. (2012). A detailed account of the first question generation shared task evaluation challenge. *Dialogue and Discourse, 3*(2), 177–204. https://doi.org/10.5087/dad

Sáenz, L.M., & Fuchs, L.S. (2002). Examining the reading difficulty of secondary students with learning disabilities: Expository versus narrative text. *Remedial and Special Education*, *23*(1), 31-41.

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C.D. (2019). Do massively pretrain ed language models make better storytellers? *arXiv preprint*. https://doi.org/10.48550/arXiv.1909.10705

Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., & Wang, S. (2018). Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3930-3939).

Tang, D., Duan, N., Qin, T., Yan, Z., & Zhou, M. (2017). Question answering and question generation as dual tasks. *arXiv preprint*. https://doi.org/10.48550/arXiv.1706.02027

Taylor, B.M., Pearson, P.D., Clark, K., & Walpole, S. (2000). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. *The Elementary School Journal, 101*(2), 121-165. https://doi.org/10.1086/499662

Taylor, B.M., Pearson, P.D., Peterson, D.S., & Rodriguez, M.C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal, 104*(1), 3-28. https://doi.org/10.1086/499740

Tivnan, T., & Hemphill, L. (2005). Comparing four literacy reform models in high-poverty schools: Patterns of first-grade achievement. *The Elementary School Journal*, *105(5)*, 419-441. https://doi.org/10.1086/431885

Wang, B., Wang, X., Tao, T., Zhang, Q., & Xu, J. (2020, April). Neural question generation with answer pivot. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 9138-9145).

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., & Zhou, M. (2017, November). Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing* (pp. 662-671). Springer, Cham.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 88–108

https://doi.org/10.21449/ijate.1104868

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Effect of formative assessment on reading comprehension

Muhammet Sonmez[1,*], **Fatih Cetin Cetinkaya**[2]

[1]Ministry of National Education, Haci Seyit Tasan Primary School, Kocaeli, Türkiye
[2]Duzce University, Faculty of Education, Department of Primary School Teaching, Duzce, Türkiye

ARTICLE HISTORY

Received: Apr. 17, 2022
Revised: Sep. 12, 2022
Accepted: Sep. 29, 2022

Keywords:
Reading assessment,
Formative assessment,
Reading,
Reading comprehension,
Primary school.

Abstract: The aim of this research is to set forth the effects of formative assessment methods on reading comprehension. To this end, reading status of a group of students was assessed with formative assessment methods, while that of another group was evaluated with traditional ones. The research was carried out by using unequalised quasi-experimental design. The experimental and control groups of the research were randomly assigned. The study group consisted of 50 3rd grade students of a primary school in the Dilovası district of Kocaeli city, Türkiye. The data of the study were obtained from the texts within 3rd grade curriculum and from the comprehension questions prepared for these texts. The data were analyzed via SPSS 22 program. Mann-Whitney and Wilcoxon signed rank tests were used during analyses. In the findings of the research, a highly significant difference was observed in favor of the experimental group. As a result of the findings of the research, it was observed that formative assessment methods contributed to reading comprehension success positively.

1. INTRODUCTION

Education of reading and reading comprehension starts from the 1st grade of primary school education and is carried out by increasing it gradually. As it is included within the aims of Turkish Language Education Program of the Ministry of National Education (MoNE) (2019), love and habit of reading and writing must be given to students and with this habit it must be ensured that they are given the opportunity to assess what they read and comprehend in a critical way. In order to reach these goals, it will be insufficient just to see, analyze, and vocalize the marks and symbols. It is therefore necessary to technically transfer reading from vocalization into meaning set up. Reading comprehension is the process of comprehending the thoughts and messages that the author intends to convey (May & Rizzardi, 2002). Comprehension is supposed to be the basic aim of reading since the basic aim of reading is to catch the meaning (Öztürk, 2019). In order for a successful reading process, the individual needs to comprehend what s/he reads. There are studies that assert that these activities need to be performed for comprehension, which is the major aim of reading, while such activities are few in numbers within schools, teachers spend less time on reading comprehension activities than it should be,

*Corresponding Author: Muhammet SONMEZ ✉ muhametsnmz41@gmail.com Ministry of Education, Hacı Seyit Tasan Primary School, Kocaeli, Türkiye

especially in primary school years, and problems arise from these reasons (Ateş & Akyol, 2013; Ness, 2011; Neuman, 2001; Pearson & Duke, 2002). However, despite these studies, reading success in international exams has not reached a sufficient level. PISA (Program for International Student Assessment), which has been implemented since 2000, aims to evaluate students' knowledge and skills. Reading skill scores that belong to Turkish students in these exams are as follows:

Table 1. *PISA 2003- 2018 Average scores of reading skills in Türkiye.*

	2003	2006	2009	2012	2015	2018
Reading Skill	441	447	464	475	428	466
Total Average	-	-	464	471	460	453
Rank	33	37	39	42	51	40
Number of participating countries	41	57	75	65	72	79

(MoNE, 2005; MoNE, 2010; MoNE, 2013; MoNE, 2015; MoNE, 2019).

When the results of PISA are analyzed, it is observed that Türkiye scored above the average only in 2018 in the field of reading skills. However, even though it ranked 40th among 79 countries participating in 2018, it shows that there is a need to carry out further studies in this area when compared to successful countries. Learning to read well is largely achieved by carrying out more reading practice in schools. If assessment is included in the natural process of reading, it allows students to use their reading skills more easily (Landauer et al., 2009).

Instead of focusing on the status of reading success while evaluating students' reading success, evaluations should be made to determine the processes that will enable them to improve their reading success (Rogoff et al., 2001). It is emphasized that measurement and evaluation practices are an inseparable whole in the MoNE’s curriculum of Turkish (2019). In addition, individuals' interests, attitudes, and values may differ over time and in this context, it is stated that the evaluation should take place with the active participation of students and teachers in the process. Reading studies and the measurement and evaluation of reading also need to be carried out as a whole.

Formative assessment is considered as a strategy to increase the success of the student or as a strategy that serves the purpose of determining the success of the student (Clarke, 2012). This type of assessment is seen as a process which is carried out through teaching rather than grading and includes determining students' prior knowledge and organizing and implementing their teaching plans according to such information (Bulunuz & Bulunuz, 2013; Keeley, 2008). Considering that formative assessment is within the teaching process, it is used to improve learning (Oosterhof et al., 2008; Vonderwell et al., 2007). With this feature, it is also called assessment for learning (Stiggins, 2002). Most of the educators agree on the idea that addressing reading with formative assessment aims to inform education and serve student needs (Piazza, 2012). Formative assessment is crucial for reading success because it reveals students’ needs to ensure the continuous improvement in reading (Roskos & Neuman, 2012). It also facilitates modification of teaching according to students’ needs and continuously provides feedback to students (Roskos & Neuman, 2012). It is known that effective reading occurs through using such skills as phonological awareness, decoding, word recognition, vocabulary, knowledge about language structures, and using inference skill (Scarborough, 2001). It is more difficult to determine in which dimension of reading the poor reader is having difficulties when compared to the problems encountered in other academic fields (Wiliam, 2006). Which students need help can be practically determined, but more detailed information is required to specify the reasons for failure. When considered within this framework, it is considered that formative assessment can be used to improve reading comprehension.

In the related literature, there are studies on strategies for improving reading comprehension and also on measures to be taken (Aktaş, 2015; Akyol & Ketenoğlu Kayabaşı, 2018; Baştuğ & Keskin, 2011; Çeliktürk Sezgin & Akyol, 2018; Çöklü Özkan, 2018; İlter, 2018; Kocaarslan, 2015; Kodan, 2015; Kuşdemir, 2014; Papatğa, 2016; Sidekli, 2010; Sözen & Akyol, 2018). Examining the international literature, the intensity of studies on reading comprehension is also observed (Coiro & Dobler, 2007; Dreyer & Nel, 2003; Gersten et al., 2001; Hock & Mellard, 2005; Ness, 2011). There are also studies in international literature that deal with reading, comprehension, and formative assessment together (Dupont, 2018; Kline, 2013; Li, 2016; Marchand & Furrer, 2014; Marcotte & Hintze, 2009; Offerdahl & Montplaisir, 2013; Roskos & Neuman, 2012). In addition to these, there are experimental studies (Boumediene & Hamazaoui-Elachachi, 2017; Gustafson et al., 2019; Hooley & Thorpe, 2017; Sanaeifar & Nafari, 2018) examining the effect of formative assessment on reading comprehension skills in the international literature; however, these studies are limited in number. In the studies conducted to evaluate reading comprehension within the national literature, there are studies that focus on questions used in comprehension (Akyol et al., 2013; Ateş, 2011; Aydemir & Çiftçi, 2008; Doğanay & Yüce, 2010; Durukan, 2009). In addition to these, there are also studies on different measurement tools used in reading comprehension (Karasu et al., 2011; Temizkan & Sallabaş, 2011). The purpose of this specific research study is therefore to set forth the effect of formative assessment methods on reading comprehension. In line with the purpose of the research, answers to the following questions are sought:

1. Is there a significant difference between the pre-test and post-test scores of the students to whom formative assessment methods were applied?

2. Is there a significant difference between the pre-test and post-test scores of the students evaluated by traditional methods?

3. Is there a significant difference between the post-test reading comprehension scores of the students to whom formative assessment methods were applied and the ones who were evaluated with traditional methods?

2. METHOD

2.1. Research Design

With a specific aim to set forth the effect of formative assessment methods on reading comprehension skills, this study was set up as a quasi-experimental design as one of the quantitative research designs. Experimental studies aim to reveal how the independent variable of the research affects the dependent variable (Karasar, 2012). In other words, the effects of different situations -set up by the researcher- on the dependent variable are examined through experimental studies (Büyüköztürk et al., 2017; Creswell, 2014; Creswell & PlanoClark, 2011). In the quasi-experimental design, groups are randomly assigned (Büyüköztürk et al., 2017). This research model can be expressed as a pre-test post-test unequalized quasi-experimental design with a control group (Karasar, 2012) as the groups were randomly assigned and the groups were partially controllable. In other words, the groups were previously formed as classroom format within the school.

2.2. Study Group

The study group of the research consisted of 2 classes of 3rd grade students in the 2019-2020 academic year in a public primary school in the Dilovası district of Kocaeli province in Türkiye. One of the classes participating in the research was randomly assigned as the experimental group and the other as the control group. In this context, the research was conducted with 50 3rd grade students. While there was a total of 22 participants, 7 female and 15 male students in the experimental group, the control group consisted of a total of 28 participants, 15 female and 13 male students. In the study, 3rd grade students were preferred because it was necessary that the participants had to complete the literacy process and be at a level to exhibit fluent reading and comprehension skills.

While determining the study group, convenience sampling method was used in order to provide speed, practicality, and economy to the research process. With this method, researchers choose situations that are easy to reach (Glesne, 2015). Considering the ease of access and the fact that the research can be followed more closely and easily, a public school in the Dilovası district of Kocaeli province, in which the researcher also worked, was preferred during sampling.

In research, firstly it is necessary to choose the tests to determine the equivalence status of the experimental and control groups. The prerequisite for this situation is the normality of the data. When the normality distributions of the pre-test comprehension scores of the research data were examined, it was concluded that the normality distribution of the comprehension scores of the control group was S-W(28)=0.04, $p<0.05$. According to this result, the control group data are not normally distributed. When the pre-test comprehension scores of the experimental group are examined, the result emerges as S-W(22)=0.125, $p>.05$. This result shows that the experimental group data are normally distributed. In this context, nonparametric tests need be used to reveal the equivalence status between the two groups.

Table 2. *Mann Whitney U test results of the pre-test scores of the participants in the experimental and control groups.*

Pre-Test		*N*	$\bar{x}$	*U*	*Z*	*p*
Comprehension Scores	Experimental	22	7.64	223.5	-1.668	0.095
	Control	28	9.57			
	Total	50				

Table 2 shows Mann Whitney U test results of the pre-test scores of the experimental and control group participants. As can be seen in the table, the study was carried out with a total of 50 participants, including 28 participants in the control group and 22 participants in the experimental group. It can also be seen that the arithmetic mean scores of comprehension of the participants in the experimental group were 7.64 and the arithmetic mean scores of comprehension of the participants in the control group were 9.57. According to the results of Mann Whitney U test, it can be concluded that the groups were equal (U=223.5, $p>.05$). Therefore, in this context, the equivalence status of control and experimental groups of the research was ensured.

2.3. Data Collection Tools

The data of this study were collected through comprehension questions prepared in line with expert opinions regarding the text titled "Mektup –The Letter", taken from the book that was approved by the Board of Education and used as a 3rd grade Turkish textbook in the 2010-2011 academic year. No taxonomy was used while measuring the reading comprehension skill. Only comprehension questions developed in line with expert opinions were used. With these questions, the pretest-posttest comprehension scores of the participants were revealed. While three of the comprehension questions measured understanding at a simple level, two of them aimed at determining in-depth understanding. Four points were awarded for the correct answer to the simple comprehension questions and five for the correct answer to the deep comprehension problems. The lowest score that can be obtained from the test is zero, whereas the highest score is 22. While choosing the text to be used in the research, it was paid attention that the participants had not encountered this text beforehand. Among the texts in the book, used as a Turkish course book before, it was decided -in line with expert opinions- to use the text titled "Mektup – The Letter" as a measurement tool.

Mistake Analysis Inventory adapted by Akyol (2006) was used as the basis for the reading comprehension questions. According to the Mistake Analysis Inventory, reading comprehension situations can be revealed through the questions asked after the text is read silently. This inventory proposes using simple comprehension and in-depth comprehension questions. In this context, five questions, three of which aim to measure literal understanding and two to measure in-depth understanding, were created by the researchers during the research process. In the process of creating the questions, a candidate question pool of 15 questions was initially created. Candidate questions were presented to the opinions of one classroom teacher and three experts who received classroom education. In line with expert opinions, the questions to be used in the research were determined. A score of 0 was given for unanswered or incorrect answers to literal comprehension questions, 2 points for partially answered questions, and 4 points for fully answered questions. In the in-depth comprehension questions, 0 points were given for unanswered or incorrect answers, 2 points for partially answered questions, 3 points for incomplete but most of the expected answers, and 5 points for fully answered questions.

Reading comprehension questions were scored by two different raters to ensure the reliability of the research. After the scoring process, the correlation method, which is one of the approaches used to ensure inter-rater reliability, was used. Because the data were not normally distributed, Spearman Brown Rank Correlation Test was performed. According to the test results, it was revealed that there was a high correlation between the raters ($r(48)= .88$, $p=.00$, $p<.05$).

During the application process with the experimental group, formative assessment methods such as cloze test's multiple-choice format (maze), sentence verification, story map, re-telling techniques (written retell), and retelling fluency were used. According to Marcotte and Hintze (2009), the multiple-choice format (maze), sentence verification method (SVM), retelling fluency, and written retell methods of fill-in-the-blank technique can be used for formative assessment applications. The story map method, on the other hand, was used as a measurement tool in the research in line with expert opinions, considering it appropriate to see which element the student's understanding deficiencies were concentrated in and to give feedback. The texts of these measurement tools used in the experimental group were taken from the book used as a textbook in the past years. Text selection and measurement tools were carried out in line with the opinions of the three-class education experts. These measurement tools used in the experimental group were the formative assessment activities of the experimental group aiming only at the evaluation of learning.

2.3.1. *Cloze test*

Cloze test is a technique developed by Wilson Taylor in 1953, inspired by the completion principle of Gesthalt (Keskin & Akıllı, 2013; Ulusoy, 2009) and includes syntactic, structural, and semantic elements of the text (Ulusoy, 2009). With this technique, considered as an extremely reliable and valid reading comprehension measure (Bormuth, 1963), it is aimed to complete the incomplete images, thoughts or words in the mind as a whole (Akyol et al., 2014).

According to Akyol et al. (2014), at the beginning of the application phase, the teacher chooses a text suitable for the grade level. After students have read the text, every 5th, 6th, 7th, 8th, 9th, 10th words are selected and deleted from the text, except for the first-last word or proper nouns (the next word is chosen from the proper name) in the text. Students are also expected to write the same words in the text in the blanks. After the application of fill-in-the-blank test, words written correctly by the students are counted and the percentage value corresponding to total words deleted from the text is calculated (Akyol et al., 2014). According to the evaluation criteria, 60% and higher scores indicate the independent reading level, those scores between 59% and 40% indicate the instructional reading level, and 40% and below scores indicate the reading level at the anxiety level (Rankin & Culhane, 1969).

2.3.2. *Sentence verification method (SVM)*

The sentence verification test was developed by Royer et al. (1979), focusing on the structural aspect of understanding (Yazıcı & Kurudayıoğlu, 2017). In this technique, each sentence in the text read and understood by the reader has its own semantic symbols (Shaughnessy, 2005). The texts in which the sentence verification test will be used should consist of 12 sentences that are meaningful in themselves or these texts should be rearranged and expressed in 12 sentences (Akyol et al., 2014; Ulusoy & Çetinkaya, 2012; Yazıcı & Kurudayıoğlu, 2017). According to Royer (2001), each of the 12 sentences in the text should be arranged in 4 different categories; namely, using the original sentence, expressing the original sentence with other words, changing the meaning of the original sentence, and distracting sentence.

Students are expected to answer the questions formed as Yes/No or True/False in the new sentences prepared in 4 categories (Ulusoy & Çetinkaya, 2012; Yazıcı & Kurudayıoğlu, 2017). Considering the 50% chance factor of the test during the interpretation of the scores obtained from the sentence verification test, it is accepted that 80% and above correct answers indicate good understanding, while 71-79% correct answers indicate poor comprehension (Royer, 2001).

2.3.3. *Story map*

Story maps emerge as an important technique in order to reveal the connections between all the elements of the story clearly and to convey to the student how the story is organized (Mathes & Fuchs, 1997). The purpose of this technique is to create a story structure with story elements in the mind and to ensure that the texts are understood (Duman, 2006). According to Akyol (2011), distinguishing the important and unimportant information in the story, enabling the students to focus on more important information, ensuring that the information is transferred to the long-term memory regularly, making forward-looking predictions in the text by making use of prior information, and intertextual reading can be done by using a story map.

2.3.4. *Written retell*

The reading-telling technique, one of the written retell techniques, is considered to be the most important of the techniques used to assess the student's comprehension level of the text (Reutzel & Cooter, 2007). According to Leslie and Caldwell (2006), answers to 4 questions should be sought during narration in order to evaluate reading comprehension:

1. Is the basic structure of the text explained? Is important information in the text mentioned during the narration?
2. Are the main ideas and supporting ideas of the text included in the narration?
3. Is the narration sequence performed in the order in the text?
4. Is the narration complete?

By looking for answers to these questions, the student's reading comprehension status can be checked.

Fuchs et al. (1998) state that rewritten expression is a more successful method in evaluating reading comprehension rather than evaluating oral expression. At the same time, the rewritten method is a method that can be used to determine the teaching goals and also to reveal the needs of the students (Fuchs et al., 1989). Although the rewritten method does not currently have a standardized format, it is shown as a formative measure of reading comprehension (Marcotte & Hintze, 2009).

2.3.5. *Retelling Fluency*

Retelling Fluency is the evaluation of reading comprehension based on oral reading fluency (Good & Kaminkski, 2002). According to this technique; When students who read more than 40 words per minute are asked to retell the text they have read, they are expected to retell what they have read with approximately 50% of the verbal fluency score or more. In this case, the student's oral reading score can be considered as an indicator of good reading comprehension, including comprehension. When a student who reads more than 40 words per minute is asked to retell the text, it is thought that if the number of words used while describing the text is 25% or less of the verbal fluent reading score, it cannot represent reading comprehension (Good & Kaminkski, 2002). For example, if the student reads 80 words in a minute and retells the text with 40 words or more when asked to retell, reading fluency represents reading comprehension. However, if the student reads 80 words per minute and retells what s/he has read with 20 words, there may be a comprehension situation that cannot be represented with fluency.

2.4. Data Collection

2.4.1. *Preparation phase*

At this stage, a text that the participants had not encountered before was selected and comprehension questions were prepared for this text. The selected text was the one named "Mektup – the Letter" from the Turkish textbook in the 2010-2011 academic year. The preparation of the comprehension questions was carried out in line with the expert opinions. Comprehension questions for the text were prepared in line with expert opinions, and with these comprehension questions, it was aimed to measure the simple and in-depth comprehension skills of the participants. In order to determine the group equivalence within the process of determining the experimental and control groups, a pre-test was applied to all the 3rd grade classes in the school. Before the pre-test application, the participants were informed about the general framework of the research and they were all told that they should not have any grade concerns. Thus, it was aimed to create an environment where they could answer the questions sincerely.

The text "Mektup – the Letter" selected as a measurement tool was distributed to the participants and they were asked to read it once. After the reading process, pre-prepared comprehension questions regarding the text were distributed to the participants and they were expected to answer them. After the answers were received, success scores of the participants were determined and analyzed with the SPSS program. After the analysis, the equivalence status of the groups was compared. Experimental and control groups were randomly determined among the classes subjected to the pre-test process. After determining the experimental and control groups, in-depth information about the research was given by interviewing the classroom teachers of the relevant classes. In addition, they were all asked to carry out the study voluntarily and sincerely as voluntary participation of the relevant teachers in the research was very important for the effective conduct of the study.

The classroom teacher in the experimental group was informed about how the implementation phase would be carried out, and it was ensured that he became aware of the time that he had to allocate for research in the Turkish lesson. In this context and within the framework of the research, the participating students were informed that during the 10-week period, reading comprehension assessment studies would be conducted for the formative assessment approach. At the same time, the assessment tools to be used in the research, sentence verification technique, multiple choice type of fill-in-the-blank technique (Maze), story map, retelling fluency, and retell writing techniques were introduced to the classroom teacher. The classroom teacher was also informed that the assessment process of reading comprehension skill would be carried out by adopting the traditional level determination approach in the control group. The Turkish lesson and the evaluation of comprehension skills continued in its normal course without any intervention in the control group. However, the teacher was informed that the texts used in the experimental group of the research should be used when applying the comprehension test.

2.4.2. *Implementation phase*

Although the research process had been planned as 10 weeks, the research implementation process was extended to 13 weeks due to the fact that the implementation phase coincided with the 1st semester break and also because of the experimental group participants' school attendance problems in some weeks of the research process. In the 14th week, the application phase of the research was concluded by applying the text applied in the pre-test and the comprehension questions about this text to the participants. At this stage, reading comprehension skills of the participants in the control group were dealt with traditional approaches and these participants were subjected to three different evaluations during the 13-week period. The researchers did not interfere with the frequency of evaluation. During the evaluations, the texts used in the experimental group that week were also applied to the control group. At the 14th week, the control group's post-test scores were obtained through the text used in the pre-test and also through the comprehension questions for this text. In the experimental group, reading comprehension skills of the participants were tested with a formative assessment method every week. The implementation process of the research was carried out as follows:

Table 3. *Implementations conducted during the implementation phase of the research.*

Week	Implementations conducted
1st Week	The reading comprehension skills of the experimental group participants were evaluated with the sentence verification technique and necessary feedback was given.
2nd Week	The reading comprehension skills of the experimental group participants were evaluated using the fill-in-the-blank technique and necessary feedback was given.
3rd Week	No evaluation could be conducted due to lack of participants.
4th Week	The reading comprehension skills of the experimental group participants were evaluated with the story map technique and necessary feedback was given.
5th Week	No evaluation could be conducted due to lack of participants.
6th Week	While the reading comprehension skills of the experimental group participants were evaluated with the rewriting technique, the reading comprehension status of the control group participants was subjected to the 1st evaluation with the classical question and answer method.
7th Week	The reading comprehension skills of the experimental group participants were evaluated using the retelling fluency method and necessary feedback was given.
8th Week	The reading comprehension skills of the experimental group participants were evaluated with the sentence verification technique and necessary feedback was given.
9th Week	While the reading comprehension skills of the experimental group participants were evaluated with the fill-in-the-blank technique and feedback was given, the second evaluation for the control group was conducted.
10 th Week	No evaluation could be conducted due to semester break.
11 th Week	The reading comprehension skills of the experimental group participants were evaluated with the story map technique and necessary feedback was given.
12th Week	The reading comprehension skills of the experimental group participants were evaluated with the rewriting technique and necessary feedback was given.
13th Week	While the reading comprehension skills of the experimental group participants were evaluated with the retelling fluency technique and the necessary feedback was given, the third evaluation was conducted for the control group.

Following each evaluation made regarding the experimental group, the participants were given feedback on where their understanding deficiencies were and how they could overcome these. While giving feedback, no judgment was made in the classroom and every attempt was made to prevent labelling students as successful or unsuccessful. No scoring was used during the evaluation. In order for the participants to see their own mistakes and shortcomings, their understanding deficiencies were resolved together in the classroom following individual feedback. After the completion of the evaluations carried out in the experimental group within a 13-week period, the post-test application was carried out and the post-test data of the research were reached.

2.4.2.1. Using the measurement tools and giving feedback. During the research process, the evaluation was carried out in the control group using the traditional question-answer method. The main purpose of the questions used in this group and the evaluation made is to reveal the reading success of the students. The assessment questions used did not focus on identifying the needs and learning deficiencies of the participants. The feedback given is limited to the exam scores and the correctness of the answers to the questions.

In the experimental group, the most basic element of the formative assessments was designed as the feedback given to the participants. After each evaluation process, feedback was provided for the needs of the participants. In the feedback given, no scoring or statements such as true or false were included. The main purpose of the questions used was to reveal the comprehension deficiencies in the text. As a result of the evaluations, the participants, who were thought not to understand enough what they read, were given information about their reading errors. For example, a participant who was thought to have a comprehension deficiency was asked to read the text aloud again, accompanied by a teacher. It was determined that the participant only focused on speed while reading and the participant was given information about how to take into account the units of meaning and how to perform prosodic reading as well.

During the research process, evaluations carried out in the experimental group by using SVM, multiple choice format of fill-in-the-blank technique, rewriting, retelling fluency, and story map techniques were completed within one course hour. Participants were asked to read the text once, and then assessments were made using measurement tools. Answers given by the participants to the measurement tools were read and their understanding deficiencies were revealed, and each participant was individually told which parts of the texts they lacked in comprehension. In this process, the texts were read again so that the students who did not answer the questions in the text or who had a lot of wrong answers could clearly see their own shortcomings as they were asked to respond to the relevant measurement tools again. In this direction, the aim was to encourage the participants to respond to the texts.

The participants were given general information about how to use SVM while the evaluations were carried out with the sentence verification technique in the experimental group. The participants were prepared for evaluation by informing them that the careful reading process would make it easier to find the answers to the questions to be asked. The texts previously prepared by the researchers were distributed to the participants. The participants were asked to choose true or false for the questions following the reading process; namely, they were asked to mark the column with "True" in case the sentence had the same meaning with the text, and to mark the column with "False" in case the sentence had a different meaning from the text. After the relevant instructions were given, the participants were told that they could start reading the text. Text reading was performed only once by each participant. SVM evaluation processes were completed by giving evaluation questions to the participants who completed the reading.

Before the assessments made with the multiple-choice format of the fill-in-the-blank technique, the measurement tool was introduced to the participants. Participants were told that a text would be given and they would read it only once. It was stated that some words of the text they would read would be removed from the form to be given after the reading process. It was also stated that these blanks would contain words in a multiple-choice form. It was added that they should choose, among these options, the same words as they appeared in the first text they had read. It was mentioned that only one word should be selected for each blank. After the general briefing, the text was distributed to the participants. The forms prepared for evaluation were distributed to the participants who had completed the text-reading process and their answers were received. After the answers were received, the evaluation processes carried out with the multiple-choice format of the fill-in-the-blank technique were completed.

In the evaluation process of the re-writing narration technique, the participants were expected to retell the text read in written form. In this process, texts were distributed to the participants and they were asked to read the texts once. After the texts were read, the participants were asked to write down everything they remembered about the text. Evaluation processes were completed by receiving the participants' re-writing narration responses.

While the evaluations were carried out with the retelling fluency technique, the number of words that the participants read during one-minute oral reading process was determined. Then, the participants were asked to verbally retell the text they read. In this process, it was measured how many words the participants used in one minute while telling the text they read. Evaluations were made by comparing the number of words they used during one-minute reading with the number of words they used during retelling. Participants who could not read enough words in one-minute period and those who explained what they read in very few words were recommended to do repetitive readings.

Story maps were used in the evaluation process. First of all, story map format was introduced to the participants. It was explained to the participants that each box was intended to identify the story elements in the text according to its title. After the texts to be used were distributed, the participants completed the reading process. The story map forms were distributed and the participants were expected to fill in the titles in the story map appropriately.

2.5. Data Analysis

The data of this study were obtained by scoring the comprehension questions and analyzing them in the SPSS program. In the data analysis process, firstly, the normality of the data was determined. Normality conditions were checked with the Shapiro-Wilk test. When the normality status of the pre-test and post-test scores of the experimental group participants was examined, it was found that the pre-test normality distribution score was S-W(22)=0.125, that is, p>.05, which shows the pre-test scores as normally distributed. When the post-test normality distribution conditions were examined, it was concluded that the post-test data were not normally distributed, with a value of S-W(22)=0.001, that is, *p*<.05. When the normality status of the pre-test and post-test scores of the control group is examined, the pre-test normality distribution is S-W(28)=0.04, *p*<.05. Considering the post-test normality, the result is observed as S-W(28)=0.175, *p*>.05. The tests to be carried out were decided according to the normality conditions. The Mann-Whitney U test was used to determine the equivalence status of the groups in the pre-test data of two independent groups, the normality of which could not be assured. Wilcoxon signed-rank test was used to compare the pre-test and post-test data of the control group and the experimental group within themselves and also to determine their significance. Finally, the Mann-Whitney U test was used to compare the post-test data of the control and experimental groups.

The $r = \frac{z}{\sqrt{N}}$ formula was used to determine the effect of the significance values that emerged as a result of the tests. According to this formula, as the effect value approaches zero, it can be mentioned that there is a low effect. As this value approaches 1, it can be interpreted that the effect increases (Green & Salkind, 2014).

3. FINDINGS

3.1. Findings Regarding the First Sub-Problem of the Research

Table 4 presents the findings that emerged as a result of the comparison of the pre-test and post-test scores of the experimental group participants with the Wilcoxon signed-rank test.

Table 4. *Wilcoxon Signed Ranks test results of the pre-test and post-test scores of the experimental group participants.*

Pre-Test / Post-Test		*N*	Mean *R.*	Total *R.*	*Z*	*p*
Comprehension scores	Negative rank	0	-	-	-3.934	0.000*
	Positive Rank	20	10.50	210.00		
	Equal	2	-	-		

*p<.05

In Table 4 it can be seen that there is no decline in the comprehension scores of any of the experimental group participants. In addition, there is no change in the comprehension scores of 2 students, while comprehension scores of 20 students increase. The mean score of the participants who show an increase is determined as 10.50. According to the results of the test, it can be concluded that there is a significant difference between the pre-test and post-test reading comprehension scores of the experimental group participants (z=-3.934, p<.05). When the effect value is calculated, the result r=-0.838 emerges, which reveals the significance value quite high.

3.2. Findings Regarding the Second Sub-Problem of the Research

Table 5 presents the findings obtained as a result of comparing the pre-test and post-test scores of the control group with the Wilcoxon Signed Ranks test.

Table 5. *Wilcoxon Signed Ranks test results of the pre-test and post-test scores of the experimental group participants.*

Pre-Test / Post-Test		*N*	Mean *R.*	Total *R.*	*Z*	*p*
Comprehension Scores	Negative rank	3	10.50	31.50	-2.152	0.031*
	Positive rank	14	8.68	121.50		
	Equal	11	-	-		

*p<.05

In Table 5, it can be seen that 3 participants experienced a decrease in their scores in the period between the pre-test and post-test and their average score of these participants was determined as 10.50. Although it is displayed in Table 5 that the scores of 11 students did not change, 14 students made progress between the pre-test and post-test processes and their mean scores reached 8.68. According to the test results, there is a significant difference between the pre-test and post-test scores of the control group participants (z=-2.152, p<.05). When the effect value is calculated, the result emerges as r=-0.390. With this result, it can be interpreted that there is a moderate significance value.

3.3. Findings Regarding the Third Sub-Problem of the Research

Table 6 shows the comparison of the reading comprehension scores of the experimental and control group participants with the Mann Whitney U test after the post-test.

Table 6. *Mann Whitney U test results of the post-test scores of the experimental and control group participants.*

Pre-Test		*N*	$\bar{x}$	*U*	*Z*	*p*
Comprehension Scores	Experimental	22	14.09	204.00	-2.064	0.039*
	Control	28	11.29			
	Total	50	12.52			

*p<.05

An analysis of Table 6 shows the arithmetic mean of the 22 participants in the experimental group as 14.09, while the arithmetic mean of 28 participants in the control group is 11.29. When the Mann Whitney U test result is examined, it is seen that the post-test reading comprehension scores of the experimental and control group participants differ significantly (U=-204.00, *p*<.05).

4. DISCUSSION and CONCLUSION

When the increase in the reading comprehension scores of the experimental group participants is evaluated individually, it is observed that the scores of only two participants did not increase, but the reading comprehension scores of the 20 participants in the experimental group improved. When considered in this context, it is seen that the formative assessment approach applied to the experimental group participants is an assessment process that is appropriate for the individual differences of the students and serves to meet the learning needs of many students. The features of effective feedback and increasing the individual time allocated to students appear as important elements in increasing the success of formative assessment in reading comprehension. Boumediene and Hamzaoui-Elachachi (2017) emphasize the conclusion that formative assessment interventions improve students' reading comprehension skills and they cite effective feedback and regular evaluation intervals as reasons for this improvement. Similar experimental studies by Gustafson et al. (2019), Hooley and Thorpe (2017) and Sanaeifar and Nafari (2018) support the current study and show that formative assessment practices improve reading comprehension. In our specific study it is observed in the control group that only half of the participants experienced improvement in their reading comprehension scores over a long period of 13 weeks, which may be the reason why the evaluation process, carried out with traditional methods, was done only to determine their development. The fact that assessment carried out in traditional methods fails to design teaching according to individual learning needs can be considered as the major reason why the reading comprehension scores of many students do not increase. As a result of this specific research, it is possible to reach the conclusion that reading comprehension skills of the experimental group participants, whose reading comprehension skills were evaluated with the formative assessment approach, improved.

Formative assessment is the process of supporting the participants by re-presenting the information according to their needs based on the data collected from the participants. In this process, teachers determine their learning goals and needs based on the information obtained before (Elden, 2019). At the same time, the fact that formative assessment is intertwined with the teaching process may cause teachers to be unaware of their assessment (Bredekamp, 2015). Reading comprehension skills are of vital importance for academic success and maintaining a quality social life. The increase in the reading comprehension success of the participants evaluated using the formative assessment approach will also affect their academic success. In his doctoral study, Ozan (2017) concluded that academic achievement increases when formative assessment practices are carried out. Considering reading comprehension as one of the most important conditions for academic success, this very study and Ozan's (2017) study point out similar results. In addition, there are many studies in the international literature stating that formative assessment positively affects academic achievement and learning (Alkharusi, 2008; Black & Wiliam, 2012; Black & McCormick, 2010; Chappuis & Chappuis, 2008; Chappuis et al., 2011; Choi et al., 2001; Clark, 2012; Fuchs & Fuchs, 1986; Gardner, 2012; Heitink et al., 2016; Herman et al., 2006; Kingston & Nash, 2011; McMillan, 2014; Peterson & Siadat, 2009). With the current research, it can be stated that formative assessment can be used to improve reading comprehension skills, which is a prerequisite for academic success.

When the formative assessment studies conducted in the national literature are examined, the strengths of formative assessment in various disciplines compared to traditional approaches have been revealed (Buluzun & Bulunuz, 2013; Doğan, 2016; Elden, 2019; İnaltun & Ateş, 2018; Metin & Özmen, 2010; Ozan, 2017; Zengin et al., 2017). This study also contributes to the national and international literature with the conclusion that when formative assessment activities are used instead of traditional assessment methods, reading comprehension skills will be positively affected. Kline (2013) concluded in his research that formative assessment contributes to secondary school students' reading success. As also seen in our specific study, it would be beneficial to use the formative assessment approach in order to develop and support reading skills.

Temizkan and Sallabaş (2011) reached the conclusion that multiple-choice tests are more successful than open-ended questions in their study in which they sought an answer to the question of whether multiple-choice tests or written exams are more effective in assessing reading comprehension skills. The most important reason for this is that multiple-choice tests eliminate the difficulty of expressing thoughts in writing (Temizkan & Sallabaş, 2011). However, if the questions are only used to perform measurements and not to identify learning deficiencies and learning goals, it will not be possible to go beyond the measurement process and as a result, a superficial measurement and evaluation application will emerge. In order to enrich the learning environment, determine learning goals, and provide effective feedback, multiple choice tests and written examinations, which are used without being included in the process, are insufficient. As can be seen in the results of this study, it was observed that the achievement scores of many students did not improve in the group that was assessed only by traditional methods and teachers evaluated their students' reading comprehension level with a result-oriented approach by using written exams. However, it is very important to pay attention to the individual differences of students for an effective reading and reading comprehension (Başaran, 2013). Formative assessment is the process that is used to identify student needs, organize, and improve education and is applied based on the interaction of student understanding (OECD, 2005). It is supported in this study that the formative assessment process implemented in this way will be more successful in determining the individual learning needs and goals of each student when compared to the success of the traditional methods.

According to Yıldırım (2012), questions are important tools for monitoring comprehension processes and the competence of teachers and students regarding questions is important for students to develop their reading comprehension skills. The emphasis should be on creating environments where students can talk about what they read for the development of high-level thinking skills (Applegate, 2007). However, according to the results of other research studies, questions are mostly used to reveal what students have learned (Ateş, 2011; Brown, 1991; Fordham, 2006; Hervey, 2006; Johnston, 1997; Knapp, 1995). Assessment type which seeks to reveal such learning situations and sees learning results rather than supporting learning is not suitable for formative assessment. As it can be interpreted from the results of our study, questions and evaluation should be included in the process and used to support learning.

In the national and international literature, there are studies comparing traditional assessment methods with assessment approaches in which students are actively involved in the process. Examining the results of these studies, it can be observed that assessment approaches that center the student in the process, such as formative assessment, are more beneficial in improving reading comprehension skills than the traditional approaches (Guthrie et al., 2006; Souvignier & Mokhlesgerami, 2006; Zipke, 2007). It can also be expressed that the results of the related studies and the results of this study are similar.

According to Roskos and Neuman (2012), the main features of formative assessment include identifying gaps between where students are and where they need to go in their reading development and it aims to create feedback loops that provide information about changes in performance gaps. It involves engaging students in meaningful and productive self-assessment process, developing a set of essential reading activities with clear criteria for success and building a culture for improving students' knowledge and skills. In such a reading climate, comprehension skills are expected to develop. In this specific research context, it is revealed that formative assessment activities should be used in education of reading and comprehension and in evaluation process by making use of the results of this research and the information in the relevant literature.

It can be suggested to use formative assessment methods that emphasize the process in the evaluation of reading studies. Formative assessment methods can be used in the evaluation of studies of reading texts within Turkish textbooks. It can be recommended that primary teachers offer effective feedback and corrections to students in reading and comprehension education. More comprehensive studies can be done on formative assessment and reading comprehension; namely, the effect of formative assessment on components such as fluent reading, reading motivation, and vocabulary can be examined; qualitative studies in which formative assessment and reading skills are handled together can be conducted; and similar studies can be carried out at different grade levels.

This research is limited to the formative assessment methods used in the process, measurement tools, and the participants of the research. Another limitation of the study is the 13-week quasi-experimental process. The possibility that the experimental and control group classroom teachers may have different qualifications and skills during the research process may also be a limitation, but the fact that the experimental and control group students had statistically equivalent comprehension scores in the pre-tests made at the beginning of the process is an indication that this situation was somewhat under control.

Acknowledgments

This research was produced from the master's thesis numbered 649617, titled "The Effect of Formative Assessment on Reading Comprehension" conducted by Muhammet Sönmez under the consultancy of Fatih Çetin Çetinkaya.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Duzce University, 05.11.2019, 2019-84.

Authorship Contribution Statement

Muhammet Sonmez: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Fatih Cetin Cetinkaya**: Methodology, Supervision, and Validation.

Orcid

Muhammet Sonmez https://orcid.org/0000-0001-6516-7635
Fatih Cetin Cetinkaya https://orcid.org/0000-0002-9843-6747

REFERENCES

Aktaş, N. (2015). *Okuma öncesi strateji öğretiminin 4. sınıf öğrencilerinin ekrandan okuduğu nu anlama düzeyine etkisi [The effect of pre-reading strategy instruction on reading on screen comprehension levels of elementary school 4th graders]* [Unpublished master's thesis]. Gazi University.

Akyol, H. (2006). *Türkçe öğretim yöntemleri [Turkish teaching methods].* Kök Yayıncılık.

Akyol, H. (2011). *Türkçe öğretim yöntemleri [Turkish teaching methods].* Pegem Akademi.

Akyol, H., & Ketenoğlu Kayabaşı, Z. E. (2018). Okuma güçlüğü yaşayan bir öğrencinin okuma becerilerinin geliştirilmesi: Bir eylem araştırması [Improving the reading skills of a students with reading difficulties: An action-research]. *Eğitim ve Bilim, 43*(193), 143-158. http://dx.doi.org/10.15390/EB.2018.7240

Akyol, H., Yıldırım K., Ateş, S., & Çetinkaya, Ç. (2013). Anlamaya yönelik ne tür sorular sorarız? *Mersin Üniversitesi Eğitim Fakültesi Dergisi* [What Kinds of Questions Do We Ask for Making Meaning?]. *9*(1), 41-56. https://dergipark.org.tr/tr/download/article-file/160845

Akyol, H., Yıldırım, K., Ateş, S., Çetinkaya, Ç., & Rasinski, T.V. (2014). *Okumayı değerlend irme: Öğretmenler için kolay ve pratik bir yol [Assessing reading: The easy and practic al way for teachers]*. Pegem Akademi,

Alkharusi, H. (2008). Effects of classroom assessment practices on students' achievement goals. *Educational Assessment,* 13(1), 243-266. https://doi.org/10.1080/10627190802602509

Applegate, M.D. (2007). Teacher's use of comprehension questioning to promote thoughtful literacy. *Journal of Reading Education, 32*(3), 12-19.

Ateş, S. (2011). *İlköğretim beşinci sınıf Türkçe dersi öğrenme-öğretme sürecinin anlama öğretimi açısından değerlendirilmesi [Evaluation of fifth-grade Turkish course learning and teaching process in terms of comprehension instruction]* [Unpublished doctoral dissertation]. Gazi University.

Ateş, S., & Akyol, H. (2013). Türkçe dersi öğrenme-öğretme sürecinin anlama öğretimi açısından değerlendirilmesi [The evaluation of Turkish language arts course with regard to comprehension instruction]. *Journal of Turkish Educational Sciences*, *11*(3), 268-300. https://dergipark.org.tr/tr/pub/tebd/issue/26091/274940

Aydemir, Y., & Çiftçi, Ö. (2008). Edebiyat öğretmeni adaylarının soru sorma becerileri üzerine bir araştırma (Gazi üniversitesi eğitim fakültesi örneği) [A research on asking question ability of literature teacher candidates (Gazi University education faculty pattern)]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 5*(2), 103-115. https://dergipark.org.tr/tr/pub/yyuefd/issue/13714/166035

Başaran, M. (2013). Okuduğunu anlamanın ölçülmesinde paragraftan anlam kurmaya dayalı çoktan seçmeli sorular [Measurement of reading comprehension using meaning-based paragraphs with multiple-choice questions]. *Eğitim Bilimleri Araştırmaları Dergisi, 3(*2), 107-121. http://dx.doi.org/10.12973/jesr.2013.327a

Baştuğ, M., & Keskin, H.K. (2011). Bilgi verici metin yapıları öğretiminin okuduğunu anlamaya etkisi [The effect of expository text structure on reading comprehension]. *E-Journal of New World Sciences Academy, 6*(4), 2598-2610. https://dergipark.org.tr/tr/pub/nwsaedu/issue/19818/212034

Black, P., & McCormick, R. (2010). Reflections and new directions. *Assessment & Evaluation in Higher Education, 35*(1), 493-499. https://doi.org/10.1080/02602938.2010.493696

Black, P., & Wiliam, D. (2012). *Developing a theory of formative assessment*. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 81-100). Sage.

Bormuth, J.R. (1963). Cloze as a measure of readability. *Proceedings of the lnternational Reading Association, 13*(1), 1-134. https://www.jstor.org/stable/1433978

Boumediene, A., & Hamzaoui-Elachachi, H. (2017). The effects of formative assessment on Algerian secondary school pupil's text comprehension. *AWEJ, 8*(3), 172-190. https://dx.doi.org/10.24093/awej/vol8no3.12

Bredekamp, S. (2015). *Erken çocukluk eğitiminde etkili uygulamalar* [*Effective Practices in Early Childhood Education].* (H. Z. İnan & T. İnan, Trans.). Nobel Yayıncılık.

Brown, R.G. (1991). *School of thoughts: How the politics of literacy shape thinking in the classroom*. Jossey Bass.

Bulunuz, M., & Bulunuz, N. (2013). Fen öğretiminde biçimlendirici değerlendirme ve etkili uygulama örneklerinin tanıtılması [Formative assessment in science teaching and demonstration of its effective implementation]. *Türk Fen Eğitimi Dergisi, 10*(4), 119-135.

Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2017). *Bilimsel araştırma yöntemleri [Scientific research methods]* (23.bs.). Pegem Akademi.

Çeliktürk Sezgin, Z., & Akyol, H. (2018). Kavram odaklı okuma öğretiminin ilkokul dördüncü sınıf öğrencilerinin okuma motivasyonuna ve okuduklarını anlamaya etkisi [Influences of concept-oriented reading instruction on reading motivation and reading comprehensi on of fourth graders]. *İlköğretim Online, 17*(2), 546-561. https://doi.org/10.17051/ilkonline.2018.418901

Chappuis, J., Stiggins, R.J., Chappuis, S., & Arter, J., (2011). *Classroom assessment for student learning: Doing it right-using it well* (2nd ed.). NJ: Merrill/Pearson.

Chappuis, S., & Chappuis, J. (2008). The best value in formative assessment. *Educational Leadership, 65*(4), 14-19.

Choi, K., Nam, J.H., & Lee, H. (2001). The effects of formative assessment with detailed feedback on students' science learning achievement and attitudes regarding formative assessment. *Science Educational International, 12*(2), 28-34.

Clark, I. (2012). Formative assessment: assessment is for self-regulated learning. *Educational Psychology Review, 24*(2), 205-249. https://doi.org/10.1007/s10648-011-9191-6

Clarke, S. (2012). *Active learning through formative assessment*. Hodder Education.

Coiro, J., & Dobler, E. (2007). Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the internet. *Reading Research Quarterly, 42*(2), 214-257. https://doi.org/10.1598/RRQ.42.2.2

Çöklü Özkan, E. (2018). Okuduğunu anlamada yaratıcı dramanın etkisi ve önemi [The effect and importance of creative drama on reading comprehension]. *Ana Dili Eğitimi Dergisi, 6*(2), 343-368. https://doi.org/10.16916/aded.399213

Creswell, J.W. (2014). *Araştırma deseni nicel, nitel ve karma yöntem yaklaşımları [Research design quantitative, qualitative and mixed method approaches]* (S. B. Demir Trans.). Eğiten Kitap.

Creswell, J.W., & Plano Clark, V.L. (2011). *Designing and conducting mixed methods research* (2.ed.). Sage.

Doğan, C.D. (2016). Biçimlendirici değerlendirmenin üniversite öğrencilerinin değerlendirme tercihleri üzerindeki etkisi: Bir ölçekleme çalışması [Effect of formative assessment on assessment preferences of the university students: A scaling study]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 16(*2), 413-431. https://doi.org/10.17240/aibuefd.2016.16.2-5000194935

Doğanay, A., & Yüce, S.G. (2010). Öğrencilerin düşünme becerilerinin geliştirilmesinde rehberli yardım: Bir öğretmenin sözel ifadelerinin analizine ilişkin durum çalışması [Scaffolding in improving students' thinking skills: A case study of the analysis of a teacher's verbal expressions]. *Kuram ve Uygulamada Eğitim Yönetimi, 2*(2), 185-214. https://dergipark.org.tr/tr/pub/kuey/issue/10334/126644

Dreyer, C., & Nel, C. (2003). Teaching reading strategies and reading comprehension within a technology-enhanced learning environment. *System, 31*(3), 349-365. https://doi.org/10.1016/S0346-251X(03)00047-2

Duman, N. (2006). *Hikaye haritasi yönteminin eğitilebilir zihinsel engelli öğrencilerin okuduğunu anlama becerileri üzerindeki etkisi [The effect of story mapping method on educable mentally retarded students' reading comprehension skills]* [Unpublished master's thesis]. Abant İzzet Baysal University.

Dupont, P. (2018). Assessing adolescent reading comprehension in a French middle school: performance and beliefs about knowledge. *Australian Journal of Teacher Education, 43*(7), 30-61. https://doi.org/10.14221/ajte.2018v43n7.3

Durukan, E. (2009). 7. sınıf Türkçe ders kitaplarındaki metinleri anlamaya yönelik sorular üzerine taksonomik bir inceleme [A taxonomic analysis on questions for understanding the texts in 7th grade Turkish textbooks]. *Milli Eğitim, 38*(181), 84-93.

Elden, A. (2019). *Okul öncesi öğretmenlerinin biçimlendirici değerlendirme uygulamaları [The formative assessment practices of early childhood teachers]* [Unpublished master's thesis]. Baskent University.

Fordham, N.W. (2006). Strategic questioning: What can you tell me about sharks? *Principal leadership, 7*(1), 33-37.

Fuchs, L.S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208. https://doi.org/10.1177/001440298605300301

Fuchs, L.S., Fuchs, D., & Hamlett, C.L. (1989). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research, 83*(2), 103−110. https://doi.org/10.1080/00220671.1989.10885938

Fuchs, L.S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28. https://doi.org/10.1177/074193258800900206

Gardner, J. (2012). Assessment for learning: A compelling conceptualization. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 197-204). Sage

Gersten, R., Fuchs, L.S., Williams, J.P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research, *Review of Educational Research, 71*(21), 279-320. https://doi.org/10.3102/00346543071002279

Glesne, C. (2015). *Becoming qualitative researchers: An introduction* (5th ed.). Pearson Education.

Good, R.H., & Kaminski, R.A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Institute for the Development of Educational Achievement.

Green, S.B., & Salkind, N.J. (2014). *Using SPSS for windows and macintosh: Analyzing and understanding data (7th ed.)* [Kindle DX version]. Retrieved from Amazon.com

Gustafson, S., Nordström, T., Andersson, U.B., Fälth, L., & Ingvar, M. (2019). Effects of a formative assessment system on early reading development. *Education, 140*(1), 17–27.

Guthrie, J.T., Wigfield, A., Humenick, N.M., Perenevich, K.C., Taboada, A., & Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *The Journal of Educational Research*, *99*(4), 232-246. https://doi.org/10.3200/JOER.99.4.232-246

Heitink, M.C., Van der Kleij, F.M., Veldkamp, B.P., Schildkamp, K., & Kippers, W.B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review,* 17, 50-62. https://doi.org/10.1016/j.edurev.2015.12.002

Herman, J., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices (CRESST Report No. 703)*. University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hervey, S. (2006). *Who asks the questions? Teaching PreK-8*, *37*, 68-69.

Hock, M., & Mellard, D. (2005). Reading comprehension strategies for adult literacy outcomes. *Journal of Adolescent & Adult Literacy, 49*(3), 192-200. https://doi.org/10.1598/JAAL.49.3.3

Hooley, D.S., & Thorpe, J. (2017). The effects of formative reading assessments closely linked to classroom texts on high school reading comprehension. *Education Tech Research Dev* 65, 1215–1238. https://doi.org/10.1007/s11423-017-9514-5

İlter, İ. (2018). Zayıf okuyucuların okuduğunu anlama becerilerinin geliştirilmesinde ana fikir belirleme becerisinin öğretimi [The instruction on identifying main ideas in improving the reading comprehension of poor readers]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 19*(2), 303-334. https://doi.org/10.21565/ozelegitimdergisi.315887

İnaltun, H., & Ateş, S. (2018). Fen bilimleri eğitiminde biçimlendirici değerlendirme: Literatür taraması [Formative assessment in science education: A literature review]. *GEFAD, 38(*2), 567-613. https://doi.org/10.17152/gefad.353975

Johnston, P.H. (1997). *Knowing literacy: Constructive literacy assessment*. ME: Stenhouse

Karasar, N. (2012). *Bilimsel araştırma yöntemi [Scientific research method]* (24.bs.). Nobel Akademik Yayıncılık.

Karasu, H.P., Girgin, Ü., & Uzuner, Y. (2011). Okuma becerilerini değerlendirmede formal olmayan okuma yöntemlerinin kullanımı [Utilizing informal reading inventories on eval uation of reading skills], *Eğitim Teknolojisi Kuram ve Uygulama, 1*(1), 108-124. https://dergipark.org.tr/tr/pub/etku/issue/6274/84243

Keeley, P.D. (2008). *Science formative assessment: 75 practical strategies for linking assess ment, instruction, and learning.* Corwin & NSTA Press.

Keskin, H.K., & Akıllı, M. (2013). Fen ve teknoloji ders kitaplarının okunabilirliğinin farklılaştırılmış boşluk doldurma teknikleriyle ölçülmesi [An assessment of the readability of science and technology textbooks through differentiated cloze tests]. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 13*(27), 47-66. https://dergipark.org.tr/tr/pub/maeuefd/issue/19400/206159

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for resear ch. Educational Measurement: *Issues and Practice, 30*(4), 28-37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kline, A.J. (2013). *Effects of formative assessment on middle school student achievement in mathematics and reading* [Unpublished doctoral dissertation]. North Carolina Universit y.

Knapp, M.S. (1995). *Teaching for meaning in high-poverty classrooms.* Teachers College Pre ss.

Kocaarslan, M. (2015). *Zihinsel imaj oluşturma öğretiminin 4. sınıf öğrencilerinin okuduğunu anlama becerilerini geliştirmeye etkisi [The impact of teaching mental imagery on reading comprehension skills of 4th graders]* [Unpublished doctoral dissertation]. Gazi University.

Kodan, H. (2015). *Koro, tekrarlı ve yardımlı okuma yöntemlerinin zayıf okuyucuların okuma ve anlama becerileri üzerine etkisi [The effects of the methods of choral, repeated and assisted reading on the reading and reading comprehension skills of poor readers]* [Unpublished doctoral dissertation]. Gazi University.

Kuşdemir, Y. (2014). *Doğrudan Öğretim Modeli'nin ilkokul dördüncü sınıf öğrencilerinin okuduğunu anlama becerilerine etkisi [The effect of Direct Instruction Model on reading comprehension skills of elementary school fourth graders]* [Unpublished doctoral dissertation]. Gazi University.

Landauer, T.K., Lochbaum, E.K., & Dooley, S. (2009). A new formative assessment technology for reading and writing, *Theory into Practice, 48*(1), 44-52, https://doi.org/10.1080/00405840802577593

Leslie, L., & Caldwell, J. (2006). *Qualitative reading inventory-4* (4th ed.). Allyn & Bacon.

Li, H. (2016). How is formative assessment related to students' reading achievement? Findings from PISA 2009. *Assessment in Education: Principles, Policy & Practice, 23*(4), 473-494. https://doi.org/10.1080/0969594X.2016.1139543

Marchand, G.C., & Furrer, C.J. (2014). Formative, informative, and summative assessment: The relationship among curriculum-based measurement of reading, classroom engagement, and reading performance. *Psychology in the Schools, 51*(7), 659-676. https://doi.org/10.1002/pits.21779

Marcotte, A.M., & Hintze, J.M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47*(5), 315-335. https://doi.org/10.1016/j.jsp.2009.04.003

Mathes, P.G., & Fuchs, D. (1997). Cooperative story mapping. *Remedial & Special Education, 18*(1), 20-27

May, F.B., & Rizzardi, L. (2002). *Reading as communication*. Prentice Hall.

McMillan, J.H. (2014). *Classroom assessment: Principles and practice for effective standards-based instruction* (5th ed.). Pearson.

Metin, M., & Özmen, H. (2010). Biçimlendirici değerlendirmeye yönelik öğretmen adaylarının düşünceleri [Prospective teachers' views about formative assessment]. *Milli Eğitim, 40*(187), 293-310. https://dergipark.org.tr/tr/pub/milliegitim/issue/36197/407045

Ministry of National Education (2019). *Türkçe dersi (ilkokul ve ortaokul 1, 2, 3, 4, 5, 6, 7, 8. sınıflar) öğretim program* [*Turkish lesson (primary and secondary school 1, 2, 3, 4, 5, 6, 7, 8th grades) curriculum*]. http://mufredat.meb.gov.tr/

Ministry of National Education (MoNE). (2005). *PISA 2003 projesi ulusal nihaî rapor* [*PISA 2003 survey national final report*]. Milli Eğitim Basımevi.

Ministry of National Education (MoNE). (2010). *PISA 2006 projesi ulusal nihaî rapor* [*PISA 2006 survey national final report*]. http://pisa.meb.gov.tr

Ministry of National Education (MoNE). (2013). *PISA 2012 ulusal ön raporu* [*PISA 2012 National Preliminary Report*]. http://pisa.meb.gov.tr

Ministry of National Education (MoNE) (2015). *PISA 2012 araştırması ulusal nihai rapor* [*PISA 2012 survey national final report*]. http://pisa.meb.gov.tr

Ministry of National Education (MoNE). (2019). *PISA 2018 Türkiye ön raporu* [*PISA 2018 National Preliminary Report*], Eğitim Analiz ve Değerlendirme Raporları Serisi.

Ness, M. (2011). Explicit reading comprehension instruction in elementary classrooms: Teacher use of reading comprehension strategies, *Journal of Research in Childhood Education, 25*(1), 98-117. https://doi.org/10.1080/02568543.2010.531076

Neuman, S.B. (2001). The role of knowledge in early literacy. *Reading Research Quarterly, 36*(4), 468-475. https://doi.org/10.1598/RRQ.36.4.6

Offerdahl, E.H., & Montplaisir, L. (2013). Student-generated reading questions: Diagnosining student thinking with diverse formative assessments. *Biochemistry and Molecular Biology Education, 42*(1) 29-38. https://doi.org/10.1002/bmb.20757

Oosterhof, A., Conrad, R.M., & Ely, D.P. (2008). *Assessing learners online*. Pearson.

Organisation for Economic Co-operation and Development. (OECD). (2005). *Formative assessment: Improving learning in secondary classrooms.* OECD.

Ozan, C. (2017). *Biçimlendirici değerlendirmenin öğrencilerin akademik başarı, tutum ve öz düzenleme becerilerine etkisi [The effects of formative assessment to students' academic achievement, attitude and self-regulation skills]* [Unpublished doctoral dissertation]. Atatürk University.

Öztürk, M. (2019). *Kelime duvarı yönteminin ilkokul 4. sınıf öğrencilerinin akıcı okuma ve okuduğunu anlama becerilerine etkisi [The effect of word wall method on 4th grade students' fluent reading and reading comprehension skills]* [Unpublished master's thesis]. Bolu Abant İzzet Baysal University.

Papatğa, E. (2016). *Okuduğunu anlama becerilerinin SCRATCH program aracılığıyla geliştir ilmesi [Developing reading comprehension skills through SCRATCH program]* [Unpublished doctoral dissertation]. Atatürk University.

Pearson, P.D., & Duke N.K. (2002). Comprehension instruction in the primary grades. In Cathy Collins Block & Sheri R. Parris (eds.), *Comprehension instruction: research-based best practices* (pp, 247-258). The Guildford.

Piazza, S.V. (2012). Cultural responsiveness and formative reading assessment: Retellings, comprehension questions, and student interviews. *Language and Literacy, 14*(3), 133-149.

Rankin, E.F., & Culhane, J.W. (1969). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading, 13*(3), 193-198

Reutzel, D.R., & Cooter, R.B. (2007). *Strategies for reading assessment and instruction: Helping every child succeed* (3rd ed.) Pearson Education, Inc.

Rogoff, B., Turkanis, C.G., & Bartlett, L. (2001). *Learning together: Children and adults in a school community*. Oxford University Press.

Roskos, K., & Neuman, S.B. (2012). Formative assessment: Simple, no additives. *Reading Teacher, 65*(8), 534-538. https://doi.org/10.1002/TRTR.01079

Royer, J.M. (2001). Developing reading and listening comprehension test based on the sentence verification technique (STW). *Journal of Adolescent & Adult Literacy, 45*(1), 30-41

Royer, J.M., Hastings, C.N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior, 11*(4), 355–363.

Sanaeifar, S.H., & Nafari, F. (2018). The effects of formative and dynamic assessments of reading comprehensions on intermediate EFL learners' test anxiety. *Theory and Practice in Language Studies, 8(*5), 533-540. http://dx.doi.org/10.17507/tpls.0805.12

Scarborough, H. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 97–110). Guilford

Shaughnessy, M.F. (2005). An interview with James M. Royer about reading and comprehension. *Educational Psychology Review, 17*(3), 273-283.

Sidekli, S. (2010). Eylem araştırması: İlköğretim dördüncü sınıf öğrencilerinin okuma ve anlama güçlüklerinin giderilmesi [An action-research: Correcting the fourth-grade students' reading and comprehension problems]. *TÜBAR*, *27*(1), 563-580. https://dergipark.org.tr/tr/pub/tubar/issue/16968/177248

Souvignier, E., & Mokhlesgerami, J. (2006). Using self-regulation as a framework for implementing strategy instruction to foster reading comprehension. *Learning and Instruction, 15*(2), 57-71. https://doi.org/10.1016/j.learninstruc.2005.12.006

Sözen, N., & Akyol, H. (2018). Rehberli okuma yöntemi: Bir eylem araştırması [Guided reading: An action-research]. *Turkish Studies, 13*(19), 1633-1658. http://dx.doi.org/10.7827/TurkishStudies.13812

Stiggins, R.J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765. https://doi.org/10.1177/003172170208301010

Temizkan, M., & Sallabaş, M.E. (2011) Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması [Comparison of multiple-choice tests and open-ended written exams in the evaluation of reading comprehension skills]. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi. 30(1)*, 207-220. https://dergipark.org.tr/tr/pub/dpusbe/issue/4772/65689

Ulusoy, M. (2009). Boşluk tamamlama testinin okuma düzeyini ve okunabilirliği ölçmede kullanılması [Using cloze test to measure students' reading levels and readability of texts]. *Türk Eğitim Bilimleri Dergisi, 7*(1), 105-126. https://dergipark.org.tr/tr/pub/tebd/issue/26140/275303

Ulusoy, M., & Çetinkaya, Ç. (2012). Cümle doğrulama tekniğinin okuma ve dinlemenin ölçülmesinde kullanılması [The use of sentence verification technique for measuring reading and listening]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H.U Journal of Education), 43*,460-471.

Vonderwell, S., Liang, X., & Alderman, K. (2007). Asynchronous discussions and assessment in online learning. *Asynchronous Discussions and Assessment in Online Learning, 39*(3), 309–328.

Wiliam, D. (2006). Formative assessment: Getting the focus right, *Educational Assessment, 11*(3-4), 283-289. https://doi.org/10.1080/10627197.2006.9652993

Yazıcı, E., & Kurudayıoğlu, M. (2017). 5. sınıf Türkçe ders kitaplarındaki dinleme metinlerinin öğrencilerin seviyesine uygunluğunun incelenmesi [Examination of the appropriateness of listening texts in the 5th grade of Turkish textbooks to students' level]. *Anadili Eğitim Dergisi, 5*(4), 967-984. https://doi.org/10.16916/aded.340839

Yıldırım, K. (2012). Öğretmen ve öğrencilerin okuduğunu anlama becerilerini değerlendirmede kullanacakları bir sistem: Barrett taksonomisi [A system to be used by teachers to evaluate students' reading comprehension skills: Barrett taxonomy]. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9(*18), 45-58. https://dergipark.org.tr/tr/pub/mkusbed/issue/19552/208323

Zengin, Y., Bars, M., & Şimşek, Ö. (2017). Matematik öğretiminin biçimlendirici değerlendir me sürecinde Kahoot! ve Plickers uygulamalarının incelenmesi [Investigation of using Kahoot! and Plickers in formative evaluation process in mathematics teaching]. *Ege Eğitim Dergisi, 18*(2), 602- 626. https://doi.org/10.12984/egeefd.318647

Zipke, M. (2007). *Metalinguistic instruction improves third graders' reading comprehension* [Unpublished doctoral dissertation]. The City University of New York.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 109–125

https://doi.org/10.21449/ijate.1132981

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of this article.

The Use of open-ended items for giving feedback during the formative assessment process

Ozge Altintas[1,*]

[1]Ankara University, Faculty of Educational Sciences, Department of Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY

Received: June 20, 2022

Revised: Oct. 3, 2022

Accepted: Oct. 4, 2022

Keywords:
Student achievement,
Formative assessment,
Feedback,
Open-ended items,
Classroom assessment,
Scoring rubric.

Abstract: Feedback plays an important role in classroom learning and teaching process. This study focuses on how feedback can be more effectively used in the formative assessment process. According to this purpose, the study first discusses the concept of student achievement and presents its changing nature in the 21st century. Subsequently, the study addresses higher-order thinking skills, the use of open-ended items in improving student achievement in the classroom, rubrics, formative assessment, and feedback. The study aims to present an exemplary measurement and assessment model that will contribute to the development of student achievement. Additionally, it examines the use of a feedback approach that will improve the power of using the knowledge of the students learned in lessons in daily life by associating it with basic life skills in the formative assessment process. Accordingly, teachers are provided with a unique means that they can easily use in improving classroom success. In the study, an open-ended item has been developed that has a real-life counterpart is used to provide information on the improving of student achievement, while a rubric used in scoring the answers to the item is also developed. The answer categories in the rubric show the place of the student in the distribution of success. Thus, teachers will be able to see what students can and cannot do, as well as be able to give accurate and realistic feedback on what needs to be done for the development of student achievement.

1. INTRODUCTION

1.1. The Concept of Student Achievement and Change in the Understanding of Student Achievement

The word “success” is frequently used in daily life to refer to individuals’ work and professional lives, academic careers, financial gains, and private lives. Gerberich et al. (1962) define success as the work undertaken in a planned and programmed manner to attain a desired result in line with set goals. Comparatively, Wolman (1973) defines success as progress made toward achieving a desired result.

When school learnings are taken into consideration, the level of students learning basic information and using what they have learned in new situations is used as academic

*Corresponding Author: Ozge ALTINTAS ✉ oaltintas@ankara.edu.tr Ankara University, Faculty of Educational Sciences, Department of Educational Measurement and Evaluation, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

achievement, unlike the concept of success. Haladyna (1997) states that the concept of achievement includes students' learning at the knowledge level (understanding the course content) and at the skill level (using the understood knowledge in practice). Koç (1978) handles the concept of success in terms of school achievement, and refers to it as the progress that the student has made in achieving those results determined by their school, class, and course.

The main aim of school education in the current century is to enable individuals to transfer the basic knowledge and skills they have learned at school to real life (Brookhart, 2010, 2014; Marzano, 1992; Nitko, 2001; Popham, 2000). Achievement in this sense, is defined as the development of those high-level thinking skills that enable students to use what they have learned in real-life situations (Haladyna, 1997; Kutlu et al., 2017).

In the 21st century, when information and technology are widely used in every field, individuals are expected to adapt to social change rapidly and may even become the initiating force of new changes. According to Aslanoğlu (2022), an education system in the 21st century is expected to incorporate students' needs in order to help them become productive and efficient users of technology, improve their critical thinking, and make them independent, autonomous, and lifelong learners.

All educational institutions have responsibilities for individuals to educate with these characteristics. Considering the socioeconomic, technological, and cultural changes in current social life, there is a need for a more inclusive definition of achievement within education. In the comprehensive report *What Matters to Student Success: A Review of the Literature*, Kuh et al. (2006) describe student success as "academic achievement, engagement in educationally purposeful activities, satisfaction, acquisition of desired knowledge, skills and competencies, persistence, attainment of educational outcomes, and post-college performance". York et al. (2015) update this definition and present a conceptual model of academic success, and define the concept based on their findings as being inclusive of "academic achievement, attainment of learning objectives, acquisition of desired skills and competencies, satisfaction, persistence, and post-college performance".

Haladyna (1997) and Kutlu et al. (2017) emphasize that student achievement should be considered in relation to an individual's mental development. Haladyna (1997) defines mental development into three dimensions: knowledge, skill, and ability. Knowledge refers to the recall or understanding of course content and skills the use of remembered and understood information for practice. Knowledge and skills are developed in short periods, such as lessons, a unit, or semester, and in most cases do not change from individual to individual until proven otherwise. Abilities, on the other hand, are high-level mental structures that can be developed throughout life.

1.2. Use of Open-Ended Items in Monitoring of Classroom Learning Achievement

This study discussed student achievement within the context of developing higher-order cognitive skills. Kutlu and Altıntaş (2021) define student achievement as the power of students to use their cognitive, intrapersonal, and interpersonal skills in realistic situations. The relevant literature describes student achievement as higher-order thinking skills whereby individuals associate different sets of information with one another (Brookhart, 2010; Haladyna, 1997; Kutlu et al., 2017; Marzano, 1992; Popham, 2000).

It is important for teachers to conduct classroom assessment and evaluation practices using open-ended items based on real-life situations that already exist in students' knowledge and experience. Brookhart (2014) describes open-ended items as those having multiple correct answers or that include multiple solutions. In this sense, teachers should pay attention to whether the topics they address in the lesson contain more than one answer in writing open-ended items. Popham (2000) emphasizes that the use of open-ended items is inevitable, especially in the measurement of certain characteristics that require originality such as problem-solving processes, writing skills, and data organization. Karakaya and Şata (2022) state that open-ended items in classroom exams require participants to respond freely, that different solutions are expected to be compared, and that these items should be preferred when focusing on higher-order thinking skills such as solving problems with multiple solutions.

The study conducted by Kintsch and Yarbrough (1982) reported that open-ended items provided more information about learning. Similarly, Kutlu (2004) states that in cases in which open-ended items are well structured, it is possible to determine whether a student can use multiple skills concurrently using a single item. Further, it should be kept in mind that open-ended items reveal the desired results, relying on new sample situations that may not be explicitly addressed in the classroom. According to Reiner *et al.* (2002), answers given in open-ended items should be formed rather than chosen, and it should be ensured that the answers comprise of at least a few sentences.

Nitko (2001) notes that the items in those books prepared for teachers and students have an easy structure that does not require considerable thinking. The researcher also refers to "structured and unstructured" problem situations whereby teachers should address in classroom assessment practices. Accordingly, *structured items* are very similar to those taught in the classroom (ordinary); however, *unstructured items* have a unique (unusual) structure. Students can understand and edit these unstructured items with what they have learned in the course, and they can see that these items may have multiple correct answers. Another aspect of open-ended items is that students must answer them in writing using their own power of expression based on the basic knowledge and skills they have acquired in the course. Such items should be designed in a way so that their answers must be given using consists of at least a few sentences. Here, it should be clear that, if an answer comprises a single one word, a few words, or a short sentence, then this will provide insufficient information as to whether the student can use the relevant skill in realistic situations.

Another point that teachers should pay attention to when writing open-ended items is the cognitive level at which the student is made to think. Studies in the literature point out that is the extent to which students use the information provided in textbooks by associating them with basic skills is more important than how much information they know (Airasian, 2001; Haladyna, 1997; Kutlu et al., 2017; McMillan, 2007; Popham, 2000, 2008). Therefore, the items should be related to the content of the course as well as the sub-dimensions of structures that require higher-order thinking such as problem-solving, analytical thinking, reasoning, and critical thinking. Thorndike (2005) emphasizes that mental processes related to the situation that is to be assessed should be well known before the responder starts to write their answers to open-ended items. He also suggests that a new material should be used at the root of the item during tis preparation, and that students should be presented with a different material than that which is taught to them in the classroom or in the textbook so that they can reproduce their knowledge.

In order for students to activate their higher-order thinking skills, they need to use the basic knowledge they have acquired and to transform it when responding to the open-ended item (Brookhart, 2010). At this point, teachers should attach as much importance to formative assessment practices that will reveal the level of acquisition of relevant life skills as they do to classroom teaching activities. Studies conducted since the end of the 20th century have revealed that means of measurement and assessment guides classroom learning. Previous studies have shown that learning success and quality increase when teachers use assessment and assessment activities correctly (Biggs & Watkins, Black & William, 1998, 2002; 1996; Clarke, 2001). Some studies have indicated that formative assessment has a positive effect on learning and teaching processes (Crooks, 1988; Harlen, 2003; Stiggins & Conklin, 1992; Torrance & Pryor, 1998).

1.3. Formative Assessment and Feedback

The learning gains defined in curricula for determining student achievement should focus on mental skills and should be associated with life skills. In addition, the emphasis on higher-order thinking skills such as doing research, questioning, critical thinking, problem-solving, etc. necessitates formative assessment in improving student achievement in the classroom. Stiggins (1994) defines formative assessment as a continuous process that aims to improve education.

Especially since the 1960s, school programs have given more attention to improving student achievement. Bloom *et al.* (1971) refer to the improvement of student learning with the concepts of *formative evaluation* and *summative evaluation*. Summative evaluation focuses on assessing the level of student learning at various stages or at the end of the teaching process, and is mostly used for grading purposes. However, according to the formative evaluation approach, it is determined whether the students have the cognitive input (prerequisite) behaviors required for the learning process (at the beginning); learning deficiencies and difficulties are determined the end of the process. This process focuses on shaping student learning and assessing whether students have sufficiently learned the topics covered rather than on grading students.

Özçelik (2014) describes the monitoring of learning as "determining which behaviors expected to be learned in a unit have been learned, which ones have not been learned, and probably why they have not been learned at the end of each unit and completing the learning deficiencies in a timely manner by considering possible difficulties". He also emphasizes the use of formative tests to assess all those new behaviors that are expected to be learned in the unit in order to identify learning deficiencies, as well as possible difficulties leading to these deficiencies.

While students' learning regarding the level of knowledge (remembering) takes place in a short period, sometimes as short as a few class hours, the process of learning skills that enable them to use the same knowledge can take months or even years (Haladyna, 1997; Kutlu et al., 2017). Therefore, it is more important to monitor the skill, not form it. Monitoring skills that develop over a long period at critical stages and giving accurate and timely feedback to the student will contribute to the adoption of such skills at the desired level of competence. In this way, students can become aware of their strengths and weaknesses while using skills. For this reason, in this study the concept of *formative assessment* was used in the meaning of *monitoring-based assessment* in this study, and the conceptualization of monitoring learning by Özçelik (2014) was developed and enriched.

Stiggins (2002; 2005) emphasizes that it is necessary to move from the understanding of *the assessment of learning* to *the assessment for learning* in school education. Kutlu and Kula-Kartal (2018) state that the understanding of assessment for learning involves more-than-frequent testing of students and that, according to this understanding, assessment and teaching process should proceed in an intertwined manner Within this process students are not expected to perform better than other students, but are expected to focus on becoming competent in the knowledge and skills they are learning.

The key point in terms of classroom assessment is to focus on the process rather than the learning outcome and to internalize an approach that prioritizes feedback. Therefore, in order for students to reach the desired learning outcome, a monitoring-based/formative assessment approach where the process is kept under control comes to the fore, rather than a level of determination at the end of the teaching process. It is important to structure those assessments that are to be carried out during the process in a way that they are both interrelated whereby they both provide rich feedback on learning.

According to Harlen (2007), feedback given to the students during the learning process helps them to organize their learning. Bloom (1976) draws attention to an effective teaching service for students to reach mastery learning, and emphasizing four elements that affect the quality of this service: *pointing and explanation, participation, enrichment, feedback, and correctness*. Bloom particularly emphasizes *feedback and correctness*; this is because feedback helps students determine their performance expectations, evaluate their level of understanding, and recognize their misconceptions. According to De Cecco (1968), feedback involves comparing student achievement with a standardized measure of achievement and informing the student of the result. In addition, feedback can give clues as to which approach can contribute to correcting students' mistakes observed in the learning process, thereby increasing their success (Attali & Powers, 2009).

Feedback is a fundamental construct for many learning and teaching theories. Understanding the conditions for effective feedback should facilitate both theoretical development and teaching practices (Bangert-Drowns et al., 1991). Kulhavy and Stock (1989) state that providing feedback based on a task is most commonly applied psychological interventions that support student achievement. A comprehensive literature review on formative feedback by Shute (2008) shows that the basic premise underlying most of the research on the subject: "only when given correctly, feedback can significantly improve learning processes and outcomes".

Gedye (2010) suggests some structures and tools that can facilitate formative feedback. Within the scope of the present study, some suggestions for teachers are given below:

- Use portfolios that allow students' need for their self-reflection.
- Have students rearrange their work after giving feedback on their draft work.
- Involve students in the process of creating assessment criteria.
- Ask students to identify the strengths and weaknesses of their work regarding established assessment criteria before submitting their work.
- Use examples to help students understand the expected standards.
- Take time to discuss and reflect on criteria and standards in the classroom.
- Before students leave the classroom, have students make a list of how they will work with an action plan based on the feedback they receive.
- Ask students the types of feedback they find most helpful and ask them to explain strategies they would follow to improve their success.

An assessment approach supported by formative feedback should aim to associate knowledge students learn using key skills. Kulhavy et al. (1976) and Kulhavy (1977) argue that feedback that does not emphasize skills cannot go beyond identifying learning deficits. Therefore, it would be appropriate for feedback to be given in formative assessment to focus on the learning gains of the courses.

Kutlu et al. (2010) emphasize that teachers should consider both the *content (scope)* and *cognitive level* of the lesson when writing items based on the learning gains of the lesson. For this reason, teachers should give feedback by making determinations based on the following two situations in student responses: first, determine whether the student has learned the information about the course content and reveal what they have learned at the desired level; second, determine whether the skill representing the cognitive level is used at the desired level in the case and situation. One of the effective elements that plays a role in determining these two situations is the *item* used, while the other is the *rubric*. Cutting and Scarborough (2006) state that determining how well an individual learns depends on how well it is measured.

Scoring rubrics show those defined criteria within which a student response or task falls, and each criterion shows the transition from competent to weak levels of achievement according to that task (Goodrich, 1997). Popham (2000) expresses the rubric as a reference scoring key that is used to evaluate the quality of student answers; however, Kutlu et al. (2017) note that a rubric is a scoring tool that shows according to which criteria a student's work is evaluated and to which level their performance will correspond. All these definitions indicate that rubrics provide detailed information about students' achievements from high to low levels. In this sense such rubrics are important in terms of drawing attention to what the students can do while scoring their answers to an item, or to what the student produces in responding or completing a performance task. In addition, learnings that each student scores regarding the rubrics correspond to the provision of important feedback to teachers about these students (Kutlu, 2004).

Miller *et al.* (2008) emphasize the reliability of the scores obtained from rubrics so that they can be used in the decision-making process. Kutlu et al. (2019) suggest that to ensure that rubrics provide reliable results teachers should examine the statements in the rubric and those answer categories to which the statements belong after the rubric has been prepared. The fact that teachers examine the response categories individually before scoring and receive opinions from other teachers in the relevant field may contribute to the reliability and validity of the scoring. McMillan (2007) emphasizes the advantage of creating a rubric before the administration of open-ended items and draws attention to the importance of preparing rubrics that include common criteria for scoring all answers.

1.4. Significance of the Study

Concepts of *formative evaluation* and *summative evaluation* introduced by Bloom et al. (1971) have played a role in the education systems of many countries, including Turkey. The concept of formative evaluation continues to be influential today and is widely used to overcome students' learning deficiencies and difficulties. This approach sees students as passive learners rather than active learners within the teaching and learning process, and focuses on what students have not learned rather than on what they have learned, as well as on their learning level in each case. Therefore, this approach mostly depends on content and repetition of content; additionally, it aims at shaping student learning in accordance with the content rather than observing the progress regarding student achievement.

From this point of view, it would not be wrong to argue that the *formative evaluation* approach has not been effective in developing the expected learning achievement in school education. However, since the last quarter of the 20th century, and especially since the 21st century, many societies have expected schools to educate students with higher-order thinking skills such as problem-solving, analytical thinking, reasoning, and critical thinking. It has been considered important to observe these skills, which develop over a long period of time, and to provide students with information about their mental strengths rather than their deficiencies. This study aims to explain to teachers how they can give feedback by using open-ended items in the formative assessment process and how they can improve student achievement by proposing a sample model.

2. METHOD

2.1. Procedure

It is emphasized in the previous sections of this study that student achievement can be considered as students' power to use the basic knowledge acquired in the courses to real-life situations. Accordingly, a sample open-ended item that can be used effectively while giving feedback in the process of formative assessment that has a counterpart in real-life, and a sample holistic rubric showing the scoring method of the aforementioned item were developed. It is suggested that the developed example should be examined by combining it with the example presented in the study of Kutlu and Altıntaş (2021). Figure 1 shows the flowchart of the feedback that can be given in the formative assessment process.

Figure 1. *Flow chart of the feedback process.*

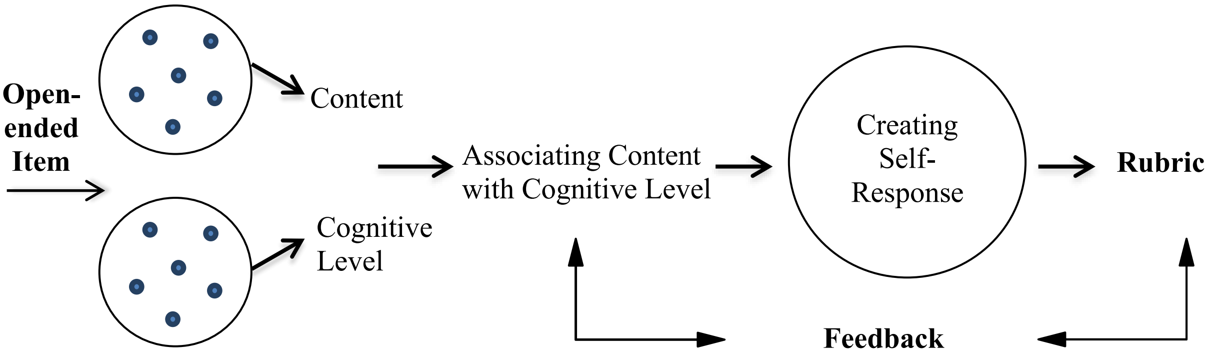


Figure 1 shows that students are expected to associate two dimensions with one another in order to give the expected answer to the feature measured by the open-ended item. As indicated in Figure 1, the student's ability to make this association depends on them having sufficiently learned the basic information about the content covered in the lesson, as well as them having acquired the skills of the relevant cognitive level in which they combine this information. For this reason, teachers should use case studies based on life situations that require the use of more than one piece of knowledge depending on the content in the feedback process that aims to improve students' learning achievement. Only in this case will teachers be able to determine both the extent to which the information is learned in the students' responses, as well as the extent to which cognitive skills will enable them to associate and use the knowledge in case studies based on real-life situations.

It is clear that students' responses will differ in the process wherein students can create their own answers. For this reason, it is important to determine those answers that are completely correct, partially correct, incorrect, or even unrelated, as well as those left blank in terms of observing the student and improving their success. Here, there are two dimensions that enrich the feedback: *the quality of the open-ended item* and *the fact that the rubric has been prepared with the expected competence*. The answer to an open-ended item usually requires writing one or more sentences. It is important that the answer is associated with one of the answer levels in the rubric in an unbiased way. Therefore, the accuracy and quality of the answer should be evaluated by a knowledgeable and talented teacher (Reiner *et al.*, 2002). To summarize, both the written item and rubric should be developed with certain accuracy and quality, and the answers should be scored by teachers who are equipped to use the rubric.

2.2. Sample Study

In order to facilitate the use of the explanations made in the previous chapters in classroom practices, *the model of giving feedback in the formative assessment process*, which is suggested in this study, is discussed through a sample study below. The learning outcome considered for the open-ended item is related to that of "the 6th grade Social Studies lesson which argues that solutions to a problem should be based on rights, responsibilities, and freedoms". The sample open-ended item consists of two parts: "situation" and "instruction". The situation part includes the problem that encourages the use of the basic knowledge learned in the lesson, makes the students think about the problem, and is as realistic as possible. Comparatively, the instruction part is that which asks the question depending on the situation. The instruction should be relevant to the situation, appropriate to the student's grade level, and should be clear and understandable.

One of the most important subjects of Social Studies lessons, which comes to the fore in interpersonal relations and daily life, is the relationship between the individual and society. The individuals' ability to lead a happy and peaceful life in society is related to them knowing the rights, responsibilities, and freedoms of both themselves and other individuals around them.

The concept of right refers to the authority of an individual to do something within the framework of certain rules and limits. Rights are also legally granted entitlements that have been given to individuals. The concept of responsibility refers to what must be fulfilled during when an individual exercises their rights; they are concerned with an individual bearing the consequences of what they do, must do, and actions they have undertaken and for which they are sometimes necessarily held accountable. Finally, freedom is the ability of an individual to do what they want without restricting other the freedoms of other people within certain limits; they are concerned with individuals making decisions according to their wishes and thoughts, independent of external influences (Şahin, 2019; TDK, n.d.).

These three basic concepts (right, responsibility, and freedom) need to be acquired by individuals at the relevant grade levels for the continuation of social life. These concepts should not only be taught to students at the descriptive level but also at the level of establishing their relationship with one another. If these skills are not adopted, necessary cooperation and solidarity among individuals living in society are not ensured, conflicts may arise between individuals, and there can be disintegration and dissolution in society in the future.

Table 1 shows an example of a formative assessment that reveals at which level life skills related to this important social issue should be addressed regarding the acquisition of classroom teaching activities. In the example shown in the Table, first the grade level, the learning field that constitutes the content and the achievement; then the cognitive level; and then the scoring method of the item were defined. The study by Kutlu (2004) was used when developing the writing style for each item. Care was given to ensure that the item was appropriate for the learning outcome, cognitive level, grade, and age level. Expert opinions were obtained once the item had been written and the rubric prepared: Three Social Studies teachers were consulted for *the scientific check*, two measurement and evaluation experts for *the psychometric check*, and one Turkish teacher and one English teacher for *the language and expression check*. These experts were asked whether the item was appropriate for the learning outcome, whether it was novel for students, whether it was appropriate for the grade and age level, about its power to represent the cognitive level, and whether the scoring key was arranged appropriately and accurately for the answers. Based on suggestions from the experts, the item and the rubric were then finalized.

Table 1. *An example of a formative assessment for a Social Studies course.*

<table>
<tr><th>Content Level</th><th>Grade Level</th><th>Cognitive Level</th><th>Scoring Method</th></tr>
<tr><td><b>Course:</b> Social Studies
<b>Learning Area:</b> Individual and Society
<b><i>Learning Gain</i>:</b>
SS.6.1.5. argues that solutions to a problem should be based on rights, responsibilities, and freedoms</td><td>Middle school
6th grade</td><td>Problem Solving (Proposing a Solution to a Problem)</td><td>Rubric</td></tr>
<tr><td colspan="4"><b><i>Item:</i></b>
Classes 5-A and 6-B in a school have a physical education lesson at the same time, and students of both classes want to play basketball in their lessons. However, there is only one basketball hoop in the schoolyard. The teachers of both classes want students to talk to one another and find a solution to this problem within the framework of “rights, responsibilities, and freedoms”.
<b>Offer the students a suggestion to solve this problem.</b> Write your <b>suggestion by associating it with the concepts of</b> “rights, responsibilities, and freedoms”.</td></tr>
<tr><td colspan="4"><b><i>Answer:</i></b></td></tr>
</table>

Concerning the open-ended item given in Table 1, it is expected for the students to offer a realistic solution to the problem and to associate this solution with the concepts of “right, responsibility, and freedom”. The crucial point that the item aims to measure is the instruction part of the item: “Write your **suggestion** ***by associating it with the concepts of*** *‘Right, responsibility, and freedom’*”. If the instruction statement had been given only in the form of “*Offer the students a suggestion that will solve the problem*”, it would have been difficult to question the correctness or incorrectness of the answers and, perhaps, it would have been necessary to accept all answers as correct. It would also not have been possible to know what background thoughts the student had when answering the item. The second part of the item is important in terms of showing whether the student has learned the concepts and whether they can associate these concepts with one another.

In order to improve student achievement, teachers need to be able to both monitor effectively and provide effective feedback. For this, it is inevitable to prepare a detailed rubric. McMillan (2007) emphasizes that teachers should create a rubric for administering the open-ended item and draws attention to whether the item should be scored holistically or analytically. Accordingly, Table 2 presents a holistic rubric prepared for the open-ended item in Table 1.

Table 2. *Holistic rubric for the sample item.*

Answers	Achievement Score
The Most Correct Answer	
The student proposes a solution to this problem within the framework of rights, responsibility, and freedom, and writes their suggestions by associating the solution with these three concepts. *Sample Answer:* I suggest that the two classes play basketball game together to solve the problem because it is the right of both classes to want to play basketball. However, if one class plays basketball the other class will not be able to play. In this case, the freedom of the second class will be denied. No one should hinder the freedom of another. Students need to take responsibility to respect one another's rights and freedoms. *Sample Answer:* To solve the problem, I suggest that one class should play basketball in the first half of the class hour and the other class in the second half. It is the right of the students of both classes to want to play basketball. However, if one class plays basketball the other class will not be able to play. In this case, the freedom of the second class will be denied. No one should hinder the freedom of another. Students need to take responsibility to respect one another's rights and freedoms.	10
Distant Correct Answers	
The student proposes a solution to this problem within the framework of the rights, responsibility, and freedom, and writes their suggestions by associating the solution with these two concepts. *Sample Answer:* To solve the problem, I suggest that one class should play basketball in one week and the other class in the other week because it is right for both classes to want to play basketball. However, if one class plays basketball, the other class will not be able to play. In that case, the freedom of that class will be denied.	8
The student proposes an indirect solution to this problem within the framework of the rights, responsibility, and freedom, and writes their suggestion by associating the solution with these two concepts. *Sample Answer:* I suggest that the two classes sit down, talk, and come to an agreement because it is the right of the students of both classes to want to play basketball. No student or class should hinder the freedom of another. They need to come to terms with one another by talking and taking responsibility.	6
The student proposes a solution with the help of someone else within the framework of the rights, responsibilities, and freedom, and writes the suggestion by associating the solution with the concept. *Sample Answer:* This is the teacher's responsibility. I tell the teacher and ask them to find a solution. The teacher should take responsibility and defend the students' rights.	4

The student proposes a solution based on coincidences without considering the concepts of the right, responsibility, and freedom and writes the suggestion by associating it with a concept. *Sample Answer:* My suggestion is that they flip a coin. Whoever gets the chance play. Let others respect their rights. No one should hinder anyone else's freedom.	2
Blank	0
Incorrect Answers	
The student writes an answer that is correct in itself but not a correct answer to the question. *Sample Answer:* I would say 6/B should play because they are higher grade level. *Sample Answer* I would say 5/A should play because they are lower grade level.	1
Irrelevant Answers	
The student writes a response that is not related to what have been taught. *Sample Answer:* Let them not play basketball, but study instead.	1

In the holistic rubric, a single point is given to the whole of the student's performance, and it is stated that it is necessary to focus on the whole performance by ignoring some minor errors in the performance. At the same time, the answers are considered holistically, and a score is given for each level after the students' answers are ranked from high to low (Kutlu *et al.*, 2017). Since rubrics describe response levels in detail, they allow more consistent scoring (Jonsson & Svingby, 2007). This increases both the validity of students' scores and of the feedback that will be given based on this determination.

In the development of the holistic rubric given in Table 2, *the most correct answer*, then *distant correct answers, blank, incorrect answers,* and then *irrelevant answers* were determined respectively. Sample student answers were given under each answer type. The score values defining the answers were as follows:

- 10 points for the *most correct answer,*
- 8, 6, 4, and 2 points for *distant correct answers,*
- 0 points for *blank answers,*
- 1 point for *incorrect* and *irrelevant answers.*

The most correct answer includes giving the expected answer to the item in full. Answers in this category are exemplary. Distant correct answers include partial accuracy, and they are scored high to the extent they are close to being the 'most correct'. An incorrect answer is logically correct but is not the correct answer to the question asked to the student. An irrelevant answer includes statements that are not related to the learning required to answer the item, or even to anything that has been taught. Nonsense and fabricated answers should be evaluated accordingly.

Scores in the rubric of the sample item are determined from *0* to *10* points. A blank answer is accepted as 0 because it indicates that the student has not answered; that is, absolute absence of an answer. However, incorrect and irrelevant answers are scored as 1 because they show that the student had an idea, even though their answer is not accurate or correct. All response categories provide information about the student's achievement.

The most important indicator for teachers to use when deciding into which response category students' answers fall is the explanation of the sample answer and the sample answer itself. It is inevitable that students will respond to an open-ended item with different explanations. For this reason, the most appropriate answer level into which the student response falls should be determined during scoring. If it cannot be decided into which of the two answer levels the response falls, the higher answer level should be preferred in favor of the student. An examination of the rubric shows that the criteria that distinguish the answer categories from one another (the explanation above the sample answer) move away from the expected answers when moving from the most correct answer to the most irrelevant one. The fact that the answer categories move away from the most correct answer shows both what the student can do and what they cannot, compared with the previous answer.

For example, a student in the answer category corresponding to 6 points in Table 2 has correctly associated two of the concepts of right, responsibility, and freedom. In addition, the suggestion for the solution to the problem was not clear and direct; it was based on an indirect situation. In this case, it would be appropriate for the teacher to show the students in this group the sample answers in the most correct answer and ask them for more concrete examples. There are also situations that students in this group need to complete regarding the concept of rights, responsibility, and freedom. For this purpose, giving students additional reading passages and repeating the subject may contribute to their development.

In another example, a student in the answer category corresponding to 2 points in Table 2 has suggested a solution without considering the concept of rights, responsibility, and freedom, and was not able to associate the concepts with one another. This student's suggestion also did not provide a solution to the related problem. In addition, the student almost never used what they were supposed to have learned the lesson. The teacher should help this student by giving more reading- and writing-based activities and supporting them to improve their learning. First, this student can be asked to read about the concepts of right, responsibility, and freedom, and write examples of the problems they observe in real life.

The students in the incorrect answer category in Table 2 are those who gave answers that can be improved more easily by providing feedback. These students learned some of the information in the lesson; however, they gave a correct answer instead of the expected or measured feature in the related item; their lack of learning may have played a role in giving incorrect answers to the item. For these students, studies similar to those that should be conducted for the answer category corresponding to 2 points can be conducted. Teachers should take more serious measures based on monitoring the students who gave irrelevant answers. These students may have learning difficulties as well as learning deficiencies. It should be kept in mind that students with learning difficulties often experience problems such as comprehending what they have read or listened to and understanding the topics taught in the lesson. Teachers should regularly monitor those students who gave incorrect answers and those who have very low scores, especially those students who gave irrelevant answers.

3. DISCUSSION and CONCLUSION

Considering its role in the teaching and learning process, feedback is not currently being used as effectively as is desired in classroom assessment processes. In particular, the fact that teacher-made tests are based on short-answers, gap-filling, true–false, and multiple-choice item types in schools in Turkey can be seen as one of the important reasons behind this issue (Kutlu & Altıntaş, 2021). These item types are more efficient in measuring basic knowledge at a recall level, which develops over a short period; therefore, they provide stronger feedback on whether this knowledge has been learned. Today, school programs are prepared to be skill-based, and school practices focus on developing students' higher-order thinking skills. For those skills that develop over a long period of time, schools should adopt an approach based on monitoring rather than formatting. Using open-ended items serves the purpose of improving understanding that requires monitoring through the provision of effective feedback in the meaningful parts of the learning process.

Studies emphasize that open-ended items are more effective than other item types in measuring higher-order thinking skills (Badger & Thomas, 1992). Brookhart (2015) states that open-ended items give individuals the opportunity to use their higher-order thinking skills and allow them to express their thoughts more freely. Similarly, Kubiszyn and Borich (2003) state that open-ended items play a more important role in making inferences about the results of complex higher-order cognitive skills, such as problem-solving, analysis, and evaluation. Comparatively, Karakaya and Şata (2022) note that open-ended items develop alternative ways of thinking in students considering that they are prepared at the level of analysis, evaluation, and creation according to Bloom's classification, as well at the levels of problem-solving, critical thinking, and creative thinking according to Haladyna's classification. Therefore, this study used an open-ended item developed based on a real-life situation to provide more effective feedback in the formative assessment process.

In their study *Prospective Teachers' Views About Formative Assessment*, Metin and Özmen (2010) examined the materials developed for the candidates. They gave feedback to each student about the mistakes and deficiencies, and asked them to set the homework again. At the end of the study, the candidates stated that they had noticed their shortcomings and strong aspects, directed their studies, and learned to evaluate themselves thanks to the feedback. In another study, Aydın (2011) scored the answers of 5th grade elementary school students to open-ended items in a mathematics lesson using rubrics, and gave feedback based on these rubrics. The study reported that the application based on giving feedback increased the success of the mathematics course and provided the students with the opportunity to see their strengths and weaknesses.

Similarly, a study by Sabilah and Manoy (2018) used open-ended items with feedback for effective learning of mathematics. In addition, they aimed to describe teachers' learning management, students' activities, and learning achievements. At the end of the study, it was seen that the teachers applied the learning management well, that each student participated in the activity, and that student success was fully achieved. It was also concluded that learning mathematics was more effective when using open-ended items with feedback.

Shute (2008), who examined the studies on formative feedback, states that feedback has been widely discussed in the literature and draws attention to the fact that these studies have many contradictory findings and that, furthermore, there is no consistent learning outcome model for feedback. The author states that this may be caused due to the fact that feedback is mostly used in the teaching process (during classroom activities) and result-oriented assessments (for level determination). This also shows that new suggestions are needed to contribute to the improvement of classroom achievement. Unlike the studies reviewed by Shute (2008), the present study focusses on students' skills that develop over a long period and emphasizes feedback in the formative assessment process. As Torrance and Pryor (1998) report, the important point that should be considered in the learning and teaching process is the integration and continuity of formative assessment within teaching processes.

In their study conducted in the United Kingdom, Harlen and James (1997) draw attention to a different point; they state that the differences between formative and summative assessment approaches in school practices and official documents have disappeared, and that all determinations made by teachers are based on the assumption that they are formative. These researchers argue that there is a need to find a way that will maintain the functional and feature differences between these two assessments and also to link them to one another. Harlen and James (1997) state that this uncertainty will negatively affect the monitoring and feedback processes. Gedye (2010) states that there are several ways in which the quality of feedback can be improved in the formative assessment process. This includes giving feedback as soon as possible, being related to predefined assessment criteria, and giving tips to help students understand how to improve their work.

The present study used the suggestions that increase the effect of feedback in the formative assessment processes. Consequently, it is hoped that the case study model discussed in the present study will contribute to the development of classroom learning success as a result of its use in future studies.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Ozge Altintas https://orcid.org/0000-0001-5779-855X

REFERENCES

Airasian, P.W. (2001). *Classroom assessment: Concepts and applications* (4th Edition). McGraw-Hill.

Aslanoğlu, E.A. (2022). Üst düzey zihinsel beceriler ve ölçülmesi [Higher-order thinking skills and their measurement]. In İ. Karakaya (Ed.), *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi [Preparation, implementation and evaluation of open-ended items]* (pp. 2-25). Pegem Academy Publishing.

Attali, Y., & Powers, D. (2009). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement, 70*(1), 22-35. https://doi.org/10.1177/0013164409332231

Aydın, S. (2011). *İlköğretim 5. sınıf matematik dersinde dereceli puanlama anahtarı kullanılarak verilen geribildirimin öğrenci başarısına etkisi [Student achievement effect of the feedback through the use of scoring rubric at the primary education fifth grade mathematics course].* [Unpublished of master's thesis, Ankara University Institute of Educational Sciences]. Publication No.: 302020. *https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=wjXtJlYgGAsLE-4DXujy5g&no=136vh6_P4eKb8uXAkFDqnQ*

Badger, E., & Thomas, B. (1992). Open-ended questions in reading. *Practical Assessment, Research & Evaluation*, *3*(4), 03-12. https://www.ericdigests.org/1993/open.htm

Bangert-Drowns, R.L., Kulik, C-L.C., Kulik, J.A., & Morgan, M.T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238. https://doi.org/10.3102/00346543061002213

Biggs, J.B., & Watkins, D. (1996). *Classroom learning: Educational psychology for the Asian teacher*. Prentice Hall.

Black, P., & William, D. (1998). Assessment and classroom learning, assessment in education. *Principles, Policy & Practice, 5*(1), 7-75. http://dx.doi.org/10.1080/0969595980050102

Black, P., & William, D. (2002). *Improved standards achieved by transforming assessment for learning*. Kings College London.

Bloom, B.S. (1976). *Human characteristics and school learning*. McGraw-Hill

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning*. Mc Graw-Hill.

Brookhart, S.M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Brookhart, S.M. (2014). *How to design questions and tasks to assess student thinking*. ASCD.

Brookhart, S.M. (2015). Making the most of multiple choice. *Educational Leadership*, *73*(1), 36-39. https://eric.ed.gov/?id=EJ1075062

Clarke, S. (2001). *Unlocking formative assessment: Practical strategies for enhancing pupil's learning in the primary classroom*. Hodder & Stoughton Educational.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481. https://doi.org/10.3102/00346543058004438

Cutting, L.E., & Scarborough, H.S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10(1)*, 277-299. https://doi.org/10.1207/s1532799xssr1003_5

De Cecco, J.P. (1968). *The psychology of learning and instruction: Educational psychology*. Prentice Hull.

Gedye, S. (2010). Formative assessment and feedback: A review. *Planet, 23*(1), 40-45. https://doi.org/10.11120/plan.2010.00230040

Gerberich, J.R., Greene, H.A., & Jorgensen, A.N. (1962). *Measurement and evaluation in the modern school*. David McKay Company, Inc.

Goodrich, H. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-17. http://www.educontinua.fciencias.unam.mx

Haladyna, T.M. (1997). *Writing test items to evaluate higher order thinking*. Viacom Company.

Harlen, W. (2003). *Enhancing inquiry through formative assessment*. Exploratorium.

Harlen, W. (2007). *Assessment of learning*. Sage Publications.

Harlen, W., & James, M. (1997). *Assessment and learning: Differences and relationships between formative and summative assessment*. *Assessment in Education: Principles, Policy & Practice, 4*(3), 365-379. https://doi.org/10.1080/0969594970040304

Jonsson A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Karakaya, İ., & Şata, M. (2022). Açık uçlu maddeler [Open-ended items]. In İ. Karakaya (Ed.), *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi [Preparation, implementation and evaluation of open-ended items]* (pp. 28-37). Pegem Academy Publishing.

Kintsch, W., & Yarbrough, J.C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology, 74*(1), 828-834. https://doi.org/10.1037/0022-0663.74.6.828

Koç, N. (1978). Liselerde öğrencilerin akademik başarılarının değerlendirilmesi uygulamaları nın etkinliğine ilişkin bir araştırma [A research on the effectiveness of the applications of evaluating the academic achievement of students in high schools]. *Education and Science, 3*(14), 28-36. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/5629

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice* (7th Edition). John Wiley & Sons, Inc.

Kuh, G.D., Kinzie, J., Buckley, J.A., Bridges, B.K., & Hayek, J.C. (2006). *What matters to student success: A review of the literature*. Commissioned Report for the National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success. National Postsecondary Education Cooperative. https://nces.ed.gov/npec/pdf/kuh_team_report.pdf

Kulhavy, R.W. (1977). Feedback in written instruction. *Review of Educational Research, 47*(1), 211-232. https://doi.org/10.3102/00346543047002211

Kulhavy, R.W., & Stock, W.A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279-308. https://doi.org/10.1007/BF01320096

Kulhavy, R.W., Yekovich, F.R., & Dyer, J.W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68*(5), 522-528. https://doi.org/10.1037/0022-0663.68.5.522

Kutlu, Ö. (2004). *Tek soruyla öğrenci performansının belirlenmesi [Assessing student performance with a single question]*. Eğitimde İyi Örnekler Konferansı. Sabancı Üniversitesi, İstanbul.

Kutlu, Ö., Yalçın, S., & Pehlivan, E.B. (2010). İlköğretim programında yer alan kazanımlara dayalı soru yazma ve puanlama çalışması [A study on writing and scoring open-ended questions based on the primary school curriculum objectives]. *Elementary Education Online, 9*(3), 1201-1215. https://dergipark.org.tr/tr/pub/ilkonline/issue/8594/106904

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2017). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme [Measurement and evaluation: Performance and portfolio-based assessment]* (5th Edition). Pegem Academy Publishing.

Kutlu Ö., & Kula-Kartal, S. (2018). The prominent student competencies of the 21st-century education and the transformation of classroom assessment. *International Journal of Progressive Education 14*(6), 70-82. https://doi.org/10.29329/ijpe.2018.179.6

Kutlu, Ö., Altıntaş, Ö., Özyeter, N.T., Alpayar, Ç., & Kula-Kartal, S. (2019). *Okuduğunu anlama becerisinin ölçülmesi ve değerlendirilmesi. [Measurement and evaluation of reading comprehension skills].* Ankara University Printing House.

Kutlu, Ö., & Altıntaş, Ö. (2021). Psikolojik ölçmelerin kısa tarihi ve 21. yüzyılda sınıf içi durum belirleme anlayışı [A brief history of psychological measurements and an approach of classroom assessment in the 21st century]. *Trakya Journal of Education, 11*(3), 1599-1620. https://doi.org/10.24315/tred.896121

Marzano, R.J. (1992). *A different kind of classroom: Teaching with dimensions of learning*. The Association for Supervision and Curriculum Development.

McMillan, J.H. (2007). *Classroom assessment: Principles and practice for effective standards-based instruction* (4th Edition). Pearson Education, Inc.

Metin, M., & Özmen, H. (2010). Biçimlendirici değerlendirmeye yönelik öğretmen adaylarının düşünceleri [Prospective teachers' views about formative assessment]. *Milli Eğitim, 187,* 293-310. https://dergipark.org.tr/tr/download/article-file/442743

Miller, D.M., Linn, R.L., & Gronlund N.E. (2008). *Measurement and assessment in teaching* (10th Edition). Prentice-Hall Inc.

Nitko, A.J. (2001). *Educational assessment of students* (3rd Edition). Upper Saddle River.

Özçelik, D.A. (2014). *Eğitim programları ve öğretim: Genel öğretim yöntemi [Curriculum and teaching: General teaching method]* (3rd Edition). Pegem Academy Publishing.

Popham, W.J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd Edition). Allyn and Bacon.

Popham, W.J. (2008). *Classroom assessment.* Allyn and Bacon.

Reiner, C.M., Bothell, T.W., Sudweeks, R.R., & Wood, B. (2002). *Preparing effective essay questions*: *A self-directed workbook for educators*. New Forums Press. https://testing.byu.edu/handbooks/WritingEffectiveEssayQuestions.pdf

Sabilah, I., & Manoy, J.T. (2018). The use of open-ended questions with giving feedback (OEQGF) for effective mathematic learning. *Journal of Physics: Conference Series, 947* 012032. https://doi.org/10.1088/1742-6596/947/1/012032

Shute, V.J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Stiggins, R.J. (1994). *Student-centered classroom assessment*. Macmillan Publishing Compan y.

Stiggins, R.J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765. https://doi.org/10.1177/003172170208301010

Stiggins, R.J. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *The Phi Delta Kappan, 87*(4), 324-328. https://doi.org/10.1177/003172170508700414

Stiggins, R.J., & Conklin, N. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* State University of New York Press.

Şahin, E. (2019). *Sosyal Bilgiler 6 ders kitabı [Social Studies 6th grade textbook]*. Anadol Publishing.

TDK. (n.d.). *Türk Dil Kurumu sözlükleri [Turkish Language Association dictionaries]*. https://sozluk.gov.tr

Thorndike, R.M. (2005). *Measurement and evaluation in psychology and education* (7th Edition). Pearson Education, Inc.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning, and assessment in the classroom*. Open University Press.

Wolman, B. (1973). *Dictionary of behavioral science.* Van Nostrand Company.

York, T.T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research, and Evaluation, 20*(5), 1-20. https://doi.org/10.7275/hz5x-tx03

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 126–145

https://doi.org/10.21449/ijate.1125823

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

The development of an online learning readiness scale for high school students

Mehmet Ramazanoglu[1,*], **Sungur Gurel**[2], **Ali Cetin**[3]

[1]Siirt University, Faculty of Education, Department of Educational Sciences, Siirt, Türkiye
[2]Siirt University, Faculty of Education, Department of Educational Sciences, Siirt, Türkiye
[3]Siirt University, Faculty of Education, Department of Mathematics and Science Education, Siirt, Türkiye

ARTICLE HISTORY

Received: Jun. 6, 2022

Revised: Sep. 21, 2022

Accepted: Oct. 6, 2022

Keywords:

Online learning readiness,
Computer self-efficacy,
Internet self-efficacy,
Self-learning,
High school students.

Abstract: Assessing students' online learning readiness is important since numerous countries have started online learning at all education levels during the Covid-19 pandemic in the 21st century. By taking students' online learning readiness level into account, it will be easier to establish on-target online learning environments. Although there are a number of online learning readiness scales available aiming at higher-education students in the Turkish setting, there is no scale available specifically for high-school students. This study, therefore, aims to develop a valid and reliable scale to identify the levels of online learning readiness for high school students in Türkiye. In order to develop an Online Learning Readiness Scale for high school students, a mixed-method exploratory sequential design was employed in this study. The first sample consisted of 916 students and the second sample consisted of 323 students who had previously experienced an online learning environment. The data were analyzed through exploratory factor analysis and confirmatory factor analysis. Validity and reliability evidences were also provided. The final version of the scale consisted of a total of 16 items in three dimensions; namely, computer self-efficacy, internet self-efficacy, and self-learning and explained 65.76% of the variance. The results of the study indicate that the Online Learning Readiness Scale (OLRS) developed in this particular study is a reliable and valid measurement tool in the assessment of online learning readiness levels of high school students in Türkiye and is expected to guide researchers and practitioners who focus on assessing high school students' online learning readiness levels.

1. INTRODUCTION

Developments in information and communication technologies have affected the fields of education and training. The introduction of the Internet into education and training worldwide has led to the creation of computer-assisted digital environments for learning-teaching activities (Richardson & Swan, 2003). Intelligent tutoring systems, interactive multimedia learning environments, computers as cognitive tools, simulations, microworlds, computer supported

*Correspondıng Author: Mehmet RAMAZANOGLU ✉ m.ramazanoglu@siirt.edu.tr Siirt University, Faculty of Education, Department of Educational Sciences, Siirt, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

collaborative learning, pedagogical agent-based environments, virtual reality environments, and online learning environmenst are the terms describing the use of technology in education as computer assisted digital envirenments (Lajoie & Naismith, 2012). Among these terms, online learning has a special importance because of its ability to supply communication and interaction between learners and teachers in a digital environment (Katz, 2000).

Online education offers a number of beneftis to teachers and students as it allows students to work collaboratively with their teachers and classmates (Katz, 2002), gives opportunuties to students to learn without place and time boundaries (Hill, 2000; Vrasidas & McIsaac, 2000), and also provides convenience and flexibility (Chizmar & Walbert, 1999; Poole, 2000). Through online learning students get quick feedback on their performance (Khan, 1997) and access information from different sources (Lin & Hsieh, 2001). Therefore, online learning is defined as a learning model in which students learn remotely by interacting with their teachers and peers, using the Internet and computer technologies. The effectiveness of online learning processes can be linked to such student characteristics as attitudes towards online learning environment (Sivo et al., 2007), attitudes towards computers (Pillay et al., 2007), perceptions on the usefulness of online learning environments, and flexibility that is obtained by taking courses online (Arbaugh, 2000). Oliver (2001) stated that sustaining online learning relies on teacher expertise, student readiness, technology infrastructure, and reusable learning objects. Given that various student characteristics influence the effectiveness of online learning processes and student readiness is a part of creating online learning, this study focuses on developing a scale that measures online learning readiness of high school students in Türkiye.

1.1. Online Learning Readiness

Online learning readiness is defined differently in various studies due to the differences in the dimensions of the online learning readiness measured and such dimensions measured can be listed as follows:

I. Watkins et al. (2004) developed online learner readiness self-assessment instrument on U.S. Coast Guard personnel and defined online learning readiness as a construct that includes the dimensions of technology access, technology skills, online relationship, motivation, online readings, online video/audio, discussion boards, online groups, and importance to your success.

II. Hung et al. (2010) developed an online learning readiness scale on college students and defined online learning readiness as a construct that includes the dimensions of self-directed learning, motivation for learning, computer/internet self-efficacy, learner control, and online communication self-efficacy.

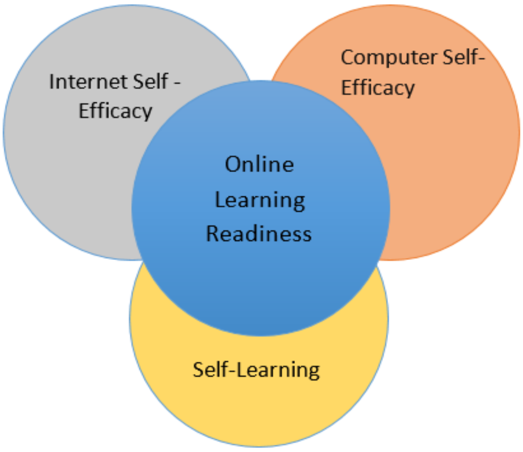
III. Pillay et al. (2007) developed a diagnostic tool for assessing tertiary level students' readiness for online learning on a sample of university students ranging from first year undergraduates to postgraduates and defined online learning readiness as a construct that includes the dimensions of technical skills, computer self-efficacy, learner preferences, and attitudes towards computers.

IV. Smith et al. (2003) adapted McVay readiness for an online learning questionnaire (McVay, 2000) on undergraduate students and defined online learning readiness as a construct that includes the dimensions of comfort with e-learning and self-management of learning.

V. Yurdugül and Demir (2017) developed a readiness for e-learning scale on undergraduate students and defined online learning readiness as a construct that includes the dimensions of autonomous learning and technology usage self-efficacy.

Online learning readiness can be redefined by using these dimensions. Watkins et al. (2004) and Hung et al. (2010) mainly discuss internet usage by using such dimensions as online relationship, online readings, online groups, online video/audio, discussion boards, and online communication self-efficacy. Instead of using various online terms, internet self-efficacy is selected for the first dimension of the study. Watkins et al. (2004), Hung et al. (2010), Pillay et al. (2007), and Yurdugül and Demir (2017) use different terms for the technological readiness of learners; namely, technology skills, computer self-efficacy, technical skills, and technology usage self-efficacy. However, instead of these terms, computer self-efficacy can be used as the other dimension of online learning readiness. All these researchers also point out learners' characteristics. Motivation, importance to your success, self-directed learning, learner control, learner preferences, self-management of e-learning, and autonomous learning are the terms used for self-learning in the relevant literature. In consideration with the common dimensions that are considered as a part of online learning readiness in the related literature, dimensions as to online learning readiness can be listed as internet self-efficacy, computer-self efficacy, and self-learning as can be seen in Figure 1.

Figure 1. *Dimensions of online learning readiness.*



Online learning readiness scales described to date are intended for assessing online learning readiness levels of university students or adults. National literature review does not reveal any scales that assess high school students' readiness for online learning in Türkiye. Assessing high school students' online learning readiness is also important as seen during the Covid-19 pandemic. In Türkiye, although all students are likely to have opportunities to access online learning, some of them cannot access it due to their own lack of readiness.

While learning a new subject or solving a problem related to the subject, students perform activities based on their existing knowledge (Senemoğlu, 2011). While readiness has been accepted as an important factor in classroom learning in the 21st century, its importance has become more understandable with the technological developments experienced to date (Demir-Kaymak & Horzum, 2013). As a matter of fact, in the report published by the International Society for Technology in Education (ISTE) (2016), students are expected to be ready in the ever-evolving technology environments in order to empower students and provide a student-oriented learning process. In this context, determining the readiness levels of students for online learning will also help them learn in online classes. Therefore, students who use digital environments for learning purposes should be ready for online learning in order to enrich their classroom learning.

Having such a tool would help researchers to reveal the readiness levels of high school students in online learning. In addition, demonstrating whether high school students are ready for online learning would help educators to establish more effective online learning environments. Developing an online learning readiness scale for high school students will therefore fill the gap in the literature and allow further research in this context.

1.2. Dimensions of Online Learning Readiness Scale

The intended dimensions of the online learning readiness scale are computer self-efficacy, internet self-efficacy, and self-learning. Self-efficacy is defined as individuals' self-judgments about their capacity to organize and implement the activities necessary to demonstrate their desired performance (Bandura, 1997). The increase in individuals' perceived self-efficacy is associated with increased performance (Bandura et al., 1977). In this context, computer self-efficacy can be defined as "a judgment of one's capability to use a computer" (Compeau & Higgins, 1995, p. 192). Prior research reveals that individuals' high computer self-efficacy levels are important in terms of being successful in online learning environments (Simmering et al., 2009). Chang and Tung (2008) concluded that one of the factors positively affecting the behavioral intention to use online learning course websites is computer self-efficacy. Lim (2001) examined adult learners' satisfaction with a web-based distance education course and their intention to attend a similar course again and concluded that the computer self-efficacy factor was the only statistically significant predictor variable. In their study, Achukwu et al. (2015) investigated 129 first-year undergraduate students' computer self-efficacy and their online learning readiness and reported that computer self-efficacy was significantly correlated with online readiness.

In addition to computer self-efficacy, internet self-efficacy in students is a second concept that needs to be investigated and is defined as the ability of individuals to communicate with their friends in online learning environments, the ability to use the environments on the Internet easily, and the ability to access the information they seek and to separate the information reached (Kim & Glassman, 2013). According to Kuo et al. (2014), it is the ability of individuals to evaluate themselves regarding their ability to organize and conduct activities that need to be done on the Internet.

In a study examining the effect of internet self-efficacy on online learning, the relationship between internet self-efficacy, and students' information-seeking strategies, it was concluded that online learning environments facilitate students' information-seeking strategies (Tsai & Tsai, 2003). In different studies, it has been stated that internet self-efficacy affects students' motivation (Liang & Wu, 2010), their academic achievement, and also their information-seeking behavior (DeTure, 2004) in online learning environments. On the other hand, it was stated that students with low internet self-efficacy levels were worried about participating in online learning environments (Livingstone & Helsper, 2010).

Finally, online learning readiness includes students' self-learning skills and is defined as the ability of students to manage their own work in online environments, to set their goals, and to evaluate themselves (Oladoke, 2006). In online learning environments, students are provided with the opportunity to work independently of time and place, to access information, and to choose and to learn individually (Lin & Hsieh, 2001). Self-learning is when students direct their own learning processes and experiences. In other words, learning can be expressed as a controlled process (Shyu & Brown, 1992). When the importance of self-learning is examined, it is stated that students should have the ability to manage their self-learning habits as well as their motivation due to the independence of online learning (Daniels & Moore, 2000).

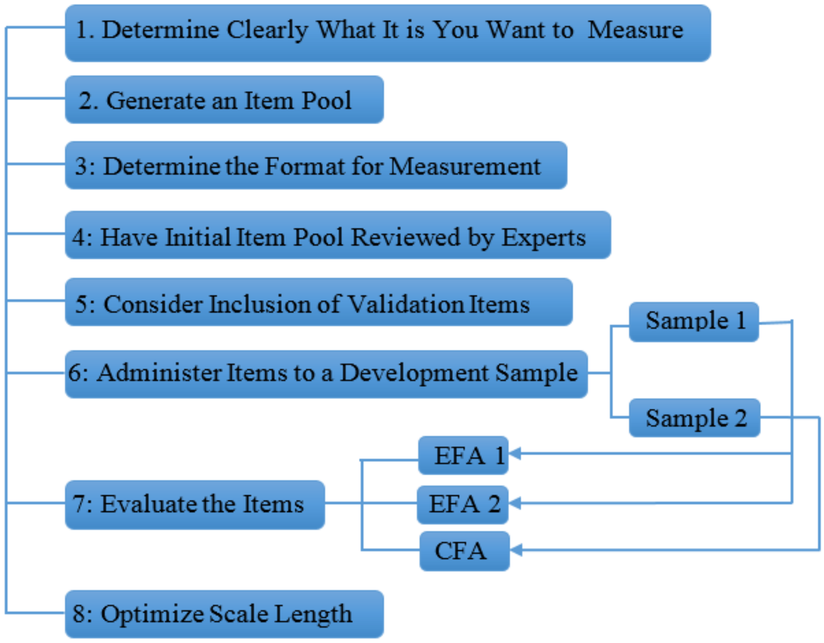
In this study, high school students' readiness for online learning is discussed in the sub-dimensions of computer self-efficacy, internet self-efficacy, and self-learning as the scale was also developed accordingly.

2. METHOD

2.1. Research Design

To develop the Online Learning Readiness Scale (OLRS) for high school students, a mixed-method exploratory sequential design was employed. Qualitative and quantitative data collection and analyses were carried out following a sequence (Creswell, 2012). Its development and validation phases (Creswell & Plano Clark, 2006) were conducted sequentially. In its development phase, the scale development process proposed by DeVellis (2017) was administered to develop OLRS for high school students. In the validation phase, two Exploratory Factor Analyses (EFA) were performed using data obtained from Sample 1 and Confirmatory Factor Analysis (CFA) was performed using data obtained from Sample 2 to test the validity and the reliability characteristics of the scores obtained from OLRS. All of these steps are summarized in Figure 2.

Figure 2. *Online learning readiness scale development steps.*


1. Determine Clearly What It is You Want to Measure
2. Generate an Item Pool
3: Determine the Format for Measurement
4: Have Initial Item Pool Reviewed by Experts
5: Consider Inclusion of Validation Items
6: Administer Items to a Development Sample
Sample 1
Sample 2
EFA 1
7: Evaluate the Items
EFA 2
CFA
8: Optimize Scale Length

2.2. Development of the OLRS for High School Students

2.2.1. *Step 1: Determine clearly what it is you want to measure*

This study aims to develop a scale to measure high school students' online learning readiness. OLRS was constructed in three dimensions: computer self-efficacy, internet self-efficacy, and self-learning. Computer self-efficacy, a judgment of computer usability, is important to be successful in online learning (Hung, 2016). The higher the computer self-efficacy level is, the higher the success on online learning is likely to be. Internet self-efficacy, the ability to use the web services on the Internet easily, allows individuals to organize and conduct the activities which they need to do (Bernard, 2014). Those students who have low internet self-efficacy may worry about participating in online learning activities (Livingstone & Helsper, 2010). Self-learning, the ability of students to manage their own work, represents the opportunity to work independently of time and place (Lin & Hsieh, 2001). Those students who can control their own learning processes are also able to behave and make decisions along with their own needs. With the measurement of these dimensions in OLRS, the online learning readiness of high school students can be determined.

2.2.2. *Step 2: Generate an item pool*

Each item to be used in the scale should be specifically designed for high school students and correspond to one of the readiness sub-dimensions (DeVeilles, 2017). For this reason, these two situations were investigated in the relevant literature while creating the item pool. However, a previously created scale for high school students was not found. An item pool was created based on the items in the relevant sub-dimensions in the studies measuring the online learning readiness of university students (Durak, 2017; Gökçearslan et al., 2017; Horzum et al., 2019; İlhan & Çetin, 2013; Yurdugül & Alsancak Sırakaya, 2013; Yurdugül & Demir, 2017). These items were revised and presented to expert opinions. The related tudies used to make up the item pool and the number of items in the initial item pool are presented in Appendix 1.

The 26 items obtained at the end of the literature review and shown in the initial item pool were written by the researchers in a way that high school students could understand. During the arrangement made by the researchers, the item that contained more than one judgment or situation (Item 1) was separated and a new item was created. Some items designed only for university students were rewritten in a more general form (Item 3, Item 4, Item 5, Item 6, and Item 7). The language of some items was also simplified so that students could understand them more easily (Item 2, Item 8, Item 9). In this way, 10 items were written under the computer self-efficacy dimension.

Some items were combined because their content was close to each other (Item 10 and Item 11). Some items were rewritten with minor adjustments (Item 12, Item 13, Item 14, Item 15, Item 16, and Item 17). A total of 7 items were obtained in the internet self-efficacy sub-dimension. Since the content of some items and the situation to be asked could be easily understood from other items, some items were removed (Item 9, Item 16, Item 22, Item 23, Item 24, Item 26). Some items were rewritten with minor changes (Item 18, Item 19, Item 20, Item 21, Item 25). By reaching a total of 6 items in the self-learning sub-dimension, 23 items were created in the entire scale.

2.2.3. *Step 3: Determine the format for measurement*

Likert-type measurement is a widely used and effective form of measurement in obtaining attitudes, beliefs or opinions (DeVellis, 2017). Thorndike (2005) also pointed out that as the number of options in the scale increases, the reliability of the scores also increases. Responses to OLRS were obtained through a 5-point Likert-type scale. Options were "*Strongly Disagree*", "*Disagree*", "*Neither Disagree nor Agree*", "*Agree*", and "*Strongly Agree*".

2.2.4. *Step 4: Have initial item pool reviewed by experts*

Two different sets of expert opinions were obtained in order to examine the appropriateness of the questions and response options of the OLRS. The first set of opinions recruited as a pilot were obtained from 8 high school students who were registered at different grade levels (one female and one male student from each of the grade levels from 9th through 12th grades). The second set of opinions were obtained from 7 academics experts in the field of measurement and evaluation in education and instructional technologies in education. Appropriate, not appropriate, and explanation statements were written next to each item in the 23-item scale. All of the opinions were obtained during school hours via a face-to-face interview with each person.

At the end of student interviews, we found that the students wanted us to add explanation texts such as writing, creating tables, and making presentations in parentheses to make the Office programs in Items 1, 2 and 3 clearer and to write "distance education system" in parentheses next to the expressions "online learning system". We also asked the students to specify which web sources or internet environments they used in addition to those web sources given in Items 11, 12, and 16.

At the end of the interviews with expert academics, it was stated that the definition of online learning should have been at the beginning of the scale. The yes/no question in item 4 was rearranged and transformed into a Likert form. Item 16, Item 23, and Item 9 were to be removed from the scale because Item 9 had various meanings and did not express a specific situation and Item 16 and Item 23 included expressions close to Item 17. In addition, the place of Item 11 and Item 12 was changed.

2.2.5. *Step 5: Consider the inclusion of validation items (preparation of the data for analysis)*

To keep the ORLS simple and short, no validation item was included. Respondents who answered the OLRS carelessly or without sufficient effort in their response were determined via an investigation of response patterns after data collection phases. Both the longest length of consecutive identical responses and the average length of consecutive identical responses were investigated. Answers that were obtained from individuals who had identical responses throughout the OLRS or who had an average length of more than 5 consecutive identical responses were excluded. Meade and Craig (2012) found that average length of 3.64 to 4.15 consecutive identical responses was found in a real data that were obtained from careless respondents. When individuals met either criterion listed above, it was assumed that these individuals responded to the items by neglecting the content of the items. 1017 students in the first sample and 397 students in the second sample responded to the ORLS initially. 101 of the responses in the first sample and 74 of the responses in the second sample were excluded due to careless or insufficient efforts in responding. Rather than including validation items, the investigation of response patterns and preparation of the data for analysis allowed us to validate the response process to some extent.

2.2.6. *Step 6: Administer items to a development sample*

Evidence based on response processes can be used as validity evidence. Specifically, internal structure of the responses was investigated to obtain valid evidence of the OLRS. Since this method relies on response processes, items were administered to two development samples. Responses from the first sample were used to explore the internal structure of responses via Exploratory Factor Analysis procedures. Responses from the second sample were used to confirm the internal structure of the responses via Confirmatory Factor Analysis. A total of 14 high schools in Siirt Province, Türkiye were selected to participate in the development and validation of OLRS via convenience sampling. All of the students registered in those schools received the OLRS form and responded voluntarily. After excluding careless or insufficient efforts on the part of respondents, the first sample consisted of 916 students and the second sample consisted of 323 responses. The descriptive characteristics of both samples are summarized in Table 1.

Table 1. *Descriptive characteristics of the development samples.*

	1st. Sample (*N*=916)		2nd. Sample (*N*=323)	
	f	*%*	*f*	*%*
Gender				
Male	303	33.08	171	52.94
Female	613	66.92	152	47.06
Grade Level				
9th. Grade	404	44.10	131	40.56
10th. Grade	203	22.16	84	26.01
11th. Grade	215	23.47	44	13.62
12th. Grade	94	10.26	64	19.81

Note: f stands for frequency, % stands for percentage.

There are various suggestions regarding appropriate sample size to estimate parameters in factor analysis. Commonly used rule of thumb suggested by Comrey and Lee (1992) and Tabachnick and Fidell (2013) is 50 as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 as excellent. Nunnally and Bernstein (1967) also suggested having at least 10 cases per question; hence, sample size of 200 would be sufficient for OLRS. Similar to this criterion, Bentler and Chou (1987) suggested that the sample size should be at least 5 times more than the number of estimated parameters. Forero et al. (2009) suggested that factor analysis of ordinal data with sample size of 200 and small factor loadings (<.40) may provide biased estimates, while a sample size of 500+ or models with moderate or high factor loadings (>.40) may provide adequate estimates. Overall, we may conclude that the sample size for the first sample meets very good criteria and sample size for the second sample meets good sample criteria in our specific study.

2.2.7. *Step 7: Evaluate the items*

To understand the validity and reliability properties of the scores obtained using OLRS, a three-step approach was administered. The first step involved Exploratory Factor Analysis (EFA) to understand the internal structure of the data obtained from the first sample. The number of dimensions in the data and items that are highly related to each dimension is determined. To determine the number of dimensions, three rules were used additively. Firstly, Kaiser (1960) and Guttman (1954) criterion was used to determine the maximum number of dimensions. Dimensions that had an eigenvalue of 1.00 or more were taken into consideration. Secondly, Cattell's scree plot rule (1966) was considered. Furthermore, the eigenvalue difference among consequent dimensions was investigated. We assumed that no new dimension emerged when the slope of the scree plot became close to flat. Finally, only those dimensions that met our theoretical expectancies were considered. After the number of dimensions was determined, we investigated the relationship between dimensions and items through the evaluation of factor loadings. As Matsugna (2010) suggested, a standardized factor loading of .40 or more in the absolute value is a common cut-off in social sciences to indicate an important relationship between the dimension and the item.

After this first EFA, we ran a second EFA using data obtained from the first sample again. The main reason to run a second EFA was to see if the structure of the responses was still the same when items that did not work as expected were excluded. The second EFA consisted of dimensions that were emergent at the first EFA and items that were found to be related to the dimension that the item was supposed to be related to. This second EFA allowed us to remove items that were unrelated to the dimension that item was supposed to be loaded theoretically. In addition, reliability evidence of the scores obtained in each dimension was obtained through Cronbach's α statistic (Cronbach, 1951). Cronbach's α statistic of .70 or greater reflects acceptable reliability, .80 or greater reflects good reliability, and .90 or greater reflects excellent reliability.

To confirm the internal structure that was reflected through the second EFA results, we collected additional data but the items that were included in OLRS were determined by the second EFA results only. Confirmatory Factor Analysis (CFA) was conducted using data obtained from the second sample to secure further validation evidence. In addition, we also calculated Raykov'sρ reliability statistic (1997) based on CFA results for further reliability evidence. Raykov'sρ was preferred here because unlike Cronbach's α, each item contributes to the composite score reliability with respect to the magnitude of its factor loading.

Model-data fit in both EFAs and in CFA was evaluated through model X2 Statistic Root Mean Square Error of Approximation (RMSEA) (Byrne, 1998), Bentler Comparative Fit Index (CFI) (Byrne, 1998), Tucker – Lewis Index (TLI) (Tucker & Lewis, 1973), and Standardized Root Mean Square Residual (SRMR) (Kline, 2011). Browne and Cudeck (1993) suggested that RMSEA value that exceeds .10 reflects a serious problem about the model. RMSEA values in between .08 and .10 reflect an acceptable level of model fit and RMSEA values that are smaller than .08 reflect a good model fit (MacCallum et al., 1996). Hu and Bentler (1999) suggested that CFI and TLI values that are smaller than .90 reflect bad fit, values that are between .90 and .95 reflect acceptable fit, and values that are greater than .95 reflect good fit. Again, Hu and Bentler (1999) suggested that SRMR values that are smaller than .08 reflect an acceptable fit.

Both EFAs and CFA were performed in Mplus version 8.6 (Muthen & Muthen, 1998-2017). In consideration of the ordinal nature of the responses, Weighted Least Squares – Mean and Variance Adjusted (WLSMV) estimator was used because Li (2016) concluded that the WLSMV estimator provides unbiased estimates when the sample size is greater than 200 with non-normal data.

Results of item evaluation based on two EFAs and one CFA, as well as reliability, are presented in the Results section in detail.

2.2.8. *Step 8: Optimize scale length*

The scale length was optimized based on two EFAs and one CFA results as presented in the Results section.

3. RESULTS

In order to evaluate items and obtain validity evidence regarding OLRS scores, two EFAs were run to determine the factor structure of the OLRS with the first sample and one CFA was run to confirm the factor structure of the OLRS with the second sample. Reliability evidence regarding OLRS scores was obtained through estimation of Cronbach's α with the first sample and Raykov'sρ with the second sample. Model fit statistics of both EFAs and CFA are summarized in Table 2; followed by results of the EFA 1, EFA 2, and CFA, respectively.

Table 2. *Model fit statistics summary.*

	χ^2	*df*	RMSEA	CFI	TLI	SRMR
EFA 1	1356.114	133	0.100	0.929	0.899	0.047
EFA 2	629.368	75	0.090	0.962	0.936	0.034
CFA	272.195	101	0.072	0.963	0.956	0.052
Acceptable fit			≤ 0.08	≥ 0.90	≥ 0.90	≤ 0.08

Note: df is the degree of freedom.

3.1. EFA 1 with Sample 1

EFA 1 included 20 items and all solutions up to the 5-factor solution were obtained. The eigenvalue of the first factor was 7.481, the eigenvalue of the second factor was 2.881, the eigenvalue of the third factor was 1.753, the eigenvalue of the fourth factor was 1.094, and the eigenvalue of the fifth factor was 0.878. According to the Kaiser-Guttman rule, it can be said that the data set is represented by the most complex 4-factor structure. However, when the scree plot was drawn, it was determined that the curve of the eigenvalues flattened after the third factor; therefore, different factors did not emerge. In addition, for this scale, which is theoretically planned to have three factors only, the 3-factor solution was primarily evaluated. The 3-dimensional structure explains 60.58% of the variance of the answers given to the indicator items. When the model-data fit statistics in Table 2 are examined, the RMSEA value of 0.100 estimated for the 3-factor structure can be seen to be at the limit indicating that the model can be developed seriously according to the RMSEA criterion. However, the CFI value of 0.929 indicates acceptable fit, the TLI value of 0.899 indicates borderline poor fit, and the SRMR statistic of 0.047 indicates a good fit. In general, it can be said that the model-data fit is borderline acceptable.

Table 3. *Results of explanatory factor analysis – factor loadings.*

		First Exploratory Factor Analysis			Second Exploratory Factor Analysis		
No	Item	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
1	I am confident in using Microsoft Office-MS Word program (a writing program).	**0.905**	0.002	-0.044	**0.907**	0.001	-0.039
2	I am confident in using Microsoft Office-MS Excel (a table creation program).	**0.802**	-0.047	0.061	**0.807**	-0.053	0.060
3	I am confident in using Microsoft Office-MS PowerPoint (a presentation program).	**0.714**	0.122	-0.028	**0.727**	0.103	-0.024
4	I know how to log in to my courses using the online learning software (distance education system).	-0.024	**0.602**	0.177			
5	I know how to log in to my courses using the online learning software (distance education system) using a computer.	0.059	**0.637**	0.101			
6	I can progress in my courses by using online learning software (distance education system) on a computer.	0.010	0.284	**0.400**			
7	I feel confident in using the operating system on a computer.	**0.571**	0.249	0.026	**0.578**	0.22	0.024
8	I can use files saved in audio, music, text, etc. formats on a computer	**0.383**	0.493	0.026			
9	I am confident in setting up programs (i.e., installing new software) on a computer.	**0.563**	0.133	0.011	**0.549**	0.146	0.018
10	I can use web browsers (Internet Explorer, Google Chrome, Safari, Mozilla, Opera, etc.) to access the Internet easily.	0.177	**0.727**	-0.032	0.187	**0.731**	-0.009
11	I am confident in using search engines such as Google-Yahoo, Bing, and Yandex on the Internet.	0.181	**0.737**	-0.017	0.198	**0.740**	0.002
12	I am able to find the information I seek on the Internet easily.	-0.016	**0.737**	0.093	-0.012	**0.759**	0.098
13	I can use social networks easily.	0.070	**0.789**	0.000	0.087	**0.796**	0.003
14	I can send e-mails using internet tools.	0.343	**0.461**	-0.021	0.351	**0.476**	-0.017
15	I can use instant messaging software (Skype, WhatsApp, etc.) to communicate with people.	-0.040	**0.734**	0.072	-0.02	**0.724**	0.073
16	I implement my own study plan in online learning (distance learning).	0.003	0.134	**0.698**	0.005	0.129	**0.702**
17	I manage time well in online learning (distance education).	-0.013	-0.042	**0.801**	-0.011	-0.04	**0.795**
18	I set my learning goals in online learning (distance education).	-0.008	0.014	**0.806**	-0.011	0.018	**0.810**
19	I can direct my own learning process in online learning (distance education).	0.014	-0.002	**0.805**	0.013	-0.004	**0.808**
20	I take a high degree of responsibility during online learning (distance learning).	0.081	0.001	**0.601**	0.079	0.002	**0.602**

Note: Factor loadings that are greater than 0.400 in absolute value are bolded. Items 4, 5, 6 and 8 are excluded from the second Explatory Factor Analysis.

The 3-dimensional structure obtained in the exploratory factor analysis was summarized in Table 3 and the items loaded heavily by each dimension were examined, and the items that did not have a factor load of at least 0.400 in the absolute value were determined. Accordingly, although items 4, 5, 6, and 8 were written to determine the level of computer self-efficacy, the factor loads in the first dimension, on which the other theoretically related items were loaded, were found to be -0.024, 0.059, 0.010, and 0.383, respectively. In addition, factor loadings of items 4, 5, and 8 in the second dimension, which was predominantly loaded by the items written to determine the internet self-efficacy level, were found to be 0.602, 0.637, and 0.493, respectively. Finally, factor loading of item 6 in the third dimension, which was predominantly loaded by the items written with the aim of determining the self-learning level, was found to be 0.400. The exploratory factor analysis was renewed by removing items 4, 5, 6, and 8.

3.2. EFA 2 with Sample 1

In consideration with the results of EFA 1, 4 items with low factor loadings with the dimension it aims to measure were removed and EFA was performed again. There were 16 items in EFA 2, and all solutions up to 5-factor solution were obtained again. The eigenvalue of the first factor was 6.103, the eigenvalue of the second factor was 2.745, the eigenvalue of the third factor was 1.673, the eigenvalue of the fourth factor was 0.825, and the eigenvalue of the fifth factor was 0.609. According to the Kaiser-Guttman rule, it can be said that the data set is represented by the most complex 4-factor structure. However, when the scree plot was drawn, it was determined that the curve of the eigenvalues flattened after the third factor; therefore, different factors did not emerge again. In addition, for this scale, which is theoretically planned to have 3 factors, the 3-factor solution was primarily evaluated. The 3-dimensional structure explains 65.76% of the variance of the item responses. When the model-data fit statistics in Table 2 are examined, the RMSEA value of 0.090 estimated for the 3-factor structure indicates that the model is acceptable based on the RMSEA criterion. In addition, the CFI value of 0.962 indicates a very good fit, the TLI value of 0.936 indicates a good fit, and the SRMR value of 0.034 indicates a good fit. In general, it can be said that the model-data fit is good.

The 3-dimensional structure that corresponds to our theoretical expectations and was obtained in the second exploratory factor analysis is summarized in Table 3. Each factor loading was examined in terms of determining which item was loaded heavily by which factor. It was found that all of the items were loaded heavily by the theoretically intended factors with factor loadings of over 0.400. The factor loadings of items 1, 2, 3, 7, and 9 that were written to reveal the level of computer self-efficacy ranged between 0.549 and .907. The factor loadings of the items 10, 11, and 12, 13, 14, and 15 that were written to reveal the level of internet self-efficacy ranged between 0.476 to 0.796, and the factor loadings of items 16, 17, 18, 19, and 20 that were written to reveal the level of self-learning ranged between 0.602 and 0.810. Thus, the first dimension is called computer self-efficacy, the second dimension is called internet self-efficacy, and the third dimension is called self-learning. The Cronbach's alpha statistics calculated to determine the internal consistency of the scores for each dimension score were found to be 0.83, 0.82 and 0.84, respectively.

3.3. CFA with Sample 2

In order to confirm the 3-factor 16-item structure obtained in the EFA 2 results, data were collected again and confirmatory factor analysis was performed using the second data set. When the model-data fit statistics in Table 2 are examined, the RMSEA value of 0.72 was estimated for the 3-factor structure that indicates a good model-data fit. In addition, the CFI value of 0.963 and TLI value of 0.956 indicate a very good fit, and the SRMR value of 0.052 indicates a good fit. In general, the model-data fit can be said to be very good. Thus, the 3-factor and 16-item structure was confirmed in another sample.

Table 4. *Summary of confirmatory factor analysis results.*

No	Item	Factor Loading	Standard Error	*t*	*p*
Computer Self-Efficacy (*Raykov'sp* = 0.860)					
1	I am confident in using Microsoft Office- MS Word program (a writing program).	0.860	0.020	41.996	<.0001
2	I am confident in using Microsoft Office- MS Excel (a table creation program).	0.779	0.023	33.640	<.0001
3	I am confident in using Microsoft Office- MS PowerPoint (a presentation program).	0.782	0.028	27.922	<.0001
7	I feel confident in using the operating system on a computer.	0.739	0.031	23.916	<.0001
9	I am confident in setting up programs (i.e., installing new software) on a computer.	0.530	0.042	12.719	<.0001
Internet Self-Efficacy (*Raykov'sp* = 0.894)					
10	I can use web browsers (Internet Explorer, Google Chrome, safari, Mozilla, Opera etc.) to access the Internet easily.	0.807	0.024	34.230	<.0001
11	I am confident in using search engines such as Google-Yahoo, Bing and Yandex on the Internet.	0.837	0.023	37.116	<.0001
12	I can find the information I seek on the Internet easily.	0.673	0.038	17.546	<.0001
13	I can use social networks easily.	0.829	0.024	34.701	<.0001
14	I can send e-mails using internet tools	0.702	0.034	20.399	<.0001
15	I can use instant messaging software (Skype, WhatsApp, etc.) to communicate with people.	0.733	0.033	22.176	<.0001
Self-Learning (*Raykov'sp* = 0.853)					
16	I implement my own study plan in online learning (distance learning).	0.667	0.034	19.468	<.0001
17	I manage time well in online learning (distance education).	0.725	0.035	20.980	<.0001
18	I set my learning goals in online learning (distance education).	0.779	0.026	29.747	<.0001
19	I can direct my own learning process in online learning (distance education).	0.846	0.024	35.930	<.0001
20	I take a high degree of responsibility during online learning (distance learning).	0.638	0.036	17.648	<.0001
Correlations among dimensions					
Internet Self-Efficacy and Computer Self-Efficacy		0.687	0.033	20.093	<.0001
Computer Self-Efficacy and Self-Learning		0.246	0.056	4.364	<.0001
Internet Self-Efficacy and Self-Learning		0.254	0.057	4.447	<.0001

CFA results are summarized in Table 4. The factor loadings of items 1, 2, 3, 7, and 9 included with the intention of revealing computer self-efficacy ranged between 0.530 and .860. The factor loadings of the items 10, 11, 12, 13, 14, and 15 included with the intention of revealing internet self-efficacy varied between 0.673 and 0.837. Finally, the factor loadings of the items 16, 17, 18, 19, and 20 included with the intention of revealing self-learning ranged from 0.638 to 0.846. In addition, a positive/moderate-high correlation of 0.687 was estimated between computer self-efficacy and internet self-efficacy dimensions. Also, a positive/small correlation of 0.254 was estimated between internet self-efficacy and self-learning dimensions. Finally, a positive/small correlation of 0.246 was estimated between computer self-efficacy and self-learning dimensions. Raykov'sp reliability statistics for the computer self-efficacy dimension was 0.860, for the internet self-efficacy dimension it was 0.894, and for the self-learning dimension it was 0.853. In conclusion, the total scores in all three dimensions were found to be reliable.

4. DISCUSSION and CONCLUSION

To investigate the online learning readiness of high school students, an OLRS for High School Students was developed and validated in this specific study. A review of the national and international literature showed some online learning readiness scales prepared for university students (Hung et al., 2010; Lin et al., 2016; Pillay et al., 2007; Yurdugül & Demir, 2017). It was seen that these studies had common characteristics in the dimensions measured and there was no scale prepared for high school students. Since many countries started online learning in all education levels during the Covid-19 pandemic in the 21st Century, the development of such a scale would fill a gap in the literature to assess the situations of high school students about online learning readiness.

Steps proposed by DeVellis (2017) on scale development were applied in a sequence as can be seen in Figure 2. While preparing the initial item pool in Appendix 1, the items in these scales, which were prepared for university students and mentioned above, were used. Then the researchers rewrote the items to make them easily understandable by high school students. The first version of OLRS was implemented to 8 high school students in a high school in Siirt, Türkiye. Additionally, expert opinion was obtained from 7 academics. After the analysis of students' and experts' opinions, the implementation form of the OLRS was obtained. For statistical analysis of OLRS, 916 high school students in sample 1 and 323 students in sample 2 participated in the study. The data taken from sample 1 were used for EFA and the data taken from sample 2 were used for CFA. The final version of OLRS is displayed in Table 4. In addition, the original Turkish version is also reported in Appendix 2. This final version includes 16 items in three dimensions: computer self-efficacy, internet self-efficacy, and self-learning.

In the studies performed with university students about online learning readiness, the number of dimensions and the content of them differs. In the present study, computer self-efficacy dimension corresponds to technology use (Watkins et al., 2004), ability to use technological tools (computer) (Hung et al., 2010), and technical interaction (Barker, 2002). Internet self-efficacy corresponds to communication (Watkins et al., 2004), willingness to interact, ability to communicate, (Bernard et al., 2004), ability to use technological tools (internet) (Hung et al., 2010), using internet sources (Choucri et al., 2003), and communication skills (Barker, 2002). Finally, self-learning corresponds to self-directed learning (Watkins et al., 2004), self-learning ability and belief, (Bernard et al., 2004), management and responsibility of self-learning (Hung et al., 2010), managing time (Pillay et al., 2007), mental and physical readiness (Borotis & Poulymenakou, 2004), and intrinsic motivation (Smith et al., 2003). Therefore, the present study corresponds to many dimensions mentioned in the related literature, while literacy and access to technology dimensions in Watkins et al. (2004) and using asynchronous and synchronous tools in Pillay et al. (2007) were not included in the study.

5. RECOMMENDATIONS for RESEARCH

As a consequence of the Online Learning Readiness Scale application, there are some recommendations for future research. Correlation between the results of OLRS and newly developed similar scales can be compared to analyze concurrent validity. Furthermore, if OLRS were applied for different levels like elementary and primary students, it would also be necessary to repeat validity and reliability analyses for the data taken from these groups. Such individual differences as gender, education level, age, and familiarity can also be searched in future studies. Additionally, new studies can be performed to adapt OLRS to new cultures through reliability and validity analyses.

Acknowledgments

The authors would like to thank the assistance of Siirt Provincial Directorate of National Education personnel and high school principals who helped researchers to get in touch with high school students. The authors received no financial support for the research.

Declaration of Conflicting Interests and Ethics

The study followed all protocols required to research in an ethical manner. The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. Ethics Committee Number: Siirt University Directorate of Ethics Committee (11/11/2020-86).

Authorship Contribution Statement

Authors contributed equally to this work.

Orcid

Mehmet Ramazanoglu https://orcid.org/0000-0001-6860-0895
Sungur Gurel https://orcid.org/0000-0003-3425-858X
Ali Cetin https://orcid.org/0000-0002-1174-6997

REFERENCES

Achukwu, C.B., Nwosu, K.C., Uzoekwe, H.E., & Juliana, A. (2015). Computer self-efficacy, computer-related technology dependence and on-line learning readiness of undergraduate students. *International Journal of Higher Education Management, 1*(2), 60-71.

Arbaugh, J.B. (2000). Virtual classroom characteristics and student satisfaction with internet-based MBA courses. *Journal of Management Education, 24*(1), 32-54. https://doi.org/10.1177/105256290002400104

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W. H. Freeman/Times Books/ Henry Holt & Co.

Bandura, A., Adams, N.E., & Beyer, J. (1977). Cognitive processes mediating behavioral change. *Journal of Personality and Social Psychology, 35*(3), 125-139. https://doi.org/10.1037//0022-3514.35.3.125

Barker, P. (2002). On being an online tutor. *Innovations in Education and Teaching International, 39*(1), 3-13. https://doi.org/10.1080/13558000110097082

Bentler, P.M., & Chou, C.P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*(1), 78-117. https://doi.org/10.1177/0049124187016001004

Bernard, R.M., Brauer, A., Abrami, P.C., & Surkes, M. (2004). The development of a questionnaire for predicting online learning achievement. *Distance Education, 25*(1), 31-47. https://doi.org/10.1080/0158791042000212440

Borotis, S., & Poulymenakou, A. (2004). E-learning readiness components: Key issues to consider before adopting e-learning interventions. In *E-learn: World conference on e-learning in corporate, government, healthcare, and higher education* (pp. 1622-1629).

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen and J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage.

Byrne, B.M. (1998). *Structural equation modeling with Lisrel, Prelis, and Simplis: Basic concepts, applications, and programming* (1st ed.). Psychology Press. https://doi.org/10.4324/9780203774762

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Chang, S.C., & Tung, F.C. (2008). An empirical investigation of students' behavioral intentions to use the online learning course website. *British Journal of Educational Technology, 39*(1), 71-83. https://doi.org/10.1111/j.1467-8535.2007.00742.x

Chizmar, J.F., & Walbert, M.S. (1999). Web-based learning environments guided by principles of good teaching practice. *The Journal of Economic Education, 30*(3), 248-259. https://doi.org/10.1080/00220489909595985

Choucri, N., Maugis, V., Madnick, S., Siegel, M., Gillet, S., O'Donnel, S., Best, M., Zhu, H., & Haghseta F. (2003). Global e-readiness- for what. In N. Choucri, V. Maugis, S. Madnick, & M. Siegel (Eds.), *Global e-readiness-for what* (pp. 1–47). Center for eBusiness at MIT: Massachusetts Institute of Technology Cambridge, MA.

Compeau, D.R., & Higgins, C.A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly, 19*(2), 189-211. https://doi.org/10.2307/249688

Comrey, A.L., & Lee, H.B. (1992). A first course in factor analysis. Lawrence Eribaum Associates.

Creswell, J. (2012). *Educational research: planning, conducting, and evaluating quantitative and qualitative research.* (4th Ed.). Pearson.

Creswell, J.W., & Plano Clark, V.L. (2006). *Designing and conducting mixed methods research.* SAGE.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

Daniels, H.L., & Moore, D.M. (2000). Interaction of cognitive style and learner control in a hypermedia environment. *International Journal of Instructional Media, 27*(4), 369-382.

Demir Kaymak, Z., & Horzum, M.B. (2013). Relationship between online learning readiness and structure and interaction of online learning students. *Educational Sciences: Theory and Practice, 13*(3), 1792-1797. https://doi.org/10.12738/estp.2013.3.1580

DeTure, M. (2004). Cognitive Style and Self-Efficacy: Predicting student success in online distance education. *American Journal of Distance Education, 18*(1), 21-38. https://doi.org/10.1207/s15389286ajde1801 _3

DeVellis, R.F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.

Durak, H. (2017). Turkish adaptation of the flipped learning readiness scale for middle school students. *Bartın University Journal of Faculty of Education, 6*(3), 1056-1068. https://doi.org/10.14686/buefad.328826

Forero, G.C., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625–641. https://doi.org/10.1080/10705510903203573

Gökçearslan, Ş., Solmaz, E., & Kukul, V. (2017). Mobile learning readiness scale: an adaptation study. *Eğitim Teknolojisi Kuram ve Uygulama, 7*(1), 143-157. https://doi.org/10.17943/etku.288492

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19*(2), 149-161. https://doi.org/10.1007/BF02289162

Hill, J.R. (2000). Web-based instruction: Prospects and challenges. *Educational media and technology yearbook, 25,* 141-55.

Horzum, M., Bektaş, M., Ayvaz Can, A., Üngören, Y., & Sellüm, F. (2019). Authentic learning readiness scale for teachers: The validity and reliability study. *Uluslararası Alan Eğitimi Dergisi, 5*(2), 94-106. https://doi.org/10.32570/ijofe.645859

Hu, L.T., & Bentler, P.M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hung, M.L. (2016). Teacher readiness for online learning: scale development and teacher perceptions. *Computers & Education, 94*, 120-133. https://doi.org/10.1016/j.compedu.2015.11.012

Hung, M.L., Chou, C., Chen, C.H., & Own, Z.Y. (2010). Learner readiness for online learning: Scale development and student perceptions. *Computers & Education, 55*(3), 1080-1090. https://doi.org/10.1016/j.compedu.2010.05.004

International Society for Technology in Education (ISTE). (2016). *ISTE Standards for Students (ebook): A Practical Guide for Learning with Technology.* Susan Brooks-Young.

İlhan, M., & Çetin, B. (2013). *The validity and reliability study of the Turkish version of an online learning readiness scale. Eğitim Teknolojisi Kuram ve Uygulama, 3*(2), 72-101.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, *20*(1), 141-151. https://doi.org/10.1177/001316446002000116

Katz, Y.J. (2000). The comparative suitability of three ICT distance learning methodologies for college level instruction. *Educational Media International, 7*(1), 25-30. https://doi.org/10.1080/095239800361482

Katz, Y.J. (2002). Attitudes affecting college students' preferences for distance learning. *Journal of Computer Assisted Learning, 18*(1), 2-9. https://doi.org/10.1046/j.0266-4909.2001.00202.x

Khan, I.M. (2009). An analysis of the motivational factors in online learning (*Doctoral dissertation, University of Phoenix*). https://www.learntechlib.org/p/127822/

Kim, Y., & Glassman, M. (2013). Beyond search and communication: development and validation of the internet self-efficacy scale (ISS). Computers in Human Behavior, 29 (4), 1421-1429. https://doi.org/10.1016/j.chb.2013.01.018

Kline, B.R. (2011). *Principles and practice of structural equation modeling.* (3rd ed.). Guilford

Kuo, Y.C., Walker, A., Schroder, K.E.E., & Belland, B.R. (2014). Interaction, internet self-efficacy, and self-regulated learning as predictors of student satisfaction in online education courses. *The Internet and Higher Education, 20,* 35-50. https://doi.org/10.1016/j.iheduc.2013.10.001

Lajoie, S.P., & Naismith, L. (2012). Computer-based learning environments. In: Seel, N. M. (eds) *Encyclopedia of the sciences of learning.* Springer. https://doi.org/10.1007/978-1-4419-1428-6_512

Lenahan-Bernard, J.M. (2014). Relationship of computer self-efficacy and self-directed learning readiness to civilian employees' completion of online courses (Doctoral dissertation, Nova Southeastern University). https://www.proquest.com/docview/1727477573?pqorigsite=gscholar&fromopenview=true

Li, C.H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavioral Researcher, 48*, 936-949. https://doi.org/10.3758/s13428-015-0619-7

Liang, J.C., & Wu, S.H. (2010). Nurses' motivations for web-based learning and the role of internet self-efficacy. *Innovations in Education and Teaching International, 47*(1), 25-37. https://doi.org/10.1080/14703290903525820

Lim, C. K. (2001). Computer self-efficacy, academic self-concept, and other predictors of satisfaction and future participation of adult distance learners. *American Journal of Distance Education, 15*(2), 41-51. https://doi.org/10.1080/08923640109527083

Lin, B., & Hsieh, C.T. (2001). Web-based teaching and learner control: A research review. *Computers & Education, 37*(4), 377-386. https://doi.org/10.1016/S0360-1315(01)00060-4

Lin, H.H., Lin, S., Yeh, C.H., & Wang, Y.S. (2016). Measuring mobile learning readiness: Scale development and validation. *Internet Research, 26*(1), 265-287. https://doi.org/10.1108/IntR-10-2014-0241

Livingstone, S., & Helsper, E. (2010). Balancing opportunities and risks in teenagers' use of the Internet: The role of online skills and internet self-efficacy. *New Media & Society, 12*(2), 309-329. https://doi.org/10.1177/1461444809342697

Maccallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how to's. *International Journal of Psychological Research, 3*(1), 97-110. http://www.redalyc.org/articulo.oa?id=299023509007

McVay, M. (2001). *How to be a successful distance learning student: Learning on the internet.* Pearson Custom Pub.

Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychologi cal Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (8th ed.)

Nunnally, J.C., & Bernstein, I.H. (1967). *Psychometric theory* (Vol. 226). McGraw-Hill.

Oladoke, A.O. (2006). Measurement of self-directed learning in online learners. *Ph.D. Thesis, Capella University*. https://www.learntechlib.org/p/118702/

Oliver, R.G. (2001). Assuring the quality of online learning in australian higher education. *Proceedings of 2000 Moving Online Conference.* (pp. 222-231). Gold Coast, QLD. Norsearch Reprographics. https://ro.ecu.edu.au/ecuworks/4792

Pillay, H., Irving, K., & Tones, M. (2007). Validation of the diagnostic tool for assessing tertiary students' readiness for online learning. *Higher Education Research & Development*, *26*(2), 217-234. https://doi.org/10.1080/07294360701310821

Poole, D.M. (2000). Student participation in a discussion-oriented online course: A case study. *Journal of Research on Computing in Education*, *33*(2), 162-177. https://doi.org/10.1080/08886504.2000.10782307

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173-184. https://doi.org/10.1177/01466216970212006

Richardson, C.J., & Swan, K. (2003). Examining social presence in online courses in relation to students' perceived learning and satisfaction. *Journal of Asynchronous Learning Networks, 7*(1), 68-88. http://dx.doi.org/10.24059/olj.v7i1.1864

Senemoğlu, N. (2011). *Gelişim, öğrenme ve öğretim: Kuramdan uygulamaya* (19. ed.). Pegem Akademi.

Shyu, H.Y., & Brown, S.W. (1992). Learner control versus program control in interactive videodisc instruction: What are the effects in procedural learning? *International Journal of Instructional Media, 19*(2), 85-95.

Simmering, M.J., Posey, C., & Piccoli, G. (2009). Computer self-efficacy and motivation to learn in a self-directed online course. *Decision Sciences Journal of Innovative Education, 7*(1), 99-121. https://doi.org/10.1111/j.1540-4609.2008.00207.x

Sivo, S.A., Pan, C.C. & Hahs-Vaughn, D.L. (2007). Combined longitudinal effects of attitude and subjective norms on student outcomes in a web-enhanced course: A structural equation modeling approach. *British Journal of Educational Technology, 3*8(5), 861-875. https://doi.org/10.1111/j.1467-8535.2006.00672.x

Smith, P.J., Murphy, K.L., & Mahoney, S.E. (2003). Towards identifying factors underlying readiness for online learning: An exploratory study. *Distance Education, 24*(1), 57-67. https://doi.org/10.1080/01587910303043

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics.* Boston, MA: Pearson Education Inc.

Thorndike, R. M. (2005). Measurement and evaluation in psychology and education. Upper Pearson Prentice Hall.

Tsai, M.J., & Tsai, C.C. (2003). Information searching strategies in web-based science learning: The role of internet self-efficacy. *Innovations in Education and Teaching International, 40*(1), 43–50. https://doi.org/10.1080/1355800032000038822

Tucker, L.R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. https://doi.org/10.1007/BF02291170

Vrasidas, C., & McIsaac, M.C. (2000). Principles of pedagogy and evaluation for web-based learning. *Education Media International, 37*(2), 105-111. https://doi.org/10.1080/095239800410405

Watkins, R., Leigh, D., & Triner, D. (2004). Assessing readiness for e-learning. *Performance Improvement Quarterly, 17*(4), 66-79. https://doi.org/10.1111/j.1937-8327.2004.tb00321.x

Yurdugül, H., & Demir, Ö. (2017). An investigation of pre-service teachers' readiness for e-learning at undergraduate level teacher training programs: The case of Hacettepe University. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education), 32*(4), 896-915. https://doi.org/10.16986/HUJE.2016022763

Yurdugül, H., & Sırakaya, D.A. (2013). The scale of online learning readiness: a study of validity and reliability. *Eğitim ve Bilim, 38*(169), 391-406. http://eb.ted.org.tr/index.php/EB/article/view/2420

APPENDIX

Appendix 1. The Initial Item Pool That is Reviewed by Experts

	Number of Items: 26	Reference
	Computer Self-efficacy	
1	I am confident in using the basic functions of Microsoft Office programs (MS Word, MS Excel, and MS PowerPoint).	İlhan, M. & Çetin, B. (2013)
2	I trust my knowledge and skill in how to manage online learning software.	İlhan, M. & Çetin, B. (2013)
3	I am confident in using the basic functions of mobile learning systems.	Gökçearslan, Ş., Solmaz, E. & Kukul, V. (2017)
4	I trust my knowledge and skills about mobile learning systems.	Gökçearslan, Ş., Solmaz, E. & Kukul, V. (2017)
5	I am confident in knowing how mobile learning systems work.	Gökçearslan, Ş., Solmaz, E. & Kukul, V. (2017)
6	I can use online note-taking technologies (Color note) to take notes or access my notes.	Durak, H. Y. (2017)
7	I can easily use Windows operating systems.	Yurdugül, H. & Demir, Ö. (2017)
8	I can view the contents of an electronic file (sound, music, text, etc.) on a computer.	Yurdugül, H. & Demir, Ö. (2017)
9	I can solve problems I encounter while using a computer.	Yurdugül, H. & Demir, Ö. (2017)
Internet Self-efficacy		
10	I can download files from the Internet.	Durak, H. Y. (2017)
11	I feel confident when using the Internet (Google, Yahoo) to obtain or collect information for mobile learning.	Gökçearslan, Ş., Solmaz, E. & Kukul, V. (2017)
12	I can easily use web browsers (Internet Explorer, Google Chrome etc.).	Yurdugül, H. & Demir, Ö. (2017)
13	I can easily find the information I am looking for on the Internet.	Yurdugül, H. & Demir, Ö. (2017)
14	I can easily ask questions in online discussion environments.	Yurdugül, H. & Demir, Ö. (2017)
15	I can ask for help by using internet tools (discussion sites, social networks, e-mail, etc.)	Yurdugül, H. & Demir, Ö. (2017)
16	I can easily communicate with voice or video on the Internet (Skype, Google hangout, Google talk, etc.).	Yurdugül, H. & Demir, Ö. (2017)
17	I can use instant messaging software (Skype, WhatsApp etc.) to communicate with people.	Durak, H. Y. (2017)
Self-learning		
18	I implement my own study plan.	Yurdugül, H. & Alsancak Sırakaya, D. (2013)
19	I manage time well.	Yurdugül, H. & Alsancak Sırakaya, D. (2013)
20	I set my learning goals.	Yurdugül, H. & Alsancak Sırakaya, D. (2013)
21	I can direct my own learning process.	İlhan, M. & Çetin, B. (2013)
22	I can direct my own learning process online.	Yurdugül, H. & Alsancak Sırakaya, D. (2013)
23	I set goals in my work and take a high degree of responsibility.	Gökçearslan, Ş., Solmaz, E. & Kukul, V. (2017)
24	In a subject that requires special field expertise, I support the student in determining the right field experts to support his/her own learning.	Horzum, M., Bektaş, M., Ayvaz Can, A., Güngören, Y. & Sellüm, F. (2019)
25	During the online education process, other online activities (chatting, surfing the Internet) do not distract me.	İlhan, M. & Çetin, B. (2013)
26	I prepare my own work plan and put it into practice.	Durak, H. Y. (2017)

Appendix 2. OLRS Final Turkish Version

LİSE ÖĞRENCİLERİ İÇİN ÇEVRİMİÇİ ÖĞRENME HAZIRBULUNUŞLUK ÖLÇEĞİ		Kesinlikle Katılmıyorum	Katılmıyorum	Ne Katılıyorum Ne Katılıyorum	Katılıyorum	Kesinlikle Katılıyorum
Bilgisayar Öz Yeterliliği						
1.	Microsoft Office- MS Word programını (yazı yazma programı) kullanma konusunda kendime güvenirim.					
2.	Microsoft Office- MS Excel (tablo oluşturma programı) programını kullanma konusunda kendime güvenirim.					
3.	Microsoft Office- MS PowerPoint (sunum yapma programı) programını kullanma konusunda kendime güvenirim.					
7.	Bilgisayardaki işletim sistemini kullanma konusunda kendime güvenirim.					
9.	Bilgisayara program (yeni yazılım kurma) kurma konusunda kendime güveniyorum.					
İnternet Öz Yeterliliği						
10.	İnternete kolay erişim için web tarayıcılarını (İnternet Explorer, Google Chrome, safari, mozilla, opera v.b.) rahatlıkla kullanabilirim.					
11.	İnternette Google-Yahoo, bing ve yandex gibi arama motorlarını kullanabilme konusunda kendime güvenirim.					
12.	İnternette aradığım bilgiye rahatlıkla ulaşabilirim					
13.	Sosyal ağları rahatlıkla kullanabilirim.					
14.	İnternet araçlarını kullanarak mail gönderebilirim					
15.	İnsanlarla iletişim kurmak için anlık mesajlaşma yazılımlarını (Skype, WhatsApp vb.) kullanabilirim.					
Kendi kendine öğrenme						
16.	Çevrimiçi öğrenmede (uzaktan eğitim sürecinde) kendi çalışma planımı uygularım.					
17.	Çevrimiçi öğrenmede (uzaktan eğitim sürecinde) zamanı iyi yönetirim.					
18.	Çevrimiçi öğrenmede öğrenme (uzaktan eğitim sürecinde) hedeflerimi belirlerim.					
19.	Çevrimiçi öğrenmede (uzaktan eğitim sürecinde) kendi öğrenme sürecime yön verebilirim.					
20.	Çevrimiçi öğrenme (uzaktan eğitim sürecinde) sırasında yüksek derecede sorumluluk alırım.					

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 146–161

https://doi.org/10.21449/ijate.1126788

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Examining the effectiveness of discussion-oriented flipped learning environments

Erdi Okan Yilmaz[1,*], Nurettin Simsek[2]

[1]Uşak University, Distance Education Application and Research Centre, Türkiye
[2]Ankara University, Faculty of Educational Sciences, Computer Education and Instructional Technologies, Türkiye

ARTICLE HISTORY

Received: June 6, 2022

Revised: Sep. 09, 2022

Accepted: Sep. 12, 2022

Keywords:

Flipped learning,
Achievement,
Satisfaction,
High-ordered thinking skill,
Discussion-oriented flipped learning environment.

Abstract: The overall aim of the study was to examine the effects of the discussion-oriented flipped learning environments on the achievements, satisfaction levels, and high-ordered thinking skills of students. This semi-experimentally planned research was prepared in accordance with the 3x2 factorial design and conducted with a group of 190 second-year coeducational students attending their undergraduate education at Uşak University. A six-week application was conducted with three groups of students, who were classified as participating in discussions in the newly-developed discussion-oriented flipped learning environments with mandatory, voluntary, and non-attendee participation status. As the data collection tool of the research, achievement tests consisting of multiple choice and open-ended questions were used together with the satisfaction scales (related to videos, discussions, and general environment) developed by the researcher. As a result of the posttests applied after the application, it was determined that the overall achievement scores of the students, who participated in the discussions in discussion-oriented flipped learning environments, were significantly higher than those who did not participate in the discussions. It was determined that there was statistically no significant difference between the satisfaction levels of students concerning the videos, while the discussion satisfaction levels of students who participated on a mandatory basis were statistically significantly higher compared to those who participated on a voluntary basis. In terms of high-ordered thinking skill scores, it was determined that mandatory or voluntary participation in discussions in flipped learning environments have a significant and positive impact on high-ordered thinking skills, in comparison to the non-participation.

1. INTRODUCTION

In parallel with the ongoing development of technology, different technological methods and techniques are developing in the education field in an attempt to include them into the teaching and learning processes. In particular, the development of communication technologies as well as devices with internet connection have paved the way for efforts to benefit from these technologies in the education field. In this continuous development and change, the meanings and expectations attributed to teaching and learning processes are changing and becoming

*Corresponding Author: Erdi Okan Yilmaz ✉ erdi.yilmaz@usak.edu.tr Uşak University, Distance Education Application and Research Centre, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

diverse. As part of this change and development, the needs and expectations of students differ, and different learning models and methods are emerging in response to these expectations (Yeşilyaprak & Partners, 2015). One of these different and new methods is the flipped class concept, which was first used by M. Lage, G. Platt and M. Treglia in the 2000s (Ng, 2015). The first studies and the first ideas about this concept were also emphasized by J. Wesley Baker (2000), who was a K12 teacher at the time (Bates et al., 2017).

In the flipped learning, the learning process in the classroom was replaced with the non-classroom processes. In this context, the classroom teaching was transferred to non-classroom environments, and out-of-school activities were taken into the classroom (Baker, 2000; Ng, 2015). Simply put, flipped learning is a learning process in which students watch the videos prepared as a course material at home and implement the practices and exercises given as homework face-to-face in the classroom environment (Bergmann & Sams, 2014).

When the literature is examined, it is seen that there are both positive and negative views about the flipped learning method. Advantageous aspects of flipped learning can be listed as follows: it supports student-centered teaching (Blau & Shamir-Inbal, 2017; Milman, 2012). Students can watch videos whenever and wherever they want (Davies et al., 2013; Enfield, 2013; Marwedel & Engel, 2014, Ramaglia, 2015). It supports students to be able to do teamwork (Blau & Shamir-Inbal, 2017; Marwedel & Engel, 2014). Students can progress at their own pace (Davies et al., 2013; Enfield, 2013; Lee & Park, 2018; Milman, 2012; Ng, 2015; Ramaglia, 2015). It saves time (Bergmann & Sams, 2012; Milman, 2012). Increases student–teacher and student–student interaction (Bergmann & Sams, 2012; Blau & Shamir-Inbal, 2017; Hung, 2018; Lee & Park, 2018). Problems experienced by students concerning non-classroom learning can be eliminated with accompaniment of teacher through classroom activities (Torun & Dargut, 2015). It is scalable, whereby it can be applied to more crowded classrooms (Davies et al., 2013). Offers students the opportunity for collaborative learning (Brewer & Movahedazarhouligh, 2018; Lee & Park, 2018; Strayer, 2012). Develops critical thinking and problem-solving skills of students (Lee & Park, 2018). It allows students to get prepared before classroom learning activities (Lo & Hew, 2017). It allows students to practice in the classroom (Topalak, 2016). It allows teachers to receive more feedback about students (Ramaglia, 2015).

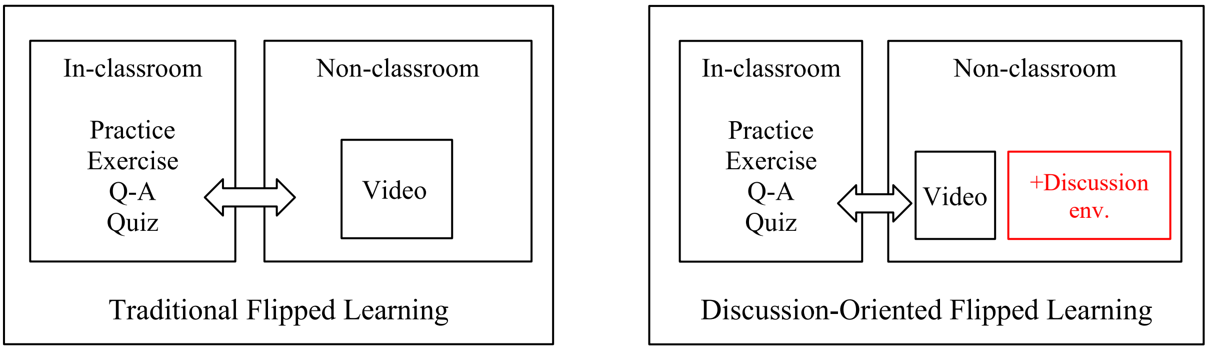
Besides the advantageous aspects of flipped learning in the literature, it was also reported that there are some limitations and disadvantages in the application and functioning of the method. The researchers reported the disadvantages of flipped learning in their findings resulting from their descriptive and experimental studies. The disadvantages of the flipped learning method can be listed as follows: Failure to be sure whether videos are watched or not (Acedo, 2019; Milman, 2012; Turan & Göktaş, 2015). The obligation for students to collaborate among themselves (Acedo, 2019). Students have difficulty in interacting with the teacher and other student friends (Aydın & Demirer, 2016; Bhagat et al., 2016; Gündüz & Akkoyunlu, 2016; Milman, 2012; Nouri, 2016; O'Flaherty & Phillips, 2015). Students feel lonely and isolated in front of the video material (Du et al., 2014; Jerkins, 2017; Milman, 2012; Nouri, 2016; Talbert, 2012). Students have no chance to ask questions to their friends or teachers (Bhagat et al., 2016; Milman, 2012; Turan & Göktaş, 2015). Students cannot receive feedback outside the classroom (Gündüz & Akkoyunlu, 2016; Turan & Göktaş, 2015). The possibility of the student to come to class without watching a video lesson (Gündüz & Akkoyunlu, 2016; Milman, 2012). Students have difficulty in establishing a relationship of meaning between subjects (Chowdhury, 2017). The method requires fast internet connection and hardware (Acedo, 2019; Du et al., 2014; Jerkins, 2017; Krueger, 2012; Ramaglia, 2015; Turan & Göktaş, 2015). It is impossible to determine to what extent the students learn outside the classroom (Du et al., 2014; Gündüz & Akkoyunlu, 2016; Krueger, 2012, Talbert, 2012). There is a need for students to be motivated and their satisfaction level can decrease (Du et al., 2014; Gündüz & Akkoyunlu, 2016; Krueger, 2012; Talbert, 2012; Yılmaz, 2017). Making videos may be difficult for teachers (Acedo, 2019; Du et al., 2014; Gündüz & Akkoyunlu, 2016; Milman, 2012; Ramaglia, 2015; Talbert, 2012).

It is foreseen that staying alone with the video material after the school, feeling themselves alone and isolated, and being unable to communicate and cooperate with other fellow students in the learning process will have a negative impact on the learning process and decrease the motivation and performance levels of the students who are attending their education in the flipped learning environments. Therefore, it was envisaged that more effective and efficient teaching–learning processes can be achieved by eliminating these disadvantages and limitations (Acedo, 2019; Aydın & Demirer, 2016; Bhagat et al., 2016; Bolat, 2016; Davis et al., 2013; Du et al., 2014; Gündüz & Akkoyunlu, 2016; Jenkins, 2017; Krueger, 2012; Milman, 2012; Ramaglia, 2015; Turan & Göktaş, 2015).

In online learning, various means of interaction and communication can be used to address such negative situations. It is reported in the literature that increasing the number of interaction tools and learner interactions in online learning environments in various dimensions can also increase success (Üstündağ, 2012). For example, if a discussion environment is applied in the flipped learning process, learners can interact with each other and with their instructors, and these interactions can also have a positive impact on the achievement and satisfaction levels of the learners (Zainuddin, 2018). Burch (2013; Quoted in Tetreault, 2013) stated that when students are alone in front of a video material available to them for teaching purposes in non-classroom learning environments, their certain needs such as asking questions, interacting, searching for different learning resources can be addressed in a discussion environment that will take place in a flipped learning environment. Chowdhury (2017) stated that students in flipped learning environments may feel isolated, which in turn may result in misunderstanding the content and inability to connect important concepts. In order to avoid this kind of limitations, it was proposed to use the online discussion media in flipped learning environments.

In this context, it is envisaged that some of the disadvantages mentioned in the literature, such as being unable to interact, feeling isolated, not being motivated, not being able to ask questions, not being able to cooperate, not being able to connect the subjects, and experiencing a decline in performance can be eliminated by an "asynchronous online discussion environment" integrated with the flipped learning environment (Figure 1).

Figure 1. *The discussion-oriented flipped learning environment.*



In this context, it emerged as a necessity to use online discussion environments to eliminate some of the disadvantages of flipped learning environments and to examine the impact of this implementation on the learning-teaching processes. From this point on, the overall aim of the study was determined as examining the impacts of undergraduate students' participation in non-classroom online discussion activities in flipped learning environments on their academic achievement, satisfaction, and high-ordered thinking skills. In line with this overall aim, answers are sought for the following questions: Is there a difference among the overall achievement scores of students based on their participation in discussions in a discussion-oriented flipped learning environment? Is there a difference among the satisfaction levels of students based on their participation in discussions in a discussion-oriented flipped learning environment? Is there a difference among the high-ordered thinking skill scores of students based on participation in discussions in a discussion-oriented flipped learning environment?

2. METHOD

This semi-experimental research was conducted in accordance with the 3x2 factorial design, taking into account the number of study groups and repeated measures. Accordingly, the first of the factors of the factorial pattern, which includes repeated measures, is the state of participation in discussion environments (mandatory, voluntarily, and non-attendance), which is the independent variable. The second factor is the two-level measurement variable consisting of "pretest and posttest", which is employed to measure the change in achievement according to tests. The dependent variables of the research are achievement, satisfaction, and high-ordered thinking skills. The symbolized version of the research model is shown in Table 1.

Table 1. *Research model.*

Study Groups	Pretest	Implementation	Posttest
GR1 (Mandatory)	$M_{1\text{-}1}$	Mandatory participation in discussions	$M_{1\text{-}2}$
GR2 (Voluntary)	$M_{2\text{-}1}$	Voluntary participation in discussions	$M_{2\text{-}2}$
GR3 (Non-attendee)	$M_{3\text{-}1}$	Not participating in discussions	$M_{3\text{-}2}$

$M_{1\text{-}1,2\text{-}1,3\text{-}1:}$ Pretest implemented to the groups: Achievement, high-ordered thinking.

$M_{1\text{-}2,2\text{-}2,3\text{-}2:}$ Posttest implemented to the groups: Achievement, high-ordered thinking, satisfaction.

The study group was comprised of 190 students who were attending Uşak University in the fall semester of 2018 academic year and who were receiving Computer Programming courses from the Faculty of Education, Computer Education and Instructional Technologies Department; Faculty of Science, Department of Mathematics; Faculty of Economics and Administrative Sciences, Department of Econometrics. Each class is divided into three groups of students who are participating in discussion activities in a flipped learning environment mandatorily (N: 69), voluntarily (N: 61), and non-attendee (N: 60).

The students in the mandatory group are the ones who are required to participate in discussion activities in a discussion-oriented flipped learning environment. Students in this group were required to submit a discussion topic / discussion question and participate in discussions opened by their friends. The students in the voluntary group are the students whose participation in discussion activities in a discussion-oriented flipped learning environment is optional. The participation of the students in this group in the discussion activities is subject to their own wishes. The students in the non-attending group did not participate in any discussion activities. There was no discussion in the flipped learning environment in which these students were present.

2.1. Data Collection Tools

In scope of the study, in order to measure the achievement, which is one of the dependent variables, achievement pretest and posttest consisting of multiple-choice questions were applied as well as high-ordered thinking skills pretest and posttest consisting of open-ended questions. Two separate achievement tests were developed to measure the students' achievements in the Go Programming course before and after the experimental procedure. The dependent variable of achievement was evaluated with the scores obtained from two basic measurements as pretest and posttest. Both achievement tests consist of questions from the same subject that meet the same gains. While the achievement pre-test was administered before the six-week application period, the achievement post-test was administered after the six-week application period. To reliability analysis of achievement tests, a draft pretest and posttest of 40 questions were applied to 28 students from Uşak University, Department of Mathematics, who had previously taken Go Programming course. Sufficient time was given to the students in their test solutions. In line with the data obtained from the answers given by the students to the test, item analysis was performed on the draft pretest and posttest achievement tests. In line with the data obtained, item difficulty and item discrimination indices were calculated. The difficulty index of the achievement pretest, which was consisting of 12 multiple choice questions developed by the researcher and the instructor, was calculated as 0.50 (medium difficulty) and the distinctiveness average as 0.56 (very good). The KR-20 reliability coefficient, one of the indicators of internal consistency of the test, was calculated as 0.70 (reliable) for the achievement pretest. Additionally, the difficulty index of the achievement posttest consisting of 12 multiple choice questions was 0.52 (medium difficulty) and the distinctiveness average was 0.57 (very good). The KR-20 reliability coefficient, which is one of the indicators of internal consistency of the test, was calculated as 0.73 (reliable) for the achievement posttest.

In order to measure the level of satisfaction, which is another dependent variable of the research, the satisfaction scales were used, which were developed by the researcher consisting of three sub-scales. During the development of the scales, the draft scales were first examined in terms of content and construct validity. Within the scope of the content validity study of the draft scales, opinions were received from 9 field experts, one of whom was a Turkish Language expert. The experts examined whether the scale items were appropriate for the purpose and whether they were understandable in terms of language. Some items have been corrected. Within the scope of the construct validity study of the draft scales, 161 students who were studying in the second year of the Faculty of Communication at Uşak University were studied. The students tested the developed environments and then answered the scales. Video satisfaction is a sub-scale developed to measure the satisfaction levels of students towards the course videos. This scale was applied to all three groups of students. As a result of the reliability analysis of the 15-item video satisfaction sub-scale, the Cronbach Alpha reliability coefficient was calculated as $\alpha=0.95$. Discussion satisfaction is a sub-scale developed to measure the satisfaction levels of students in the discussion environment embedded in the flipped learning environment and learning processes therein. This sub-scale was applied only to two groups of students who participated in the discussion environment on a mandatory and voluntary basis. As a result of the reliability analysis of the 10-item discussion satisfaction sub-scale, the Cronbach Alpha reliability coefficient was calculated as $\alpha=0.96$. General environment satisfaction is a sub-scale developed to measure the satisfaction levels of students about the flipped learning system developed by the researcher. This sub-scale was applied to all the students in three study groups. As a result of the reliability analysis of the 10-item general environment satisfaction sub-scale, the Cronbach Alpha reliability coefficient was calculated as $\alpha=0.94$.

High-ordered thinking skills pretest and posttest, each consisting of 5 open-ended questions, were used in order to reveal the overall achievement scores of students experiencing the newly developed environment, and to investigate its reflection on the higher-ordered thinking skills of them. Demirtaşlı (2010) stated that written exams consisting of open-ended questions, projects or self-assessments can be used to measure students' high ordered thinking skills. Similarly, Wright (2010) stated that open-ended questions can be used to measure higher-order thinking skills. Open-ended questions are those that allow the student to answer freely, and the correct answer can be expressed in different ways. The test, which consists of open-ended questions, is a parallel measurement tool with a similar scope to the achievement tests consisting of multiple-choice questions prepared for the Computer Programming course. In order to test the high-ordered thinking skills of students, two measurement tools which were consisting of a total of 10 open-ended questions prepared by two field experts were developed following the content validity analysis. In the process of developing open-ended questions, a content validity study was conducted with five field experts. In line with the feedback from the experts, a revision study was carried out on the open-ended questions. Answers to open-ended questions consist of texts in which students convey their free thoughts and experiences and may reflect all or part of the ideal one-to-one answer (Karadeniz, 2016). Therefore, different types of methods such as classification according to other question types (good-moderate-poor) or grading (0-5) can be used while scoring. Within the scope of this research, a rubric was used.

2.2. Newly Developed Environments

The newly developed learning environment was designed as three different environments under two types: *with-discussion* and *without-discussion* environments. While there was a discussion environment in the settings of the students who participated in the discussions either mandatorily or voluntarily, there was no discussion environment in the flipped learning environment in which the students of the non-attendee group participated. The environment was developed for teaching the Go programming language within the scope of Computer Programming course. The six-week course videos were shot and prepared in a professional studio environment by the researcher together with the course instructor, and then they were placed in the three newly-developed environments. In addition, questions are embedded in the videos in order to ensure that the videos are viewed. The newly-developed discussion-oriented flipped learning environments were examined by nine field experts before the application, and they were asked to make an assessment. In accordance with feedback from the experts, a student group consisting of 42 students apart from the study group was asked to experience the environment, participate in the preliminary applications, and then make an evaluation. After taking into account the feedback from the students, the environment was put into its final form with necessary revisions.

2.3. Application Process

Discussion environments are prepared in asynchronous structure. There were no moderators in the discussion environments. Discussions were conducted within the framework of the Go Programming Language in which this application is run. The students were able to open any discussion topic they wanted and answered the discussion topics of their friends.

Before the six-week application process began, orientation meetings were held with all students. Detailed information was provided in the orientation meetings in certain subjects such as access to the system, use of the system, information about videos, and a number of activities that students can do within the system (watching video, answering video questions, participating in discussions, scoring, etc.). A different meeting was arranged with the students in the mandatory participation group in a different time, and they were guided about that participation in the discussion in the system is mandatory, they should participate in the discussions throughout the process and initiate discussion topics, and it is also mandatory to ask questions and write answers for the subjects initiated by other friends.

Students whose participation in discussions was mandatory within the framework of non-classroom application activities watched the course videos and answered questions while they were watching. Students in the mandatory group watched the videos, mandatorily participating in discussions and responding the subjects their friends addressed. Students of the non-attendee discussion group watched course videos and answered the questions embedded in videos without any discussion environment. The lecturer did not participate in the discussions, preventing the existence of any authority or moderator in the discussion environments.

Within the framework of classroom application activities, students carried out face-to-face weekly applications with the instructor in line with the course follow-up process. In the flipped learning environment based on the video course content concerning the subjects specific to the Go Programming Language, students carried out activities by writing codes in the laboratory environment. Sample code writing exercises have been performed continuously in the classroom environment. The class learning process was carried out in the same way in all groups.

2.4. Data Collection

Before the application, the achievement pretest including multiple-choice questions was administered to the students, and similarly, the high-ordered thinking skill pretest including open-ended questions was implemented in order to measure the achievement levels. After the six-week application process, students were administered the achievement posttest consisting of multiple-choice questions, the high-ordered thinking skills posttest consisting of open-ended questions, and satisfaction sub-scales (concerning the videos, discussions, and general environment). Satisfaction sub-scales concerning the videos and general environment were administered to the whole study groups, while the discussions satisfaction sub-scale was applied to the students participating in the discussions in the voluntary and mandatory groups, but not to the students from the non-attendee group that did not participate in the discussions. All activity records of students during the six-week discussion-oriented flipped learning environment were obtained from their logs on the system.

2.4. Data Analysis

The overall success score was calculated by adding 50% of the achievement test scores consisting of multiple-choice questions and 50% of the achievement test scores consisting of open-ended questions. One-way variance analysis (ANOVA) was used in the analysis of the overall achievement pretest and posttest scores. When the pre-application overall achievement pretest scores were analyzed, it was determined that there was no difference among the groups, and since the groups demonstrated a homogenous distribution, the analyses were made over the posttest scores. Therefore, instead of analysis of covariance, one-way variance analysis (ANOVA) was employed for the three groups through posttest scores. The possible differences among the satisfaction and overall achievement scores of the three participant groups in the study were interpreted as a result of their participation in the discussions.

In the analysis of the data obtained from satisfaction sub-scales (video, discussion, general environment), it was examined whether they were suitable for parametric analysis, and it was decided to employ one-way variance analysis (ANOVA). Independent samples t-test was used in the analysis since the data obtained from the satisfaction scale concerning discussions were applied only to the two groups of students participating in discussions mandatorily (GR1) and voluntarily (GR2).

10 open-ended questions (five pretests and five posttests) prepared to measure high-ordered thinking skills were rated by four different experts. The high-ordered thinking skill score was obtained by taking the average of the scores given by these four experts. The reliability between the scorers was calculated through the intraclass correlation coefficients. The analysis about whether the scores of high-ordered thinking skills pretest and posttests, which were consisting of open-ended questions, differ among the groups was tested through one-way variance analysis (ANOVA).

3. FINDINGS

3.1. Findings and Interpretations Concerning the Achievement Variable

The findings of the students concerning the achievement variable were obtained from the pretest implemented before the application and the posttest after the application.

Table 2. *Mean and standard deviation values of the students in the groups concerning the pretest-posttest overall achievement scores.*

Groups	*N*	Pretest		Posttest	
		$\bar{X}$	*Sd*	$\bar{X}$	*Sd*
GR1 – Mandatory	69	14.31	9.66	42.41	14.69
GR2 – Voluntary	61	14.25	8.12	42.58	13.23
GR3 – Non-Attendee	60	13.96	8.85	33.43	14.59
Total	190				

Examining Table 2, according to the pretest and posttest overall achievement scores, the average achievement scores of students who participated in discussions in flipped learning environments was $\bar{X}$=14.31 before the application, whereas it was $\bar{X}$=42.41 after the application. The mean achievement score of students participating in the discussions was $\bar{X}$=14.25 before the application, while it was $\bar{X}$=42.58 after the application. The mean achievement score of students in the group that did not participate in the discussions was $\bar{X}$=13.96 before application, and $\bar{X}$=33.43 following the application. Based on the assessment of these mentioned figures, it can be stated there is an increase in the overall success scores of all students.

As a result of the one-way variance analysis (ANOVA), which was implemented to determine whether there was a significant difference among the overall achievement scores of the students participating in the learning process in three different experimental environments, it was determined that there was statistically no significant difference [F(2,187)=0.027; *p*>.05]. This finding was interpreted that the prior knowledge levels of students about Computer Programming course before the application were similar. The results of the one-way variance analysis (ANOVA), which was implemented to determine whether there was a significant difference among the overall achievement scores of the students participating in the learning process in three different experimental environments after the application, are given in Table 3.

Table 3. *One-way variance analysis (ANOVA) of the posttest overall achievement scores of the student groups.*

Source of the Variance	Sum of Squares	*Sd*	Mean of Squares	F	*p*	Significant Difference
Intergroup	3336.257	2	1683.129	8.334	.000	GR1-GR3
Intragroup	37767.173	187	201.963			GR2-GR3

As can be seen in Table 3, as a result of the one-way variance analysis (ANOVA), which was implemented to determine whether there was a significant difference among the post-application overall achievement scores of the students participating the learning process in three different experimental environments, it was determined that there was a statistically significant difference [F(2,187)=8.334; *p*<.05]. The effect size (eta squared) calculated as a result of the test was determined as η^2 = 0.08. This eta-squared value demonstrate that the effect was in "medium" level. In other words, it can be mentioned that the 8% of the observed variance in the *posttest achievement score* dependent variable can be explained by the experimental conditions, and that it depends on the *participation* independent variable. Following this process, the complimentary post-hoc analysis techniques were applied in order to determine the source group of the significant difference detected through the ANOVA (Table 4).

Table 4. *Post-Hoc Scheffe Test results following the one-way variance analysis (ANOVA) that was employed to determine which sub-groups differed according to the posttest achievement scores.*

Groups		Difference in Means	*p*
GR1-Mandatory	GR3-Non-attendee	8.973*	.002
GR2-Voluntary	GR3-Non-attendee	9.145*	.002

**p<.01*

As a result of the Post-Hoc Scheffe Test following the one-way variance analysis (ANOVA), which was employed to determine which sub-groups differed according to the posttest achievement scores, it was determined that there was a statistically significant difference (at *p*<.01 level) between the mandatory participants and non-attendee participants in favor of the mandatory participants. Additionally, it was determined that there was a statistically significant difference (at *p*<.01 level) between the voluntary participants and non-attendee participants (Table 4). In line with these findings, it can be stated that the overall achievement levels of the students who participated in the discussions in the discussion-oriented flipped learning environments were significantly higher compared to those who did not participate in the discussions.

The effect size (Cohen’s d) obtained from the pretest-posttest mean scores of the students from the mandatory participation group was *d* = 1.89 (large effect). The effect size (Cohen’s *d*) obtained from the pretest-posttest mean scores of the students from the voluntary participation group was *d* = 2.18 (large effect). The effect size (Cohen’s d) obtained from the pretest-posttest mean scores of the students from the non-attendee group was calculated as *d* = 1.71 (large effect). Accordingly, it was interpreted that the effect of the participation independent variable on the pretest-posttest achievement mean scores was large effect.

3.2. Findings and Interpretations Concerning the Satisfaction Variable

3.2.1. *Findings concerning video satisfaction scores*

One-way variance analysis (ANOVA) was employed in order to determine whether the video satisfaction mean scores of the students in the groups differ on a group basis. As a conclusion of the analysis, the descriptive statistics concerning the video satisfaction variable comprising of 15 items are given in Table 5.

Table 5. *Mean and standard deviation values of video satisfaction scores of the student groups.*

Groups	*N*	$\bar{X}$	*Sd*	%	Min	Max
GR1 – Mandatory	69	59.29	11.31	79.05	17.00	75.00
GR2 – Voluntary	61	59.43	11.14	79.24	20.00	75.00
GR3 – Non-attendee	60	57.47	12.82	76.62	15.00	73.00
Total	190					

Examining Table 5, it is seen that the video satisfaction mean scores of the students participating in the discussion-oriented flipped learning environment in the mandatory group was $\bar{X}$=59.29 (79.05%), while it was $\bar{X}$=59.43 (79.24%) for those in the voluntary group. The video satisfaction mean scores of the students in the non-attendee group was $\bar{X}$=57.47 (76.625). The results of the one-way variance analysis (ANOVA) which was implemented to determine whether there was a significant difference among the video satisfaction mean scores of the students participating in the learning process in three different experimental environments are given in Table 6.

Table 6. *One-way variance analysis (ANOVA) of the video satisfaction scores of the students in the groups.*

Variance Source	Sum of Squares	*Sd*	Mean of Squares	F	*p*
Intergroup	146.809	2	73.404	.531	.589
Intragroup	25848.054	187	138.225		

As a result of the one-way variance analysis (ANOVA) which was implemented to determine whether there was a significant difference among the video satisfaction scores of the students participating in the learning process in three different experimental environments, it was determined that there was statistically no significant difference [F(2,187)=0.531, *p*>.05]. This finding is interpreted that the participation status of the students in the discussion environments did not cause a significant difference in the video satisfaction mean scores.

3.2.2. *Findings concerning the discussion satisfaction scores*

The results of the t-test which was employed in order to determine whether there was a significant difference among the discussion satisfaction levels of the students participating in discussions in two different experimental environments in the flipped learning environment are presented in Table 7.

Table 7. *Mean and standard deviation values of the student groups concerning their discussion satisfaction levels.*

Groups	*N*	$\bar{X}$	SS	%	Min	Max
GR1 – Mandatory	69	37.99	10.29	75.98	10.00	50.00
GR2 – Voluntary	61	33.75	9.83	67.50	11.00	50.00
Total	130					

In line with the data obtained from the discussion satisfaction scale, which was comprised of 10 items, it was determined that the discussion satisfaction level of the students participating in the discussions on a mandatory basis was $\bar{X}$=37.99 (75.98%), which was higher compared to $\bar{X}$=33.75 (67.50%), the mean score of those participated on a voluntary basis (Table 7). Paired sample t-test analysis was conducted in order to determine whether this difference was significant (Table 8).

Table 8. *t-test analysis results of the discussion satisfaction scores of the student groups.*

Participation Status	*N*	$\bar{X}$	*Sd*	*Sd*	*t*	*p*
GR1- Mandatory	69	37.99	10.291	128	2.390	.018
GR2- Voluntary	61	33.75	9.825			

Examining the *t*-test analysis results concerning the discussion satisfaction scores of the students in the groups (Table 8), it was determined that there was a statistically significant difference between the discussion satisfaction levels of students from the mandatory group and students from the voluntary group [*t*(128)=2.390, *p*<.05]. Accordingly, the discussion satisfaction levels of the students from the mandatory group were higher compared to those of the students in the voluntary group. In line with the results of the *t*-test, the effect size (eta squared) was calculated as $\eta^2 = 0.04$. Based on this effect size, it was interpreted that mandatory or voluntary participation status of students had a "low level" effect size on the discussion satisfaction scores.

3.3. Findings and Interpretations Concerning the High-Ordered Thinking Skills Variable

Descriptive statistics concerning the one-way variance analysis (ANOVA) results, which was used in order to determine whether there was a statistically significant difference in the high-ordered thinking skills of students participating in the learning process in three different experimental environments, are given in Table 9.

Table 9. *Mean and standard deviation values of student groups concerning the high-ordered thinking skill pretest-posttest scores.*

Groups	*N*	Pretest		Posttest	
		$\bar{X}$	*Sd*	$\bar{X}$	*Sd*
GR1 – Mandatory	69	2.78	6.51	27.71	17.81
GR2 – Voluntary	61	2.55	5.36	29.44	17.69
GR3 – Non-attendee	60	2.80	7.80	13.00	14.21
Total	190				

Examining Table 9 and as a result of the evaluation concerning the high-ordered thinking skills pretest and posttest scores, it can be stated that there is a general increase in the high-ordered thinking skill scores of all the students. As a result of the one-way variance analysis (ANOVA), which was employed to determine whether there was a statistically significant difference among the high-ordered thinking skill pre-application scores of students participating in the learning process in three different environments, it was determined that there was statistically no significant difference [F(2,187)=0.026; *p*>.05]. The results of the one-way variance analysis (ANOVA), which was employed to determine whether there was a statistically significant difference among the high-ordered thinking skill post-application scores of students participating in the learning process in three different environments, are given in Table 10.

Table 10. *One-way variance analysis (ANOVA) concerning the high-ordered thinking skill posttest scores of student groups.*

Source of the Variance	Sum of Squares	*Sd*	Mean of Squares	F	*p*	Significant Difference
Intergroup	9989.464	2	4994.732	17.863	.000	GR1-GR3
Intragroup	52287.252	187	279.611			GR2-GR3

As can be seen in Table 10, one-way variance analysis (ANOVA) was implemented in order to determine whether there was a significant difference among the post-application scores in high-ordered thinking skills of the students participating in the learning process in three different experimental environments. As a result of the analysis, it was determined that there was a statistically significant difference between the high-ordered thinking skills of students [F(2,187)=17.863; *p*<.05]. The effect size calculated after the test was $\eta^2 = 0.16$. This eta-squared figure demonstrated that there was a large effect. Subsequent to this process, the complimentary post-hoc analysis methods were implemented in order to determine the source group of the difference (Table 11).

Table 11. *Post-hoc Scheffe test results following the one-way variance analysis (ANOVA) that was employed to determine which sub-groups differed according to the high-ordered skill posttest scores.*

Group		Differences in Means	*p*
Mandatory	Non-attendee	14.710*	.000
Voluntary	Non-attendee	16.442*	.000

**p<.01*

According to Table 11, as a result of the Post-Hoc Scheffe Test following the one-way variance analysis (ANOVA) which was employed to determine which sub-groups differed according to the high-ordered thinking skill scores, it was determined that there was a statistically significant difference (at p<.01 level) between the mandatory participants and non-attendee participants in favor of the mandatory participants. Additionally, it was determined that there was a statistically significant difference (at p<.01 level) between the voluntary participants and non-attendee participants (Table 11). In line with these findings, it can be stated that participation in discussions in the flipped learning environments regardless of participating mandatorily or voluntarily, have a positive influence on the high-ordered thinking skills.

4. DISCUSSION and CONCLUSION

In this semi-experimental research, the impacts of participation status of students in the discussions in a discussion-oriented flipped learning environment on their achievement, satisfaction and high-ordered thinking skills were examined. The results obtained from the findings based on the experimental processes are listed below.

There is a significant difference among the pretest-posttest achievement scores of all student groups (mandatory, voluntary, non-attendee), who had a six-week learning experience in a discussion-oriented flipped learning environment. In other words, it can be mentioned that learning was experienced in all groups.

Comparing the overall achievement scores of the students based on their participation status in the discussions in the flipped learning environment, it was determined that the achievement levels of the students who mandatorily or voluntarily participated in the discussions compared to the non-attendees. According to this finding, it can be stated that turning the flipped learning environments into discussion-oriented environments will increase the achievement levels of students. Using discussions in flipped learning environments influences the learner interactions, and it can influence the achievement performances in a positive way. Zainuddin (2018) reported that using discussion environments in the flipped learning environments influenced the interactions of the learner in a positive way, which in turn, increased the achievement and satisfaction levels. Lack of interaction in flipped learning, which was the starting point of this study, was tried to be eliminated through a discussion environment that was integrated into flipped learning process. Thus, it can be stated that turning the flipped learning process into a discussion-oriented environment can provide an enhancement in the learner achievement level. There was no significant difference between the video satisfaction levels of the student groups participating in the flipped learning environment. Accordingly, when the video satisfaction mean scores of students are examined, it can be said that students who watch videos in a flipped learning environment are generally satisfied with the videos. Based on the fact that there was statistically no significant difference among the groups concerning the video satisfaction levels, it can be suggested to be emerging from that all groups were provided with the same video material.

A statistically significant difference was determined between the mandatory and voluntary participant groups in the discussions of the flipped learning environment, in favor of the mandatory participants. Accordingly, it can be stated that making it mandatory for the students to participate in discussions in the flipped learning environment can increase their discussion satisfaction levels.

It was determined that there was statistically no significant difference among the general environment satisfaction levels of the participant student groups in the flipped learning environment. Though not significant, the general environment satisfaction level of the students mandatorily participating in the discussions was higher compared to the other groups. Based on this finding, it can be stated that students in all groups were satisfied with the general environment. Davies et al. (2013) emphasized that flipped learning environment increased the satisfaction levels of students, which in turn had a positive impact on the achievement levels of the learner. In this study, the flipped learning method was applied to the three groups of students. Having a positive satisfaction level in all groups is a finding that is in parallel to those of similar studies in the literature.

While there was no difference among the high-ordered thinking skill pretest mean scores of the student groups participating in the flipped learning environment, it was determined that there was a significant difference among the high-ordered thinking skill posttest mean scores of the groups after the application. Accordingly, it was determined that at the end of the six-week application, the high-ordered thinking skill scores of the students who participated in the discussions regardless of participating mandatorily or voluntarily were significantly higher compared to those not participating in the discussions. It can be stated that regardless of voluntarily or mandatorily, participation in the discussions in a flipped learning environment has a positive impact on the high-ordered thinking skill levels compared to that of non-participation. Online discussion environments are the medium where students can practice their high-ordered thinking skills. As a conclusion of this study, it is considered that using the discussion environment has a positive impact on the development of high-ordered thinking skills of students.

In this research study, it was concluded that using a discussion platform in the flipped learning environment increases the achievement level of the learner. Based on this finding, it can be stated that the developers who will use the flipped learning method and prepare a flipped learning environment can create a more efficient teaching-learning environment by using the discussion environments together with the course videos.

In this quasi-experimental study, there is a limitation due to the pre-test and post-tests administered at six-week intervals. This situation, which is one of the weaknesses of the research, could not be controlled. It is recommended that subsequent investigators perform similar studies over a larger time period with a completely random sample distribution.

Acknowledgments

This research is derived from the doctoral thesis prepared by Erdi Okan Yılmaz (Uşak University) under the supervision of Prof. Dr. Nurettin Şimşek (Ankara University).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. Ethics Committee Number: Ankara University/Social and Humanities Ethics Committee, 2019-11/344.

Authorship Contribution Statement

All authors have equally contributed to all sections of this study.

Orcid

Erdi Okan Yilmaz https://orcid.org/0000-0002-7423-725X
Nurettin Simsek https://orcid.org/0000-0002-9319-1875

REFERENCES

Acedo, M. (2019). *10 pros and cons of a flipped classroom* [TeachThought]. *https://www.teachthought.com/learning/10-pros-cons-flipped-classroom/*

Aydın, B., & Demirer, V. (2016). Flipping the drawbacks of flipped classroom: Effective tools and recommendations. *Journal of Educational and Instructional Studies in The World*, *6*(1), 33-40.

Baker, J.W. (2000). *The "classroom flip": Using web course management tools to become the guide by the side.* 11th. International Conference on College Teaching and Learning. *https://digitalcommons.cedarville.edu/media_and_applied_communications_publications/15/*

Bates, J.E., Almekdash, H., & Gilchrest-Dunnam, M.J. (2017). The flipped college classroom. In Green, L.S., Banas, J.R., & Perkins, R.A. (Eds.), *The flipped college classroom conceptualized and re-conceptualized* (pp. 3-11). Springer.

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day.* International Society for Technology in Education.

Bergmann, J., & Sams, A. (2014). *Flipped learning: gateway to student engagement.* International Society for Technology in Education.

Bhagat, K.K., Chang, C.-N., & Chang, C.-Y. (2016). The impact of the flipped classroom on mathematics concept learning in high school. *Journal of Educational Technology & Society, 3*(19), 134-142. http://www.jstor.org/stable/jeductechsoci.19.3.134

Blau, I., & Shamir-Inbal, T. (2017). Re-designed flipped learning model in an academic course: The role of co-creation and co-regulation. *Computers & Education, 115*(1), 69-81. https://doi.org/10.1016/j.compedu.2017.07.014

Bolat, Y. (2016). Ters yüz edilmiş sınıflar ve eğitim bilişim ağı (EBA) [The flipped classes and education information network (EIN)]. *Journal of Human Sciences, 13*(2), 3373-3388. http://dx.doi.org/10.14687/jhs.v13i2.3952

Brewer, R., & Movahedazarhouligh, S. (2018). Successful stories and conflicts: A literature review on the effectiveness of flipped learning in higher education. *Journal of Computer Assisted Learning*, *34*(4), 409-416. https://doi.org/10.1111/jcal.12250

Chowdhury, T.R. (2017). *Engaging in isolation: Student engagement in a flipped classroom* [TeachThought]. *https://www.teachthought.com/technology/student-engagement-in-flipped-classroom/*

Davies, R.S., Dean, D.L., & Ball, N. (2013). Flipping the classroom and instructional technology integration in a college-level information systems spreadsheet course. *Educational Technology Research and Development, 4*(61), 563-580. https://doi.org/10.1007/s11423-013-9305-6

Davis, L., Neary, M.A. & Vaughn, S.E. (2013). Teaching advanced legal research in a flipped classroom. *Teaching Legal Research and Writing, 22*(1), 13-19.

Demirtaşlı, N. (2010). Üst düzey düşünme becerilerinin ölçülmesinde gündelik yaşam unsuru [Daily life element in measuring higher-order thinking skills]. *CİTO Eğitim: Kuram ve Uygulama*, *7*(1), 9-26. https://docplayer.biz.tr/60122283-Ust-duzey-dusunme-becerilerinin-olculmesinde-gundelik-yasam-unsuru.html

Du, S.-C., Fu, Z.-T., & Wang, Y. (2014). *The flipped classroom–advantages and challenges.* International Conference on Economic Management and Trade Cooperation. Atlantis Press.

Enfield, J. (2013). Looking at the impact of the flipped classroom model of instruction on undergraduate multimedia students at CSUN. *TechTrends, 6*(57), 17-27. https://doi.org/10.1007/s11528-013-0698-1

Gündüz, A.Y., & Akkoyunlu, B. (2016). Dönüştürülmüş sınıftan dönüştürülmüş öğrenmeye [From flipped classroom to flipped learning]. In A. İşman, F. Odabaşı & B. Akkoyunlu (Eds.), *Eğitim Teknolojileri Okumaları 2016*. (pp. 237 - 251). TOJET.

Hung, H.-T. (2018). Gamifying the flipped classroom using game-based learning materials. *ELT Journal.* 1-13. https://doi.org/10.1093/elt/ccx055

Jenkins, C. (2017). *The advantages and disadvantages of the flipped classroom* [Echo360]. *http://blog.echo360.com/blog/bid/59158/The-Advantages-and-Disadvantages-of-the-Flipped-Classroom*

Kardaş, F., & Yeşilyaprak, B. (2015). Eğitim ve öğretimde güncel bir yaklaşım: teknoloji destekli esnek öğrenme (flipped learning) modeli [A current approach to education: flipped learning model]. *Journal of Faculty of Educational Sciences, 48*(2), 103-121. https://doi.org/10.1501/Egifak_0000001366

Krueger, J. (2012). *Five reasons against the flipped classroom* [Stratostar]. *https://stratostar.net/five-reasons-against-the-flipped-classroom/*

Lage, M.J., Platt, G.J., & Treglia, M. (2000). Inverting the classroom: a gateway to creating an inclusive learning environment. *The Journal of Economic Education, 31*(1), 30-43. https://doi.org/10.2307/1183338

Lee, M.K., & Park, B.K. (2018). Effects of flipped learning using online materials in a surgical nursing practicum: a pilot stratified group-randomized trial. *Healthcare Informatics Research, 24*(1), 69-78. https://doi.org/10.4258/hir.2018.24.1.69

Lo, C.K., & Hew, K.F. (2017). A critical review of flipped classroom challenges in K-12 education: possible solutions and recommendations for future research. *Research and Practice in Technology Enhanced Learning, 4*(12), 1-22. https://doi.org/10.1186/s41039-016-0044-2

Marwedel, P., & Engel, M. (2014). Flipped classroom teaching for a cyber-physical system course – an adequate presence-based learning approach in the internet age. *IEEE Explore,* 11-15. https://doi.org/10.1109/EWME.2014.6877386

Milman, N.B. (2012). The flipped classroom strategy: What is it and how can it best be used? *Distance Learning, 3*(9), 85-87.

Ng, W. (2015). *New digital technology in education.* Springer. https://doi.org/10.1007/978-3-319-05822-1

Nouri, J. (2016). The flipped classroom: for active, effective and increased learning – especially for low achievers. *International Journal of Educational Technology in Higher Education, 13*(33), 1-10. https://doi.org/10.1186/s41239-016-0032-z

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *Internet and Higher Education, 25*(1), 85-95. https://doi.org/10.1016/j.iheduc.2015.02.002

Ramaglia, H. (2015). *The flipped mathematics classroom: A mixed methods study examining achievement, active learning, and perception*. Kansas State University, USA.

Strayer, J.F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learning Environments Research, 2*(15), 171-193. https://doi.org/10.1007/s10984-012-9108-4

Talbert, R. (2012). *Inverted classroom* [Scholar Works]. *https://scholarworks.gvsu.edu/cgi/viewcontent.cgi?article=1183&context=colleagues*

Tetreault, P.L. (2013). *The flipped classroom: Cultivating student engagement.* University of Victoria. http://hdl.handle.net/1828/5086

Topalak, Ş. (2016). *Çevrilmiş öğrenme modelinin başlangıç seviyesi piyano öğretimine etkisi [The effect of flipped classroom model on the beginner level piano teaching]* [Doctoral dissertation]. Inönü University.

Torun, F., & Dargut, T. (2015). Mobil öğrenme ortamlarında ters yüz sınıf modelinin gerçekleştirilebilirliği üzerine bir öneri [A proposal on the feasibility of the flipped classroom model in mobile learning environments]. *Adnan Menderes University Faculty of Education Journal of Education Sciences, 6*(2), 20-29.

Turan, Z., & Göktaş, Y. (2015). Yükseköğretimde yeni bir yaklaşım: öğrencilerin ters yüz sınıf yöntemine ilişkin görüşleri [A new approach in higher education: the students' views on flipped classroom method]. *Journal of Higher Education and Science*, *2*(1), 156-164. https://doi.org/10.5961/jhes.2015.118

Üstündağ, M.T. (2012). *Çevrimiçi öğrenme ortamlarında uyarlanmış sosyal etkileşim araçlarının öğrencilerin akademik başarılarına ve sosyal bulunuşluk algılarına etkisi [The effects of adaptable social interaction tools on students' academic achievements and perceptions of social presence in online learning environments]* [Doctoral dissertation]. Gazi University.

Wright, R.J. (2010). Multifaceted assessment for early childhood education. Sage Publications.

Yılmaz, R. (2017). Exploring the role of e-learning readiness on student satisfaction and motivation in flipped classroom. *Computers in Human Behavior, 70*(1), 251-260. https://doi.org/10.1016/j.chb.2016.12.085

Zainuddin, Z. (2018). Implementing Moore's model of interaction in a flipped-class instruction. *The Online Journal of Distance Education and e-Learning, 6*(3), 10-20.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 162–178

https://doi.org/10.21449/ijate.1132980

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of this article.

The role of teacher support and in-class teaching practices on reading performance: Evidence from PISA 2018 outcomes for Türkiye

B. Umit Bozkurt[1,*]

[1]Bolu Abant Izzet Baysal University, Faculty of Education, Department of Turkish Language Education, Bolu, Türkiye

ARTICLE HISTORY

Received: Jun. 20, 2022
Revised: Oct. 3, 2022
Accepted: Oct. 11, 2022

Keywords:
PISA 2018,
Reading performance,
Teacher support,
Feedback,
Stimulation,
In-class teaching practices
.

Abstract: The study deals with the variation of Turkish students' reading comprehension performance according to *perceived teacher support* and *reading activities in the classroom.* This study, which is grounded on the data drawn from the PISA 2018 database, investigates the relationship between certain variables. In the analyses performed on the PISA IDE server, the PISA 2018 reading literacy general averages of Türkiye were associated with the identified variables, and the differences in the averages were examined. As a result, perceived teacher support, teacher's adaptation of the course, and stimulation of reading engagement have a positive relationship with reading comprehension; however, it was found out that the frequency of receiving feedback had a negative relationship with reading performance. In addition, the general reading average of the students who reported that they had not performed activities such as *summarizing, comparing the content of the text with their own experiences, comparing the text they have read with other texts written on similar topics,* and *writing about the text that has been read* was much higher than those who reported that they had performed these activities. These results have strengthened the conclusion that teachers give feedback to poor readers more frequently. On the other hand, it is possible that good readers may find the learning activities in the course inadequate. In summary, reading comprehension performance is positively or negatively affected by teacher support, adaptive instruction, feedback, and engagement in reading activities in the classroom.

1. INTRODUCTION

Reading comprehension is a skill that develops in the process, includes various stages, and deepens with different layers. The monitoring-based guidance of teachers makes this process effective and efficient. In addition to its cognitive multilayeredness, the reading process can reach an effective level with pre-reading, reading and post-reading activities inside and out-side the classroom. Kutlu *et al*. (2019) point out that reading comprehension is a multi-dimensional process that is affected by the characteristics of the individual, the text and the context. It also consists of many subcomponents and emphasizes that the ways to be followed for the evaluation

*Corresponding Author: B. Umit BOZKURT ✉ umitbozkurt@gmail.com Bolu Izzet Baysal University, Faculty of Education, Department of Turkish Language Education, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

of this skill should be versatile and comprehensive. An effective reading process should be built on a supportive classroom climate in which the teacher monitors the student and gives feedback, motivation, and encouragement.

An important dimension of the in-class studies is assessment activities aimed at monitoring the student's development, strengths, and weaknesses. It is known that the assessment affects the academic success of the student not only with its cognitive dimensions but also with its affective dimensions. Students' interactions with their teachers play an important role in their learning and attitude. As Federici and Skaalvik (2014) point out, students need to feel that their teachers care about them and their success in order to fully participate in learning activities and perform at their best. The work of Klem and Connell (2004) and Wang and Holcombe (2010) also show that teacher support is important for student engagement and that students' perceptions for the school environment affect their academic achievement directly or indirectly. Teachers support their students by encouraging, motivating, listening helping them, and providing them with the necessary resources of knowledge and materials.

Teacher support is conceptualized in the literature with various contents. Briefly, it is framed as 'information, instruments, feelings or evaluation support for the student. Malecki and Demaray (2003) explain that most of the classifications used can fit into the following common framework: informational support is to give suggestions in a specific area; instrumental support is to provide the necessary resources. While *emotional support* is to inspire confidence, interest or empathy, appraisal support is the giving evaluative feedback to each student. Providing feedback is an important part of teacher support (Sukhram & Monda-Amaya, 2017).

Teacher support can also be classified in two types as emotional support (empathy, sincerity, encouragement, interest, etc.) within the classroom and instrument support (for instance, teachers help students to solve a problem or accomplish a difficult task). Instrument support includes students' perceptions of resources and practical help. These may include teachers' questioning, clarification, correction, elaboration, and modelling behaviours that contribute to comprehension, problem solving, or skill development (Federici & Skaalvik, 2014).

Various studies reveal that emotional support from teachers is associated with students' positive emotions, attitudes, and behaviours such as class participation, effort, low anxiety levels, and high internal motivation (Federici & Skaalvik, 2014; Guess & McCane-Bowling, 2016; Lee, 2012; Ruzek *et al.,* 2016; Sakiz, Pape, & Hoy, 2012). Instrumental support is in the form of concrete and practical assistance that has a strong and direct relationship with students' low level of anxiety, effort, and internal motivation (Federici & Skaalvik, 2014). Supportive teacher-student relationships are significantly associated with student engagement (Lee, 2012). When teachers are more emotionally supportive, there is an increase in students' behavioural engagement and motivation (Ruzek *et al*., 2016). Sakiz, Pape and Hoy (2012) indicated that the emotional support that students perceive encourages academic self-efficacy and academic effort. Guess and McCane-Bowling (2016) argue that supportive teachers create students who are more satisfied with their lives. Lei, Cui, and Chiu (2018) who conducted a meta-analysis (effect size, 121) of 65 studies found that teacher support was significantly associated with students' academic emotions (emotional experiences such as fun, hopelessness, boredom, anxiety, and anger, which can affect learning outcomes). They also reported that these relationships could be treated as positive and negative connections.

Studies also highlight the link between teacher support and students' academic success. The supportive teacher-student relationship influences student achievement, both directly and indirectly, with a greater sense of commitment to school (Hughes *et al*., 2008; Klem & Connell, 2004; Lee, 2012; Reyes *et al.,* 2012; Wang & Holcombe, 2010). Malecki and Demaray (2003) found out that perceived emotional support from teachers was the important and only predictor of students' social skills and academic competence. In addition, Dolapçıoğlu (2019) pointed out that students' relationship levels with their teachers were higher in the courses they were successful in.

When the subject is customized in the context of reading ability, the relationship of teacher support on reading performance stands out. It is important to note that the teacher-student relationship (Lee, 2012) and teacher support perceived by students (Ma, Luo, & Xiao, 2021; Ma, Xiao, & Hau, 2022), have an impact on reading skills.

The teacher's instructional activities in the classroom are another variable that has an impact on reading comprehension. These activities include encouraging students with questions, giving feedback, relating the text to the preliminary experiences, establishing in-text and out-of-text relationships, making intertextual comparisons, writing, and summarizing. These are effective in maintaining *engagement in reading*. The stimulation of reading engagement refers to supporting students' motivation and providing them with opportunities (Afflerbach & Harrison, 2017; Merga, 2020). Participation/dedication in reading is vital for reading performance (Lee *et al.,* 2021). Lei, Wen, Li, Kong, Chen, and Li (2019) concluded that teacher support through metacognitive strategies improved reading comprehension. Gambrell (1996) also emphasizes the critical role of the teacher in creating a classroom culture that encourages reading motivation.

In Türkiye, the interest shown in the role of the teacher in students' reading performance is little if any. However, the reasons why Turkish students' reading comprehension levels are far below expectations in international and national student monitoring programs should be investigated from various aspects. Approximately 67% of the 4th and 8th-grade students in the field of Turkish language in the ABIDE (Monitoring and Evaluation of Academic Skills) project (2018) were in the intermediate and below levels (Parlak, 2019; Yıldırım and Ozgurluk, 2019). A similar situation was observed in the central examinations carried out to be placed in secondary education schools. In 2022 and 2021, the average number of correct answers of students in the Turkish language test was 9 out of 20, and the number of correct answers for 63% of students was between 0 and 10. In 2020, there was an average of 7 correct answers in the Turkish language test (MEB, 2020; 2021; 2022). The PISA 2015 and 2018 results also showed that there were some fundamental problems in reading comprehension. The reading literacy average of 15-year-old Turkish students was below the OECD average, and more than half of the students were at the second level or below (OECD, 2016; 2019).

With the data of large-scale monitoring projects such as PISA, PIRLS, or ABIDE, significant inferences on the depths of the education system can be obtained. In studies carried out in Türkiye, reading comprehension achievement was widely examined in relation to the number of read books, the educational background of parents, and socio-economic level. Although the relationship between reading performance and the role of the teacher and in-class activities was clearly shown in the literature, this issue has not been sufficiently emphasized as a part of the classroom teaching and evaluation process.

Within the scope of PISA 2018, the classroom climate, the teacher's initiatives, behaviours, and the effect of classroom teaching practices on reading performance were discussed in detail in the language courses. The change in categories such as teacher enthusiasm, teacher support, and teacher behaviour from the point of view of the students was examined in general terms in terms of countries. In this study, *teacher support, teacher feedback, adaptive instruction, teacher stimulation of reading engagement,* and *in-class reading activities* were discussed in regard to Turkish students' reading performance in PISA 2018. In this respect, it is foreseen that significant inferences can be made for the development of reading in the Turkish education system from the results of the study.

2. METHOD

2.1. The Database and Sample

This study has a sectional design that examines the relationship between student reading achievement and certain variables in PISA 2018 dataset. The reading scale and student survey data of Türkiye sample were taken from PISA 2018 database (https://pisadataexplorer.oecd.org/ide/idepisa/) by analyzing the relationship of the variables to be investigated.

Nomenclature of Territorial Units for Statistics (NUTS) was used to determine the Türkiye sample of the PISA 2018 project. Accordingly, 186 schools representing 12 regions and 6890 students participated in the study with stratified sampling. 44% of the 15-year-old students representing Türkiye are educated in Anatolian High Schools, 31% in Vocational and Technical Anatolian High Schools and 14% in Anatolian Imam-Hatip High Schools. 0.3% of the students are at the secondary school level. 49.6% of the sample of Türkiye is female and 50.4% is male.

2.2. Data Analysis

In the process, through the data analysis tool offered by OECD, our analyses that provided the basis of this research have been carried out, and reports were generated from PISA datasets. In the secondary analyses conducted on the server, Türkiye's PISA 2018 *reading ability scale*: *Overall Reading*, the following variables reported by the students were correlated: *Teacher support, emotional support, feedback, adaptive instruction, teacher stimulating of reading engagement,* and *in-class reading activities* in Turkish/Turkish Language and Literature course.

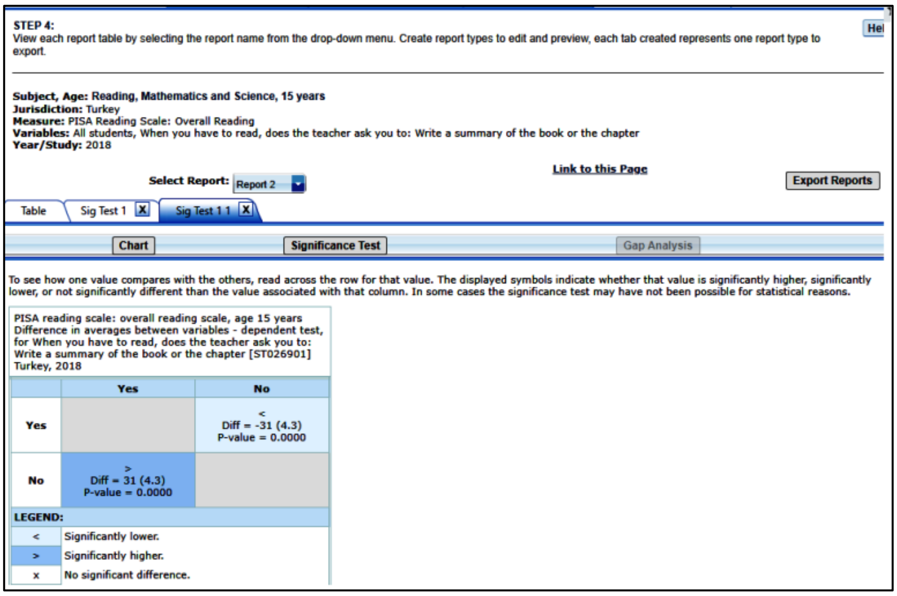
Adaptive instruction is inferred from students' responses to the question of ST212; teacher feedback was obtained using students' responses to ST104 that a trend question; teachers' stimulation of reading engagement was obtained based on a trend question (ST152) from PISA 2009; teacher support was inferred from students' responses to ST100; and teacher-directed instruction was gathered from ST102. The details of other variables can be seen in Table 1.

Table 1. *List of categories and items numbers.*

Categories of analysis		PISA ITEMS ID
Student-teacher relations (reported by students)	Teacher support	ST100
	Teacher feedback	ST104
Engagement in reading activities	Teacher-directed instruction	ST102
	Teaching practices in Turkish language course	ST153
Classroom instruction in reading-teacher strategies	Teachers' stimulation of reading engagement	ST152
Self-related cognition related to learning	Teacher emotional support	ST211
	Adaptive instruction	ST212

The screenshot of the system enabling secondary analysis at PISA 2018 database was presented in Figure 1. In the analysis, it was determined whether there was a significant difference between the variables in terms of average reading scores. The p values were presented in the tables.

Figure 1. *PISA IDE data analysis tool.*



STEP 4:
View each report table by selecting the report name from the drop-down menu. Create report types to edit and preview, each tab created represents one report type to export.

Subject, Age: Reading, Mathematics and Science, 15 years
Jurisdiction: Turkey
Measure: PISA Reading Scale: Overall Reading
Variables: All students, When you have to read, does the teacher ask you to: Write a summary of the book or the chapter
Year/Study: 2018

Select Report: Report 2

To see how one value compares with the others, read across the row for that value. The displayed symbols indicate whether that value is significantly higher, significantly lower, or not significantly different than the value associated with that column. In some cases the significance test may have not been possible for statistical reasons.

PISA reading scale: overall reading scale, age 15 years
Difference in averages between variables - dependent test,
for When you have to read, does the teacher ask you to:
Write a summary of the book or the chapter [ST026901]
Turkey, 2018

	Yes	No

Yes		<
		Diff = -31 (4.3)
		P-value = 0.0000

No	>
	Diff = 31 (4.3)
	P-value = 0.0000	

LEGEND:

<	Significantly lower.
>	Significantly higher.
x	No significant difference.

3. RESULTS

In addition to reading performance in PISA 2018, the results obtained from the data collected for the 'classroom climate perceived by the students' in Turkish/Turkish Language and Literature courses were discussed under the subheadings of *teacher support (help and emotional support), feedback, adaptive instruction, the stimulation of reading engagement, and in-class reading practices.*

3.1. Teacher Support, Feedback, Adaptive Instruction, and Reading Performance

According to the perception of receiving help as an indicator of teacher support, the average overall reading score of the Türkiye sample varies. In the Turkish/Turkish Language and Literature course, the average reading score of the students who stated, "Teacher helps students with their learning" and those who have a negative perception of help were different.

Table 2. *Help perception and reading performance.*

	Every lesson (468)	Most lessons (472)	Some lessons (452)
Most lessons (472)	Diff = 4 (3.0) *p*-value = 0.2072		
Some lessons (452)	**Diff = 16 (4.6)** ***p*-value = 0.0005**	**Diff = 20 (4.1)** ***p*-value = 0.0000**	
Never or hardly ever (452)	Diff = 15 (9.1) *p*-value = 0.0897	**Diff = 19 (8.2)** ***p*-value = 0.0190**	Diff = 1 (7.9) *p*-value = 0.9445

As can be seen in Table 2, the average reading score of students who reported that the teacher helped in "most lessons" was considerably higher than those who reported that the teacher "sometimes" helped or "never" helped.

A similar situation was with regard to additional assistance. According to the answers given to the question "The teacher gives extra help when students need it", Türkiye's general reading scale average scores varied.

Table 3. *Perception of extra help and reading performance.*

	Every lesson (466)	Most lessons (475)	Some lessons (459)
Most lessons (475)	**Diff = 9 (2.8)** ***p*-value = 0.0017**		
Some lessons (459)	**Diff = 7 (3.3)** ***p*-value = 0.0244**	**Diff = 16 (3.7)** ***p*-value = 0.0000**	
Never or hardly ever (451)	**Diff = 15 (6.3)** ***p*-value = 0.0178**	**Diff = 24 (6.6)** ***p*-value = 0.0003**	Diff = 7 (5.5) *p*-value = 0.1785

As can be seen in Table 3, there is a significant difference between the reading comprehension performance of the students who reported that the teacher gives extra help in every or most lessons and the students who stated that they hardly helped. The average reading score of students who report that the teacher helped in "most lessons" is considerably higher than students who reported that the teacher "sometimes" helped or "never" helped. In other words, when the perception of receiving help is positive, reading performance is also high.

When we look at the relationship the students establish with the teacher, which is *the emotional support*, it is seen that there is a difference in reading performance. The resulting difference points to a complex situation (see Table 4).

Table 4. *Teacher listening to and paying attention to students' views.*

	Strongly disagree (448)	Disagree (472)	Agree (468)
Disagree (472)	**Diff = 24 (4.8)** ***p*-value = 0.0000**		
Agree (468)	**Diff = 20 (4.7)** ***p*-value = 0.0000**	Diff = -3 (3.5) *p*-value = 0.3225	
Strongly agree (470)	**Diff = 23 (5.2)** ***p*-value = 0.0000**	Diff = 1 (4.6) *p*-value = 0.8169	Diff = 2 (3.5) *p*-value = 0.4945

There is a significant difference between the reading success of the students who stated that, "I strongly disagree" with the statement "The teacher listened to and paid attention to my views on how to do things" and those who stated, "I do not agree", "I agree" and "I totally agree", and this difference is statistically significant.

A similar situation is seen with students who reported that, “The teacher made me feel confident in my ability to do well in the course”.

Table 5. *Ensuring that the teacher has confidence in the students' abilities.*

	Strongly disagree (446)	Disagree (481)	Agree (469)
Disagree (481)	**Diff = 35 (4.2)** ***p*-value = 0.0000**		
Agree (469)	**Diff = 23 (4.3)** ***p*-value = 0.0000**	**Diff = 12 (2.8)** ***p*-value = 0.0000**	
Strongly agree (458)	**Diff = 13 (5.4)** ***p*-value = 0.0193**	**Diff = 22 (4.8)** ***p*-value = 0.0000**	**Diff = 10 (4.4)** ***p*-value = 0.0190**

As can be understood from Table 5, the big difference, here, is poor reading performance, especially among students who firmly stated that "the teacher does not listen to their views' and "the teacher doesn't enable them to feel confident" in class. However, student responses do not indicate a linear development.

The feedback perception of the students in the Turkish/Turkish Language and Literature course is also seen to be related to the average scores of the general reading scale (see Table 6).

Table 6. *Feedback: Powerful aspects.*

	Never or almost never (468)	Some lessons (459)	Many lessons (474)
Some lessons (459)	**Diff = 9 (2.9)** ***p*-value = 0.0016**		
Many lessons (474)	Diff = 6 (4.5) *p*-value = 0.1959	**Diff = 15 (3.5)** ***p*-value = 0.0000**	
(Almost) every lesson (468)	Diff = 0 (5.4) *p*-value = 0.9671	Diff = 9 (4.6) *p*-value = 0.0575	Diff = 6 (3.8) *p*-value = 0.1146

The reading score of the students who thought that they receive feedback on their good aspects in "most courses" is significantly higher than the those who thought that they receive feedback on their good aspects in "some courses". The scores of the students who thought that they had never received any feedback have not changed compared to those who thought that they had received some feedback in each lesson.

An inverse relationship emerged between students who reported *receiving feedback from the teacher on how to improve themselves* and students who reported that they did not.

Table 7. *Feedback: Aspects that could be improved.*

	Never or almost never (473)	Some lessons (467)	Many lessons (466)
Some lessons (467)	**Diff = 6 (3.1)** ***p*-value = 0.0458**		
Many lessons (466)	Diff = 8 (4.9) *p*-value = 0.1218	Diff = 1 (3.4) *p*-value = 0.7072	
(Almost) every lesson (459)	**Diff = 14 (4.7)** ***p*-value = 0.0033**	**Diff = 8 (3.6)** ***p*-value = 0.0366**	Diff = 6 (3.8) *p*-value = 0.0953

As can be seen in Table 7, students with a negative perception of feedback on the aspects that could be improved have a higher average reading score than students with positive feedback. It should be noted that as the perception regarding the rate of reporting feedback decreases, so does the reading performance score. The same situation was also revealed in the perception of feedback about which areas students can still improve themselves (see Table 8).

Table 8. *Feedback: Areas for improvement.*

	Never or almost never (482)	Some lessons (462)	Many lessons (461)
Some lessons (462)	**Diff = 19 (3.0)** ***p*-value = 0.0000**		
Many lessons (461)	**Diff = 21 (4.9)** ***p*-value = 0.0000**	Diff = 1 (3.6) *p*-value = 0.6986	
(Almost) every lesson (459)	**Diff = 22 (4.6)** ***p*-value = 0.0000**	Diff = 3 (3.7) *p*-value = 0.3849	Diff = 2 (3.9) *p*-value = 0.6368

It is understood that students' reading performance varies according to the perception of positive or negative feedback. As reading performance improves, the frequency of receiving feedback decreases. This suggests that feedback expectations of students who are successful in reading are also high. On the other hand, there is a high probability that teachers give feedback to poor readers more frequently.

Reading comprehension performance shows a linear relationship with *the teacher's adaptation of the instruction* according to the level and needs.

Table 9. *The teachers' adaptation of the instruction to the needs and level of the class.*

	Never or almost never (442)	Some lessons (452)	Many lessons (475)
Some lessons (452)	Diff = 9 (5.4) *p*-value = 0.0836		
Many lessons (475)	**Diff = 33 (5.2)** ***p*-value = 0.0000**	**Diff = 24 (2.8)** ***p*-value = 0.0000**	
(Almost) every lesson (483)	**Diff = 41 (6.5)** ***p*-value = 0.0000**	**Diff = 31 (4.1)** ***p*-value = 0.0000**	**Diff = 8 (3.0)** ***p*-value = 0.0094**

As can be seen in Table 9, the average reading score of students who stated that "almost every lesson" was organized according to the level and need of the class was much higher than the students who thought that the lesson was "almost never" adapted to the class, and the difference was significant.

From the students' point of view, individual assistance to students who had difficulties in the course made a significant difference in reading scores. The reading performance of the students who reported that they were helped when they had difficulty in "almost every lesson" was higher than the others. In terms of performance level, there were students reporting that they were "almost never" helped or "sometimes" helped when they had difficulties. This can be seen from Table 10.

Table 10. *Helping the student who is struggling individually.*

	Never or almost never (461)	Some lessons (461)	Many lessons (469)
Some lessons (461)	Diff = 1 (4.2) *p*-value = 0.8608		
Many lessons (469)	Diff = 9 (5.1) *p*-value = 0.0957	**Diff = 8 (3.4)** ***p*-value = 0.0230**	
(Almost) every lesson (474)	**Diff = 13 (4.9)** ***p*-value = 0.0064**	**Diff = 13 (3.4)** ***p*-value = 0.0003**	Diff = 5 (3.5) *p*-value = 0.1746

3.1. Teachers' Stimulation of Reading Engagement, Classroom Reading Practices and Reading Performance

Teachers' stimulation of reading engagement is significant in reading performance. The average reading score seems linear, as the teacher stimulates the student to explain his or her views on the text read in the lesson. The difference that arises in this regard is also very remarkable.

Table 11. *Stimulate: Express opinion.*

	Never or hardly ever (442)	Some lessons (451)	Most lessons (480)
Some lessons (451)	**Diff = 9 (4.7)** ***p*-value = 0.0464**		
Most lessons (480)	**Diff = 38 (5.5)** ***p*-value = 0.0000**	**Diff = 28 (3.4)** ***p*-value = 0.0000**	
All lessons (483)	**Diff = 41 (5.6)** ***p*-value = 0.0000**	**Diff = 32 (3.6)** ***p*-value = 0.0000**	Diff = 3 (3.6) *p*-value = 0.3745

As can be seen in Table 11, students who reported that they were not encouraged to express their own opinions have a significantly lower reading average. Students who stated that they were encouraged to express their views on “every course or most courses” had higher reading performance.

Stimulating students *to associate the read text with their own experiences* also affects their reading comprehension performance (see Table 12).

Table 12. *Stimulate: Relate to lives.*

	Never or hardly ever (465)	Some lessons (461)	Most lessons (471)
Some lessons (461)	Diff = 4 (3.2) *p*-value = 0.2478		
Most lessons (471)	Diff = 6 (3.6) *p*-value = 0.0843	**Diff = 10 (3.1)** ***p*-value = 0.0014**	
All lessons (474)	**Diff = 9 (4.7)** ***p*-value = 0.0476**	**Diff = 13 (3.7)** ***p*-value = 0.0006**	Diff = 3 (4.0) *p*-value = 0.4556

Motivating the participation in the course with questions also increases the level of reading comprehension (see Table 13).

Table 13. *Strategies: Motivating questions.*

	Never or hardly ever (464)	Some lessons (462)	Most lessons (469)
Some lessons (462)	Diff = -1 (4.8) *p*-value = 0.7609		
Most lessons (469)	Diff = 5 (5.1) *p*-value = 0.3037	**Diff = 7 (2.9)** ***p*-value = 0.0191**	
All lessons (470)	Diff = 7 (5.8) *p*-value = 0.2453	**Diff = 8 (3.7)** ***p*-value = 0.0256**	Diff = 1 (3.1) *p*-value = 0.6379

In the Turkish/Turkish Language and Literature course, as the frequency of motivating the student's participation in the course with questions increases, the level of reading comprehension also increases. There is a significant difference between the general reading scores of the students who stated that they were motivated by questions in "some courses" and those who stated that they were motivated in "all courses."

The reading average scores of Turkish students participating in PISA 2018 differ according to how the teacher evaluates their in-class practices in reading activities. There is a big difference between the reading scale scores of the students who stated, "the teacher makes a short summary of the previous lesson at the beginning of the lesson" and the students who had negative opinions on this subject, and this difference indicates an inverse relationship (see Table 14).

Table 14. *Teacher giving a summary of the previous lesson at the beginning of the lesson.*

	Every lesson (449)	Most lessons (467)	Some lessons (476)
Most lessons (467)	**Diff = 18 (3.2)** ***p*-value = 0.0000**		
Some lessons (476)	**Diff = 28 (3.3)** ***p*-value = 0.0000**	**Diff = 9 (2.6)** ***p*-value = 0.0003**	
Never or hardly ever (492)	**Diff = 43 (6.9)** ***p*-value = 0.0000**	**Diff = 25 (5.7)** ***p*-value = 0.0000**	**Diff = 16 (5.9)** ***p*-value = 0.0082**

There is a 43-point difference between the students who reported that the teacher "every lesson" made a short summary of the previous lesson at the beginning of the Turkish/Turkish Language lesson and the students who reported that this practice was never made. What is remarkable is that the reading performance score increases as the frequency of teacher reports concerning the summary of the lesson decreases. This may indicate that the expectations of students with high reading performance have not been met. From another point of view, the positive perceptions of students with poor reading performance suggest that their awareness of classroom activities is poor.

Positive and negative responses to *activities related to reading a book or a chapter* result in different appearances in reading performance.

The average reading score of the students who reported that the *summary* of the book or book chapter read in the course was written is lower than the students who reported that the summary activity was not done. This difference is high and significant. This can be seen in Table 15.

Table 15. *Summarizing.*

	No (489)
Yes (458)	**Diff = 31 (4.3)** ***p*-value = 0.0000**

As with the summarization activity, *small group discussion* also indicates an inverse relationship. In the Türkiye sample, the average reading score of the students who reported that small group discussions were held with students reading the same book was lower than the students who reported that they did not, and this difference was significantly higher (see in Table 16).

Table 16. *Small group discussion with students reading the same text.*

	No (476)
Yes (455)	**Diff = 22 (2.8)** ***p*-value = 0.0000**

The same situation is seen in *the activity of comparing the content of the text read with their own experiences*. The reading score of students who reported that this activity was not done was significantly higher than the students who stated that it was done (see in Table 17).

Table 17. *Comparing the content of the text with their own experiences.*

	No (482)
Yes (453)	**Diff = 29 (2.9)** ***p*-value = 0.0000**

In addition, the reading performance of the students who reported that *the text read in the courses was compared with other texts written on similar topics* is lower and statistically significant This can be seen in Table 18.

Table 18. *Comparison with other texts on similar topics.*

	No (481)
Yes (455)	**Diff = 26 (2.6)** ***p*-value = 0.0000**

The average reading score of students who gave a positive opinion about *the writing that was done on the text that was read* was considerably lower than that of students who gave a negative opinion, and this difference is significant. the values can be seen in Table 19.

Table 19. *Writing a text related to the text being read.*

	No (476)
Yes (458)	**Diff = 19 (2.9)** ***p*-value = 0.0000**

As can be seen, the average overall reading score of the students who reported that the *activities of summarizing the text, comparing their own experiences with the content of the text, comparing the text with other texts on similar topics,* and *writing* were not done are much higher than those students who reported that these activities were done. This may be due to the fact that good readers find the teacher's activities inadequate in the lesson, or it may be due to the poor readers' inability to correctly define the activities in the classroom.

4. DISCUSSION

Based on the PISA 2018 data, this study focuses on the role of teacher support, feedback and teaching practices in Turkish/Turkish Language and Literature courses on reading comprehension performance in the Türkiye sample. According to the findings obtained from PISA 2018, there is a significant difference between the average reading score of students whose *perception of the help from the teacher* is positive and those who are negative in the Turkish/Turkish Language and Literature course. The average reading score of students whose perception of receiving help from the teacher is positive is considerably higher than the others. Accordingly, when the perception of help from the teacher is positive, reading performance is also high. Karip (2020) also evaluated the findings in Türkiye in general and found out that students' reading performance scores increased as the teacher support increased. Across OECD, students who reported receiving more teacher support scored lower in reading. For example, participants in schools where teachers often show interest in each student's learning scored an average of 479, while students in schools where teachers report little interest in each student's learning scored an average of 491 (OECD, 2019).

When the relationship that the students establish with the teacher is examined, namely *the emotional support*, it is seen that there is a significant difference in reading performance. There is a remarkable difference between the students who stated ($\bar{x}$ =448) that the teacher "absolutely did not listen" to the students' opinions about how to do something and the reading performance of the other students. The same is true for the students who "strongly disagree" that their teacher builds a sense of confidence that they can succeed. The big difference, here, stands out as poor reading performance, especially for students who firmly state that the teacher does not listen to their opinions and "don't make them feel confident" in class. However, student responses do not indicate a linear development. Karip (2020), in his study, reported that while 64% of students in Türkiye stated that their teacher created a sense of confidence in them that they could succeed; 62% of students thought that the teacher listened to their own views on how to do something. These findings show that the emotional support provided by teachers according to students' statements in Türkiye remains at a lower level than the OECD average. Meşe-Soytürk (2020) investigated teacher support including emotional support and found out that the highest impact on the reading skills of 15-year-old students studying in Türkiye was positively related to the classroom discipline, family support, reading competence perceptions, feeling of a sense of belonging to the school, respectively, and negatively related to teacher support. Karaman (2022) examined the relationship between teacher behavior and reading performance in PISA 2018 and found that the students who felt supported by their teachers showed higher performance in reading literacy.

The feedback perception of the students in the Turkish/Turkish Language and Literature course was also seen to be related to the average scores of the general reading scale. Students who think they have received feedback on their good attributes in "most subjects" have a reading score significantly higher than those who think they have received feedback on their good attributes in "some lessons". There is a negative relationship between students who reported that they received feedback from the teacher on how to improve themselves and those who reported that they did not. Students with a negative perception of feedback for the purpose of improvement had higher average reading scores than students with positive feedback. The same situation arose in the perception of feedback given so that students could improve themselves. As reading performance increased, the frequency of those stating that he or she received feedback decreased. This suggests that students who were successful in reading also have high feedback expectations. It is also possible that teachers give more frequent feedback to poor readers. Karaman (2022) stated that the teacher feedback was negatively associated with reading performance. Safari (2020) found that teachers in countries above the OECD average often provide feedback and better reading materials to their students than teachers in countries below the average. This result also explains the negative relationship seen in Türkiye. Göçer and Şentürk (2019) pointed out that Turkish teachers used descriptive, process-based, and written feedback less than giving evaluative and verbal feedback for the whole class, and that Turkish language teachers had consensus on the importance of giving feedback in the text processing operation, and they had problems with when, how and which type of feedback could be given to which skill area. Karip (2020) stated that approximately one-fifth of students in Türkiye could not receive feedback from their teachers about their strengths, and how they could improve their performance and weaknesses which they could improve themselves. When the PISA 2018 results are evaluated in terms of the participating countries in general, it is seen that only from 10% to 15% of the students received feedback. In OECD economics specifically, less than 10% of students reported receiving feedback on their strengths "every or almost every lesson", and more importantly, many students reported that they received feedback "never or almost never" (OECD, 2019).

Adaptive instruction is another variable associated with reading comprehension. Reading comprehension performance shows a linear development as *the teacher adapts the lesson according to the level and needs*. The average reading score of the students who stated that "almost every lesson ($\bar{x}$=483)" is organized according to the level and needs of the class is much higher than the students who think that the lesson is "almost never ($\bar{x}$=442)" adapted to students, and the difference is significant. From the students' point of view, individual assistance to the students who had difficulties in the lesson also made a significant difference in the reading scores. Students who report to have been helped when they had difficulty in "almost every lesson" have higher reading performance than others. In terms of performance level, students who report to have been "almost never" helped or "sometimes" helped when they have difficulties are at the bottom. Karaman (2022) found out that the adaptation of instruction showed a positively significant relationship with reading literacy in Türkiye. Adapting the course requires expert knowledge. Vaughn (2019) found that teachers who made adaptation to the specific needs of their students could change their teaching according to the individual situation and the students they worked with. Houtveen *et al.* (1999) found that adapting the instruction during the initial reading process provided more successful reading results. Qian and Lau (2022) also found out that adaptive instruction was associated with reading performance.

Teachers' stimulation of reading engagement has been monitored since PISA 2009. According to the findings, the encouragement of teacher to express opinions and associating the content with their schemata in classroom reading practices and motivating student participation with questions are important parameters in reading performance. For example, the average reading score develops linearly when the teacher encourages the student to express his or her views on the text they have read in the lesson. The average reading score of the students who stated that they were not encouraged to express their opinions was significantly lower. Encouraging students to relate the text they have read to their own lives also affects reading comprehension performance. In addition, as the frequency of motivating the student's participation in the lesson with questions increases in the Turkish / Turkish Language and Literature course, the level of reading comprehension also increases. Based on the PISA 2018 findings, Qian and Lau (2022) showed that teacher encouragement was positively related to reading performance at both student and school levels. Guthrie *et al.* (2006) identified that stimulating tasks in reading increased interest, internal motivation, and reading comprehension. Studies show that teachers have a critical role in promoting motivation to read intrinsically (e.g., De Naeghel *et al.,* 2014; Gambrell, 1996; Guthrie, McRae, & Klauda, 2007). Verdegaal (2021) suggests that the decline in the Netherlands' PISA reading performance is related to reading motivation. Finally, there is a 43-point difference between the students who reported that they "never or almost never" wrote a short summary of the previous lesson at the beginning of the lesson and those who reported that this practice was done "every lesson." As the frequency of the teacher's reporting of summarizing the lesson decreased, the reading performance score increased.

The average reading score of students who reported that there were no *summarizing the text, comparing their own experiences with the content of the text, comparing the text with other texts on similar topics, and writing activities related the text* was much higher than the students who reported that these activities were carried out. The average score of the students who pointed out that they summarized the text they read in the course (72%) was 458, while the average score of those who reported that they did not summarize the text (27%) was 489. There is a 22-point difference between the average of students who reported that small group discussions "was done" (45%) and students who reported that, "it was not done" (53%), and the difference is significant. There was also a 29-point difference in the statements for comparing the text to their own experience (yes= 52%, no=46). Intertext comparison (yes= 53%, no=45) and text-related writing (yes= 49%, no=49) show a 26-point difference in favor of those who reported negatively. It can be thought that the expectations of the good readers may not be met by the teacher, and that the awareness of the poor readers about the classroom activities is weak. From another point of view, the quality of in-class reading activities can be discussed. In the literature, the opinion that summarizing, criticizing, and evaluating the text affects reading performance is dominant. Kutlu *et al.* (2011) pointed out that the probability of predicting whether the reading comprehension was successful or not was influenced by the variable that the teacher had them write a summary about the texts they read. Dilidüzgün (2013) identified that the frequency of teachers' summary studies was limited to the summary studies in the book (97%). In addition, 31% of the teachers argued that the ability to summarize was not taught, and 47% argued that it was partially taught. Erdağı-Toksun (2017) pointed out that 4 out of 15 teachers had their students write a summary during reading-comprehension activities. In the project conducted by Kutlu et al. (2019), there was an increase in teachers' initiatives and behaviors such as giving feedback to students about reading comprehension, encouraging for discussion, encouraging them to express their opinions, making them associate it with their own experiences, writing something about what they read and summarizing what they read.

The present study has identified that teacher-related variables play crucial roles in students' reading achievement. Reading comprehension performance is positively or negatively associated with teacher support, teacher's adaptive instruction, teacher feedback, engagement in reading activities and in-class teaching practices. In order to increase reading performance, it can be recommended to focus on the teacher's behavior in the classroom.

4.1. Limitations

The study does not include a comparison with the data of the countries in Türkiye's economic bracket; it has limitations in terms of not addressing the differences that may occur in terms of gender, school type, reading habits and socio-economic variables.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

B. Umit BOZKURT https://orcid.org/0000-0003-2532-9104

REFERENCES

Afflerbach, P., & Harrison, C. (2017). What is engagement, how is it different from motivation, and how can I promote it? *Journal of Adolescent & Adult Literacy, 61*(2), 217-220. https://doi.org/10.1002/jaal.679

De Naeghel, J., Valcke, M., De Meyer, I., Warlop, N., Van Braak, J., & Van Keer, H. (2014). The role of teacher behavior in adolescents' intrinsic reading motivation. R*eading and Writing, 27*(9), 1547-1565. https://doi.org/10.1007/s11145-014-9506-3

Dilidüzgün, Ş. (2013). Ortaokul Türkçe derslerinde oku(ma)dan özet yaz(ma)ya [From reading to summary writing in secondary school Turkish lessons]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 46*(2), 47-68. https://shorturl.at/bFO57

Dolapçıoğlu, S. (2019). Teacher support for a classroom setting that promotes thinking skills: an analysis on the level of academic achievement of middle school students. *Cukurova University Faculty of Education Journal, 48*(2), 1429-1454. https://doi.org/10.14812/cufej.557616

Erdağı-Toksun, S. (2017). Türkçe öğretmenlerinin okuma stratejileri bilişsel farkındalık becerilerini kullanma düzeylerine ilişkin görüşleri [The views of Turkish teachers on the level of using their cognitive awareness skills of reading strategies]. *e-Kafkas Journal of Educational Research, 4*(2), 10-18. https://doi.org/10.30900/kafkasegt.310416

Fraser, B.J. (1998). Classroom environment instruments: Development, validity and applications. *Learning Environments Research, 1,* 7-34. https://doi.org/10.1023/A:1009932514731

Federici, R.A., & Skaalvik, E.M. (2014). Students' perceptions of emotional and instrumental teacher support: Relations with motivational and emotional responses. *International Education Studies, 7*(1), 21-36. http://dx.doi.org/10.5539/ies.v7n1p21

Gambrell, L.B. (1996). Creating classroom cultures that foster reading motivation. *Reading Teacher, 50*, 14-25. https://www.jstor.org/stable/20201703?seq=1

Göçer, A., & Şentürk, R. (2019). Türkçe öğretmenlerinin metin işleme sürecinde kullandiklari geribildirim türlerine yönelik bir araştirma [A research on the feedback species of the Turkish teachers used in the process of text processing]. *Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 31*, 40-92. https://doi.org/10.14520/adyusbd.461313

Guess, P., & McCane-Bowling, S. (2016). Teacher support and life satisfaction: An investigation with urban, middle school students. *Education and Urban Society, 48*(1), 30-47. http://dx.doi.org/10.1177/0013124513514604

Guthrie, J.T., McRae, A., & Klauda, S.L. (2007). Contributions of concept-oriented reading instruction to knowledge about interventions for motivations in reading. *Educational Psychologist, 42,* 237–250. http://dx.doi.org/10.1080/00461520701621087

Guthrie J.T., Wigfield, A., Humenick, N.M., Perencevich, K.C., Taboada, A., & Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *The*

Journal of Educational Research, 99(4), 232-246. http://dx.doi.org/10.3200/JOER.99.4.232-246

Houtveen, A.A.M., Booij, N., de Jong, R., & van de Grift, W.J.C.M. (1999). Adaptive instruct ion and pupil achievement. *School Effectiveness and School Improvement, 10*(2), 172-192. http://dx.doi.org/10.1076/sesi.10.2.172.3508

Hughes, J.N., Luo, W., Kwok, O.M., & Loyd, L.K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psy chology, 100*(1), 1-14. http://dx.doi.org/10.1037/0022-0663.100.1.1

Karaman, P. (2022). Examining non-cognitive factors predicting reading achievement in Türk iye: Evidence from PISA 2018. *International Journal of Contemporary Educational Res earch, 9*(3), 450-459. https://doi.org/10.33200/ijcer.927884

Karip, E. (2020). *PISA'da okuma performansı ve öğrencilerin okul yaşamı [Reading perform ance and students' school life in PISA*]. TEDMEM. http://shorturl.at/elpCS

Klem, M.A., & Connell, J.P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74*(7), 262-273. http://dx.doi.org/10.1111/j.1746-1561.2004.tb08283.x

Kutlu, Ö., Yıldırım, Ö., Bilican, S., & Kumandaş, H. (2011). İlköğretim 5. sınıf öğrencilerinin okuduğunu anlamada başarılı olup-olmama durumlarının kestirilmesinde etkili olan değ işkenlerin incelenmesi [Investigation of factors that affect 5th graders' success in readin g comprehension]. *Journal of Measurement and Evaluation in Education and Psycholo gy, 2*(1), 132-139. https://dergipark.org.tr/tr/pub/epod/issue/5806/77235

Kutlu, Ö., Özyeter, N.T., Alpayar, Ç., & Kula-Kartal, S. (2019). *Okuduğunu anlama becerisinin ölçülmesi ve değerlendirilmesi [Measurement and assessment of reading comprehension skills]*. Ankara Üniversitesi Ölçme ve Değerlendirme Uygulama ve Araştırma Merkezi. Ankara Üniversitesi Basımevi. http://shorturl.at/orIQX

Lee, J. (2012). The effects of the teacher–student relationship and academic press on student engagement and academic performance. *International Journal of Educational Research, 53*, 330-340. http://dx.doi.org/10.1016/J.IJER.2012.04.006

Lee, Y., Jang, B.G., & Conradi Smith, K. (2021). A systematic review of reading engagement research: What do we mean, what do we know, and where do we need to go? *Reading Psychology, 42*(5), 540-576. http://dx.doi.org/10.1080/02702711.2021.1888359

Lei, H., Cui, Y., & Chiu, M.M. (2018). The relationship between teacher support and students' academic emotions: A meta-analysis. *Frontiers Psychology, 8*, 1-12. https://doi.org/10.3389/fpsyg.2017.02288

Lei, Y., Wen, Z., Li, J., Kong, Y., Chen, Q., & Li, S. (2019). Teacher support, reading strategy and reading literacy: A two-level mediation model. *Best Evid Chin Edu, 2*(1), 157-170. https://doi.org/10.15354/bece.19.ar1036

Ma, L., Luo, H., & Xiao, L. (2021). Perceived teacher support, self-concept, enjoyment and achievement in reading: A multilevel mediation model based on PISA 2018. *Learning and Individual Differences, 85*, 1-9. https://doi.org/10.1016/j.lindif.2020.101947

Ma, L., Xiao, L., & Hau, K.T. (2022). Teacher feedback, disciplinary climate, student self-concept, and reading achievement: A multilevel moderated mediation model. *Learning and Instruction, 79,* 1-12. https://doi.org/10.1016/j.learninstruc.2022.101602

Malecki, C.K., & Demaray, M.K. (2003). What type of support do they need? Investigating student adjustment as related to emotional, informational, appraisal, and instrumental support. *School Psychology Quarterly, 18*(3), 231-252. http://dx.doi.org/10.1521/scpq.18.3.231.22576

Ministry of National Education. (2022). *2022 ortaöğretim kurumlarına ilişkin merkezi sınav [2022 central examination for secondary education institutions].* The Series of Educatio n, Analysis, and Evaluation Reports, Ministry of National Education, Türkiye.

Ministry of National Education. (2021). *2021 ortaöğretim kurumlarına ilişkin merkezi sınav [2021 central examination for secondary education institutions].* The Series of Education, Analysis, and Evaluation Reports, 16. Ministry of National Education, Türkiye.

Ministry of National Education. (2020). *2020 ortaöğretim kurumlarına ilişkin merkezi sınav [2020 central examination for secondary education institutions]*. The Series of Education, Analysis, and Evaluation Reports, 12. Ministry of National Education, Türkiye.

Merga, M.K. (2020). Fallen through the cracks: Teachers' perceptions of barriers faced by struggling literacy learners in secondary school. *English in Education, 54*(4), 371-395. http://dx.doi.org/10.1080/04250494.2019.1672502

Meşe-Soytürk, M. (2020). *Yapısal eşitlik modelleri ve 2018 PISA verileri ile örnek bir uygulama [Structural equation models and a case study using 2018 PISA]* [Master's dissertation, Yildiz Technical University].

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results (volume I): Excellence and equity in education.* OECD Publishing.

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results (volume I): What students know and can do.* OECD Publishing. https://doi.org/10.1787/5f07c754-en

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results (volume III): What school life means for students' lives.* PISA OECD Publishing. https://doi.org/10.1787/acd78851-en

Qian, Q., & Lau, K.L. (2022). The effects of achievement goals and perceived reading instruction on Chinese student reading performance: Evidence from PISA 2018. *Journal of Research in Reading, 45*(1), 137-156. https://doi.org/10.1111/1467-9817.12388

Parlak, B. (prepared by) (2019). *Akademik becerilerin izlenmesi ve değerlendirilmesi (Monitoring and assessing of academic skills Project 2018 report for 4th grades).* Ministry of National Education General Directorate of Measurement, Assessment, and Examination Services, Türkiye.

Reyes, M.R., Brackett, M.A., Rivers, S.E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, *104*(3), 700-712. http://dx.doi.org/10.1037/a0027268

Ruzek, E.E.A., Hafen, C.A., Allen, J.P., Gregory, A., Mikami, A.Y., & Pianta, R.C. (2016). How teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction, 42*, 95-103. http://dx.doi.org/10.1016/J.LEARNINSTRUC.2016.01.004

Safari, N.F.N. (2020). Students' perception of teacher guidance on reading learning based on results of PISA 2018. *Indonesian Journal of Educational Assessment, 3*(1), 32-41. https://doi.org/10.26499/ijea.v3i1.56

Sakiz, G., Pape, S., & Hoy, A. (2012). Does perceived teacher affective support matter for middle school students in mathematics classrooms? *Journal of School Psychology, 50*(2), 235-255. http://dx.doi.org/10.1016/J.JSP.2011.10.005

Sukhram, D., & Monda-Amaya, L.E. (2017). The effects of oral repeated reading with and without corrective feedback on middle school struggling readers. *British Journal of Special Education, 44*(1), 95-111. http://dx.doi.org/95-111. 10.1111/1467-8578.12162

Vaughn, M. (2019). Adaptive teaching during reading ınstruction: A multi-case study. *Reading Psychology, 40*(1), 1-33. http://dx.doi.org/10.1080/02702711.2018.1481478

Verdegaal, A.L. (2021). *The Dutch decline in PISA reading performance explained: Exploring ICT-use, reading motivation, reading frequency, and reading strategies* [Master's dissertation, University of Twente]. https://purl.utwente.nl/essays/86741

Wang, M., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal, 47*(3), 633-662. http://dx.doi.org/10.3102/0002831209361209

Yıldırım, A., & Özgürlük, B. (prepared by) (2019). *Akademik becerilerin izlenmesi ve değerlendirilmesi (Monitoring and assessing of academic skills project-2018 report for 8th grades)*. Ministry of National Education General Directorate of Measurement, Assessment, and Examination Services, Türkiye.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 179–196

https://doi.org/10.21449/ijate.1127815

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Click here for the Turkish version of this article.

Examining the effect of peer and self-assessment practices on writing skills

Aslihan Erman Aslanoglu[1,*]

[1]Ufuk University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY

Received: June 08, 2022

Revised: Sep. 02, 2022

Accepted: Sep. 27, 2022

Keywords:

Peer-assessment,
Self-assessment,
Writing skills,
Quasi-experiment method.

Abstract: This study aims to reveal how peer- and self-assessment practices influence the writing skills of 9th grade students. The study adopted mixed-methods explanatory design. The participants were 102 students attending a public school in Ankara. The quantitative data were collected through a quasi-experimental method, and qualitative data were collected through a case study. There were three groups of participants in this study: the 1st experimental group in which peer-assessment was carried out with 34 participants; the 2nd experimental group in which self-assessment was conducted with 34 students, and 34 students in the control group. The interventions lasted 7 weeks. Writing performance tasks and rubrics were used to gather quantitative data while a Semi-Structured Interview Form was used to collect the qualitative data. For the analysis, paired samples *t*-test, ANOVA, and content analysis were used. The findings revealed that there was a significant difference between pre-test and post-test scores of experimental groups in which peer and self-assessments were conducted whereas there was not a significant difference between the scores of the control group. The findings of ANOVA, the post-test results of the experimental and control groups showed that there was a significant difference between all groups in favor of the 1st experimental group in which peer assessment was applied. The qualitative findings of the study corroborate the quantitative findings. Hence, we can conclude that peer and self-assessment practices were effective both in the development of students' writing skills and on their attitudes and interests towards writing.

1. INTRODUCTION

Writing skill, which is a language skill that students are required to gain and improve from their first year of educational life, is one of the most significant skills used while expressing oneself. It is deemed critically vital for students in terms of their academic success in other courses, expressing their thoughts effectively through writing and noting down what they have learnt (Sperling & Freedman, 2001). Since approximately half of the practices in the school environment require writing, the activities used to improve this skill become more important than any other skills.

*Corresponding Author: Aslihan ERMAN ASLANOGLU ✉ aslıhanerman@yahoo.com 🖃 Ufuk University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye.

e-ISSN: 2148-7456 /© **IJATE** 2022

Writing skill is considered a skill that encompasses steps including designing, organizing thoughts, drafting, formation, and editing (Chamot, 2009). Writing skill, with these aspects, is a higher order thinking skill, which is also simultaneously regarded as a process that incites metacognitive skills (Earl & Katz, 2006). One can describe higher order thinking skills as one's ability to use several skills holistically associating with their personal characteristics. Thinking method used by the students during writing constitutes the cognitive aspect of writing, and the checking technique used in the process of writing constitutes the metacognitive aspect of writing (Collins, 2000). Metacognitive skills can be defined as the level of awareness or knowledge that the individual has of their thinking or cognitive abilities (Desoete & Roeyers, 2002). Metacognitive skills are conceived as important factors to develop the concept of life-long and life-wide learning, and it is asserted that students with improved metacognitive skills will be more successful than others in their future lives (Edwards et al., 2002).

The assessment phase, which comprises the metacognitive aspect of writing skill, is one of the most valuable parts of a writing practice. Students can improve their own writing ability, fix their mistakes, and gain prevalent articulacy in writing through the feedback given as a result of the evaluation (Black et al., 2003). Information about the practices and the impact of these practices are limited in Turkey since there is no distinct writing approach to follow and assess writing skills in our country (Karatay, 2013). However, the development and improvement of students' writing skills necessitate the inclusion of processes such as planning, regulating at certain intervals, reviewing, correcting, and re-writing the teaching of writing (Collins, 2000). During these processes, when students receive feedback particularly on what they have written, they can be aware of the impact their writing has created on their readers and find the opportunity to improve themselves.

Teacher is mostly the primary evaluator in the assessment of students' written products. However, feedback should not be provided by a single source, but multiple and different sources are required. It is especially emphasized that diversification of sources that provide feedback is a necessity in order to have effective feedback practices (Ferris, 1997). These sources can be teachers, peers or even students themselves (Sun & Feng, 2009).

Peer- and self-assessments are metacognitive strategies helping students create recognition in what works and what they are supposed to improve regarding their performances. They ensure that the students make their mind in problem solving and decide for themselves regarding their attitude and attitudes of their peers. Once a teacher gives assignments for peer- and self-assessment, students will have the opportunity to reveal things and draw implications regarding their writing ability. They can improve their metacognitive skills by assessing not only their peers' but also their own work (Kulm, 1994). Additionally, thanks to these approaches, students will have the opportunity to criticize their learning and make it more permanent by taking the responsibility of their learning process (Sadler & Good, 2006). This condition, thus, creates a positive learning environment for students (Noonan & Duncan, 2005).

Peer assessment is defined as giving feedback to peers regarding a particular task, problem or performance on the basis of a standard set of criteria (Boud & Falchikov, 2007). Students already assess themselves and their peers in the educational environment. With the help of these assessments, they compare what they have learnt with that of the others and use it in order to make inferences about their own learning process. To include peer- and self-assessment to existing assessment and evaluation process allows students to systematize and formalize the assessments they have already made.

Researchers state that peer feedback has a significant role in students' educational life to improve their written products (Ruegg, 2015). Thanks to peer assessment, students not only get feedback from their peers and give feedback to them. With the help of this approach, students get the opportunity to compare their writing with those of the others and to widen and deepen their grip of writing process and language use. In return, their critical reading skills, as a reader, are improved, and general critical thinking skills are developed (Moussaoui, 2012).

Self-assessment is an evaluative process in which students critically make reflections their works' quality, comment on what extent their work reflects the explicitly stated aims, and review their writing performance accordingly. In other words, self-assessment can be explained as a skill to criticize and decide upon one's thoughts and skills as a way of reinforcing their learning skills (Noonan & Duncan, 2005). With this aspect, self-assessment enables students to become autonomous learners and to mirror their progress and criticize their work (Pierce, 2003).

Self-assessment in writing practices is considered a necessity rather than a preference (Lam, 2010). By means of self-assessments, students grasp the performance expected from them and improve their writing skills by determining their weaknesses and strengths about writing (Oscarson, 2009). If self-assessment activities are carried out effectively, student grading may help the teacher save time, and provide feedback in the shortest time (Boud 1989; Sadler & Good, 2006). Self-assessment gives students the chance to analyze their writing skills and make alterations accordingly (Boud, 1989; Mistar, 2011). Academic success of the students who find the opportunity to notice their shortcomings and work on them is positively affected (Desoete & Roeyers, 2002; Gardner, 2000).

While students fulfill performance tasks that require higher order thinking skills like writing, the rubrics are instructive in evaluating these tasks. Rubric is a kind of rating tool that shows the dimension of the quality to be assessed in the evaluation of students' performances, and it comprises assessment criteria, criteria definitions and a rating strategy (Popham, 2006). Rubrics help not only the teachers but also the students capture the criteria to be deployed to assess a work and realize the level of the present performance of the students (Kutlu et al., 2010).

Studies on classroom assessment have demonstrated that peer- and self-assessment based upon a rubric improve students' writing performance and enhances the reliability of the grades by providing concrete criteria for performance evaluation (Andrade et al., 2008; Ross et al.,1999; Weigle, 2002)

The assessment phase that constitutes the metacognitive aspect of writing skill which has critical importance for students is one of the most important parts of an effective writing practice. Even though peer- and self-assessment are recommended to be used from primary school to higher education in evaluating writing skills, researchers indicate that there are restricted number of experimental studies in the international literature on this matter (Nielsen, 2021; Ruegg, 2015; Strijbos & Sluijsmans 2010). First group studies addressing self and peer assessment and writing skill are mainly based upon the comparison of the rating of teachers, peers and the students themselves in order to make evaluations about the reliability of peer- and self-assessment scores (Cho et al., 2006; Eckes, 2008; Falchikov & Goldfinch, 2000; Topping, 2003). These studies depend upon the hypothesis that if there is resemblance between teacher's scores and the feedback given to oneself or peers, then it is reliable. Second group studies involve the teachers' and students' opinions of peer- and self-assessment practices (Brown et al., 2009; Cheng & Warren, 1997; Fallows & Chandramohan, 2001; Hanrahan & Isaacs, 2001; Young & Jackman 2014). Third group studies focus on how the use of rubric influences students' peer- and self-assessment practices during the evaluation of students' writing performance (Andrade et al., 2008; Ross et al., 1999; Weigle, 2002). Studies in the international literature regarding peer- and self-assessment in writing skill is predominantly centered around writing skills in teaching English as a second/foreign language (Javaherbashsh, 2010; Meihami & Varmaghani, 2013, Nielsen, 2021, Wang et al., 2017). Similarly, studies in the national literature regarding peer- and self-assessment are associated with writing skills in foreign language teaching (Cömert & Kutlu 2018; Uysal, 2008). Additionally, the reliability of peer, self and teacher rating in the assessment of writing skills have also been addressed (Erman Aslanoğlu et al., 2021).

Previous literature shows that there is a necessity to conduct studies with regard to the influence of feedback based upon peer- and self-assessment on writing skills in mother tongue and to observe the influence of the process of peer- and self-assessment on writing skills following its application in the classroom environment. Therefore, this study attempts to illuminate the influence of peer- and self-assessment practices on the writing skills of high school freshmen year students. In this respect, the present study seeks answers to the questions given below:

1. Is there a significant difference between the pre-test and post-test writing task scores of the students in the experiment group in which peer-assessment has been implemented?
2. Is there a significant difference between the pre-test and post-test writing task scores of the students in the experiment group in which self-assessment has been implemented?
3. Is there a significant difference between the pre-test and post-test writing task scores of the students in the control group in which peer- and self-assessment methods have not been implemented?
4. Is there a significant difference between the pre-test and post-test writing task scores of the students in the self-assessment, peer-assessment and control groups?
5. What are the opinions of the students regarding the effect of peer-assessment practices on writing skills?
6. What are the opinions of the students regarding the effect of self-assessment practices on writing skills?

2. METHOD

2.1. Research Model

This research adopted mixed methods design in which quantitative and qualitative research techniques are jointly used. Mixed methods, the joint use of qualitative and quantitative methods, serve to carry out a thorough analysis and interpretation of the research problem (Yıldırım & Şimşek, 2011). This study implemented the "Exploratory Research Design" of mixed method designs. Accordingly, quantitative data of the study was analyzed first, then qualitative data were obtained and analyzed. The findings obtained were interpreted in correlation to one another.

As a quantitative dimension of the study, quasi-experimental design was used. Out of the quasi-experimental groups, pretest-posttest matched control-group approach was chosen for the study, and among the groups that showed similar qualities as a result of the analyses conducted, one control group and two experimental groups were objectively appointed. Quasi-experimental design studies with pre-test and post-test groups require the objective selection of the groups. The researcher objectively chose a control and an experimental group out of the existing groups and applied the pretest to both groups. Within this context, following the experimental activities carried out in the experiment group, posttest were administered in both groups and the differences between them were evaluated (Creswell, 2005).

The second phase of the research was based upon the interviews conducted with the students. Case study was chosen for the analysis of qualitative data. Case study is a qualitative research method in which a case or cases, namely a program, a social group or systems that are linked to one another are thoroughly investigated, and themes dependent on these cases are defined (Merriam, 2015).

2.2. Study Group

The study group comprises 102 students attending the 9th grade in a state high school in Ankara. Prior to determining the experiment and control groups, the students' average grade point in the

Turkish Language course in the previous term was taken into consideration. General Turkish language course average grade point of 9th grade students of 6 groups was calculated to be 72.01 on the scale of 100.

One-way ANOVA test was employed in the analysis of the data since variance homogeneity could be met in the class divisions identified (Levene test F=.68, p>.05), score distributions were normal, and there were more than two groups. ANOVA analysis detected that the average grade point of the Turkish language course of the class divisions did not show a significant difference [F(5-226)=.28; p>.05]. This result demonstrates that there is no significant difference among the 6 class divisions regarding Turkish language grade point mean scores. Following these results, three of the class divisions were randomly selected as the study group. Moreover, prior to the experimental procedures carried out in the experimental groups, ANOVA test was used again to detect if there was a significant difference between pre-test scores of the study groups related to the writing skills. Table 1 illustrates the result of the ANOVA test conducted.

Table 1. *ANOVA results regarding the comparison of the pretest scores for writing skill.*

Group	N	$\bar{X}$	S_x	sd	F	p
1st Experimental (Peer) group	34	12.56	5.06			
2nd Experimental (Self) group	34	12.79	4.33	2-99	0.023	.98
Control group	34	12.65	4.24			

When Table 1 was reviewed, a significant difference was not detected between the groups regarding the mean scores for writing skills [F(2-99)=.023; p>.05]. As a result of the analyses performed, one control group two and experimental groups were randomized out of the three groups. In this study, among 102 students, there were 34 students in the First Experimental Group (Peer Assessment), 34 in the Second Experimental Group (Self Assessment) and 34 in the Control Group. Table 2 summarizes gender distribution of the students attending the control and experimental groups.

Table 2. *Distribution of the students to experimental and control groups by gender.*

Grup	Gender	N	Toplam
1st Experimental group	Female	16	34
	Male	18	
2nd Experimental group	Female	15	34
	Male	19	
Control group	Female	18	34
	Male	16	

Table 2 indicates that 47.1% of the students in 1st experimental group were female, and 52.9% of it were males. In the 2nd experimental group, females comprised the 44.1% of the group while males formed the 55.9% it. In control group, females formed the 52.9% while males comprised 47.1% of the group.

2.3. Procedures

Writing skill pre-test was primarily administered to all the groups within the scope of the research. Having completed the writing skill pre-test, writing skills of each group were rated by two raters, and their mean scores were used as the pre-test scores of the students. Following the application of the pre-test, the learning and teaching process in the study was conducted differently in the experimental groups where peer- and self-assessment were conducted, and in the control group where normal education was continued. The intervention phase of the research took 7 weeks (21 hours in total). The following section presents the practices applied in the experimental and control groups during this process.

2.3.1. *Procedure steps in the first experimental group (peer-assessment) and second experimental group (self-assessment)*

Fachikov (2005) recommends an effective guide oriented at carrying out writing skills practices with peer- and self-assessment approaches in the classroom environment. Peer- and self-assessment studies in this study were performed based upon these steps. The steps and the practices carried out are as follows:

1. **Informing the students on peer- and self-assessment practices:** The students had no prior knowledge of peer- and self-assessment practices. Within the scope of this step, the students in the 1st Experimental Group were informed on what peer- assessment was, how it was made, and the benefits of peer-assessment in the first week. The students attending to the 2nd Experimental Group were informed on the self-assessment approach.

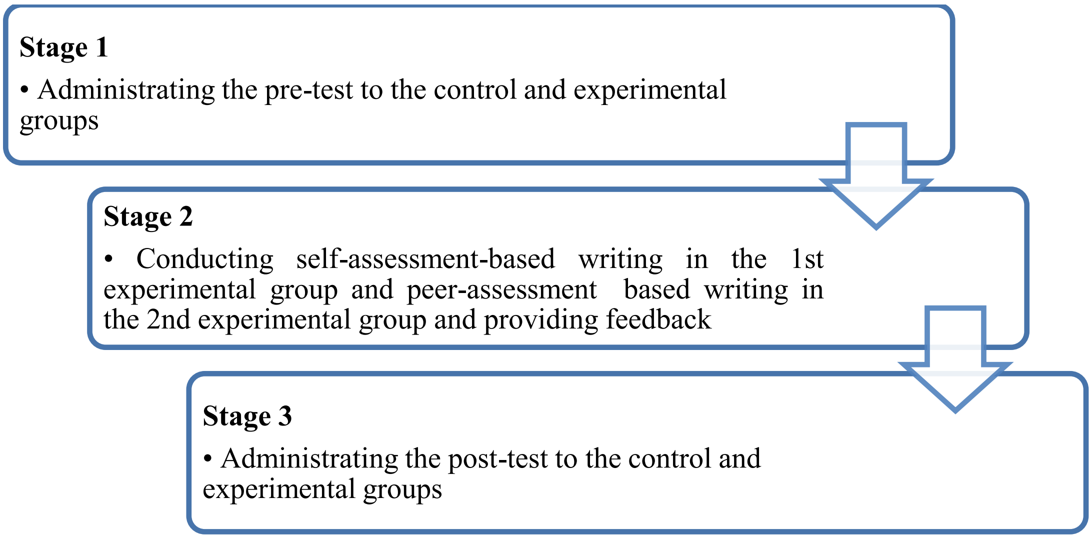
2. **Explaining students that participating in peer- and self-assessment is beneficial and providing evidence:** Within the scope of this step, the students were enlightened about what feedback was and that feedback could be provided from different sources (teacher, peer, self) and examples on how peer- and self-assessment could be made were introduced to the 1st and 2nd Experimental Groups in the second week. How the students would be involved in the assessment was also explained at this phase.

3. **Explaining the assessment criteria to students:** Writing assessment rubric was introduced to peer- and self-assessment groups, and information was provided on the criteria and criteria definitions found in the rubric.

4. **Conducting sample studies:** It is important to carry out studies as examples so that students can gain practicality and see their shortcomings in peer and self-assessment practices. Within this scope, the 1st Experimental Group (peer assessment) and the 2nd Experimental Group (self-assessment) were asked to write two more narratives during the process. Students attending to the 1st Experimental Group were randomly divided into groups of 3 or 4. The written product of each student in the group was assessed by two friends in the group, and feedback was given. When peer feedback had been completed, the teacher laid specific examples that carried perfect, average and weak qualities on the table and provided feedback on these matters. The students in the 2nd Experimental Group assessed their own written products. The teacher laid specific examples that carried perfect, mediocre and weak qualities on the table and provided feedback on these matters. During this process, the attention of the students was drawn to the mistakes they had made so that they could gain and improve their auto-control skill.

Following the completion of the above-mentioned processes in the peer- and self-assessment groups, the last test in which they were required to write a narrative was administered. Writing skills of the groups were rated by two raters, and the mean scores were used as post-test scores of both groups. Afterwards, interviews using a semi-structured form were administered to 15 students from varying levels of writing skills. A flowchart including the three-stage experimental process is presented in Figure 1.

Figure 1. *Flowchart illustrating the stages of the study.*



Stage 1

• Administrating the pre-test to the control and experimental groups

Stage 2

• Conducting self-assessment-based writing in the 1st experimental group and peer-assessment based writing in the 2nd experimental group and providing feedback

Stage 3

• Administrating the post-test to the control and experimental groups

2.3.2. *Control group*

Writing practices of the control group were implemented with regard to the curriculum of the relevant course. The teacher was asked to use a rubric in assessing students' writing tasks, and the essays of the students were evaluated accordingly using a rubric, and feedback was provided to the students as such.

2.4. Data Collection Tools

This section provides information about the data collection tools used during the research.

2.4.1. *Writing performance*

Four writing performance tasks were prepared to be used during peer- and self-assessment activities and to assess students' writing ability. Writing performance tasks were based on writing narratives. It is known that students mainly deal with narratives as text types in schools (Ateş, 2011). Equality in difficulty and class-level appropriacy of the writing performance tasks were considered. Two of the writing performance tasks that had equal difficulty levels were used in the pre-test and post-test practices of the experimental and control groups. The other two equally difficult writing performance tasks were used during the process for the writing practices of the control and experimental groups. Opinions were sought from two experts of the field, three Turkish literature teachers and two measurement and evaluation experts regarding the writing performance tasks prepared, and the tasks were put into their final form according to the received feedback.

2.4.2. *Rubric evaluating narrative writing*

A rubric was prepared following the steps recommended by Andrade (2001) so that the students' writing skills could be assessed and evaluated by peers, teacher, and themselves. The following are the steps and their explanations:

1) Identifying the criteria to be utilized in the assessment of writing skills: Since the students were going to be asked to write narratives, literature of the subject was reviewed, and 6 criteria were determined that provide the opportunity to assess students' writing skills as content-wise and format.

a) Textual Structure: Text should contain exposition, complication and resolution parts, and transition between the parts should be logically employed.

b) Characters: The name and physical-mental qualities of the characters should be given.

c) Setting and Time: The setting and time of the incident should be given in detail.

d) Chain of Events: The text should contain a chain of events, and transition from an event to another should be logical.

e) Language and Narration: Rich vocabulary should be used, statements should be clear and easy-to-comprehend, and meaningful connections between the statements should be sought.

f) Spelling and Punctuation Rules: Spelling and punctuation rules should be sought, words should be spelled correctly, and appropriate punctuation marks should be used.

2) Determining the rubric type: In evaluating a written product, different rubrics including holistic and analytic ones can be used. Analytic rubrics provide better results compared to holistic rubrics since they give more detailed feedback in assessing students' performance and ensure intra-rater and inter-rater reliability (Knoch, 2009). Due to these qualities, analytic rubric was used in this study.

3) Defining the criteria: Considering the level and age of the students, the criteria determined in order to assess the writing ability of the students were ranked between 1 and 4; 1 is the lowest and 4 is the highest. Detailed definitions were also written considering the criteria and ranking. Consequently, the rubric that was developed consisted of 6 criteria, and each criterion is scored from 1 to 4. One can get 24 points at most from this rubric.

4) Expert opinion: The rubric prepared was sent to 3 experts in the field, 2 Turkish Literature teachers and 3 measurement and evaluation experts, and the experts were asked to evaluate the rubric as "adequate, partly adequate and inadequate" in terms of content validity (content, structure, criteria), appropriateness to the level of the class, and spelling and narration mistakes. The rubric was organized again compatible with the recommendations of the experts.

Receiving expert opinion is of vital importance in terms of evaluating the validity of the analytic rubric developed. Rubric development steps were followed to ensure validity, and using formula recommended by Miles and Huberman (1994), compatibility percentages of the expert opinions was found to range between 89% and 97%. These compatibility percentages were evaluated as evidence of the content validity of the rubric prepared.

For satisfying the reliability of the scores obtained from the rubrics, inter-rater coherency was investigated. To that end, writing performance tasks of the students were scored by two teachers, and inter-rater coherency of the total scores the students received from the test was analyzed through Kendall's W test. Kendall's W coefficient receives values between 0 and 1. If the value calculated is closer to 0, it indicates an inter-rater incoherency, and if the value is closer to 1, it indicates an inter-rater coherency (Howell, 2002). As a result of the calculations, inter-rater coherency for the pretest and posttest was found as 0.87 and 0.89, respectively.

Furthermore, intra-rater agreement coefficient was also calculated to ascertain if there was a difference between the rating made by the same rater at different time frames. To that end, responses belonging to randomly selected student were re-scored by a randomly selected rater at three-weeks intervals. The result was found as 0.92 applying the formula recommended by Miles and Huberman (1994).

2.4.3. *Interview*

An interview form was utilized in the research to unearth students' opinions of the influence of peer- and self-assessment practices on their writing skills. Within this scope, a semi-structured interview form with two items was prepared. The items were sought to be easily understood by the students, fit the purpose of the interview and not to contain any controlling expressions. Opinions of two expert linguists were asked to evaluate the quality of the items. Amendment was made compatible with the recommendations, and the form was completed. Interviews were performed with the students at the end of the data-collection process. When an open response could not be received from the students, the questions were paraphrased in a different way considering the level and age of the participants.

2.5. Data Analysis

The pre-test and post-test scores of the students had normal distribution. Two factors of normality are skewness and kurtosis. Having a skewness coefficient within the limits of ±1 can be interpreted as the fact that scores do not show any important deviance (Tabachnick & Fidell, 2013). In this context, the pre-test and post-test scores were found to be within the limits and meet normality hypothesis. Therefore, statistical approaches were used in the analysis of data. In data analysis, t-test was used for dependent groups in the comparison of pre-tests and post-test scores since pre-test and post-test scores showed normal distribution, variances were homogenous, and covariance matrixes were equal. In inter-group comparisons, one-way ANOVA was used. Since a significant difference was detected between the groups after ANOVA analysis, Scheffe's test was used based on variance homogeneity. Statistical significance was set at 0.05 in all analyses conducted in the research. Moreover, in the event of a significant difference between the groups, effect size was calculated to determine how significant this difference was between the variables. While determining effect size, eta-squared ($\eta 2$) was used for the dependent group t test that analyzed the difference between the average of the two groups, and Cohen's f value was calculated in variance analysis (Creswell, 2005). $0.01 \leq \eta 2 < 0.06$ eta-squared value is interpreted as small effect, $0.06 \leq \eta 2 < 0.14$ range is considered as moderate effect, and values ranging between $0.14 \leq \eta 2$ show large effect. Cohen's f value belonging to the data was interpreted as small at .10, moderate at .25 and large at .40 (Cohen, 1988).

Content analysis was used to analyze qualitative data. The most general definition of content analysis is a systematic coding of qualitative or quantitative data within a specific theme or classifications (Creswell, 2005). In content analysis, the main aim is to reach notions that could explicate the collected data, thus similar data are brought together and interpreted in relation with the notions and themes determined (Yıldırım & Şimşek, 2010).

3. FINDINGS

The findings of the analyses are given in this section.

3.1. Findings Related to the Pretest and Posttest Score of the 1st Experimental Group (Peer Assessment)

Following the experimental procedures carried out in the 1st Experimental Group within the scope of the question: "Is there a significant difference between the pre-test and post-test writing task scores of the students in the experimental group in which peer-assessment method has been implemented?" paired samples t-test was used to unearth if there was a significant difference between pretest and posttest scores belonging to writing skills, and the results were illustrated in Table 3.

Table 3. *Paired Samples t-test results regarding pretest and posttest scores of the 1st experimental group.*

Grup	Test	*N*	$\bar{X}$	S_x	*sd*	*t*	*p*
1st Experimental Group	Pretest	34	12.56	5.06	33	-12.058	0.000*
	Posttest	34	19.53	3.82			

* $p<0.05$

As it is illustrated in Table 3, a significant difference was found between the pretest and posttest scores of writing skills of the 1st Experimental Group [$t(33)= -12.058$, $p< .05$]. According to the findings obtained, it was found out that the mean score of the posttest scores ($\bar{X}$=19.53) of the 1st Experimental Group was significantly higher than the pretest scores ($\bar{X}$=12.56). These findings indicate that peer assessment has a positive effect on the improvement of writing skills of the students. Eta-squared effect size was found as *η2*= 0.815. This value is an evidence that peer-assessment has a "large effect" on the enhancement of the students' writing ability.

3.2. Findings related to the Pretest and Posttest Score of the 2nd Experimental Group (Self-Assessment)

Following the experimental procedures carried out in the 2nd Experiment Group within the scope of the question: "Is there a significant difference between the pre-test and post-test writing task scores of the students in the experimental group in which self-assessment method has been implemented?," paired samples t-test was used to unearth if there was a difference between pre-test and post-test scores related to writing skills, and the results were illustrated in Table 4.

Table 4. *Paired Samples t-test results regarding pretest and posttest of the 2nd experimental group.*

Group	Test	*N*	$\bar{X}$	S_x	*sd*	*t*	*p*
2nd Experimental Group	Pretest	34	12.79	4.33	33	-7.983	0.00*
	Posttest	34	16.09	4.13			

* p<0.05

Table 4 shows that a significant difference was found between the mean scores of the pre-test and post-test scores of writing skills of the 2nd Experimental Group [$t(33)$=-7.983, p< .05]. Accordingly, it was seen after experimental practices that the mean score of the post-tests ($\bar{X}$=12.79) of the 2nd Experimental Group was significantly higher than the pre-test scores ($\bar{X}$=16.09). These findings suggest that self-assessment has a positive effect on the improvement of writing skills of the students. Eta-squared effect size was found as *η2*= 0.658. This value is an evidence that peer-assessment has a "large effect" on the improvement of the students' writing ability.

3.3. Findings Related to the Pretest and Posttest Score of the Control Group

Following the educational procedures carried out in the Control Group (no peer and self-assessment) within the scope of the question "Is there a significant difference between the pre-test and post-test writing task scores of the students in the control group in which peer- and self-assessment methods have not been implemented?," paired samples t-test was used to compare and find out the pre-test and post-test scores related to writing skills of the students, and the findings were illustrated in Table 5.

Table 5. *Paired Samples t-test results regarding pretest and posttest scores of writing skills of the control group.*

Group	Test	*N*	$\bar{X}$	S_x	*sd*	*t*	*p*
Control Group	Pretest	34	12.64	4.24	33	-1.496	0.144
	PostTest	34	13.18	4.21			

As can be seen in Table 5, there was no significant difference between the mean scores of pre-test and post-test of writing skills of the Control Group [$t(33)$=-1.496, p> .05]. According to this finding, it can be inferred that the current education process carried out in the control group has no significant effect on writing skills.

3.4. Findings Related to the Posttest Score of the Experimental and Control Groups

In order to answer the question "Is there a significant difference between the pre-test and post-test writing task scores of the students in the self-assessment, peer-assessment and control groups?" one-way ANOVA was carried out to illuminate if there was a difference between the posttest scores of the students belonging to the control and experimental groups. The findings were illustrated in Table 6.

Table 6. *Results of the ANOVA of the posttest scores in the experimental and control groups.*

Group	*N*	$\bar{X}$	S_x	*sd*	*F*	*p*
1st Experimental Group (Peer)	34	19.53	3.82	2-99	20.857	0.00
2nd Experimental Group (Self)	33	16.09	4.13			
Control Group	34	13.18	4.22			

As seen in Table 6, writing skill post-test scores of the groups significantly differed between the groups [*F*(2-99)=20.857. Since group variances were homogenous, Scheffe's test was used. The results suggest that writing skills of the students who attended to the 1st Experimental Group (peer assessment) were significantly higher than those of the 2nd Experimental Group (peer assessment) (p < .05) and Control Group (p< .05). Moreover, the writing skills of the students in the 2nd Experimental Group (self-assessment) were found to be higher than those of the Control Group (p< .05). When Cohen's *f* effect size value (Cohen's *f* =.30) of the difference between groups is investigated, it is found out that the difference has a "large effect" size.

3.5. Findings Related to Student Opinions Regarding the Effect of Peer Assessment on Writing Skills

Responses to two questions found in the interview form to answer the question "What are the opinions of the students regarding the effect of peer-assessment practices on writing skills?" were analyzed using content analysis. The 1st item of the interview form was the question "Do you think that peer assessment practices carried out to improve your writing skills have contributed to improve your writing skills? Please explain.", and findings related to the responses are presented below:

Students' opinions regarding the contribution of peer-assessment to writing skills were reviewed, and it was found out that these opinions could be brought together under two dimensions: cognitive and affective. These findings are illustrated in Table 7.

Table 7. *Opinions of the students in the 1st experimental group regarding writing processes.*

Category	Code	Frequency
Cognitive	Feedback given provided for realizing shortcomings and correcting mistakes	12
	Identifying the shortcomings of one's own work while assessing the work of others	7
	Receiving quick feedback	6
Affective	Positive emotions (Enjoying the process, finding it enjoyable, being happy, having a fruitful time)	12
	Decreased anxiety towards writing	6
	Increased motivation for writing	10

12 of the students stated that feedback given by peers during writing practices provided for realizing the shortcomings of their writings and contributed to their correction. 7 of the students indicated that they also identified their shortcomings while assessing the writings of others. 6 of the students remarked that quick feedback contributed to their studies. Regarding to the affective characteristics, 12 of the students found it positive to receive peer feedback while 6 students realized a decrease in the anxiety they had towards writing practices. 10 students specified that their motivation to write increased. Opinions of some students regarding this subject are as follows:

"…. *While assessing the work of my friends, I realized the shortcomings in my own work and could correct them.*" (Student A)

"…. *Talking to my friends and getting help from a rubric in essay writing decreased my anxiety. When the teacher told us to write an essay in the past, I would feel anxious about where and how to start.*" (Student B)

"…. *In the past, I could not decide on what to write and just wrote down a few sentences. Now, I started to write longer and more carefully since my friends would be the ones to assess me*." (Student C)

"… *I became aware of my shortcomings thanks to the feedback I received from my friends. It helped me focus on these points in my future writing practices*." (Student D)

"… *The practices were fun. Normally, I would only learn my grade after having written an essay but now I could quickly see my mistakes*." (Student E)

The 2nd item on the interview form was "What were the things that gave you a difficult time in making peer assessment? Please explain." The findings obtained herein indicated that the students had a difficulty in assessment, rubric use, and writing skills. The opinions are illustrated in Table 8.

Table 8. *Opinions of the 1st experimental group regarding the situations they had most difficulty in during peer assessment.*

Kategori	Kod	Frekans
Assessment	Disliking being assessed by a friend	2
	Feeling insufficient in assessing a friend	1
Rubric	Finding it hard to use a rubric since it was the first time	1
Writing skill	Having problems with writing	2

Among participants, 2 of the students expressed that they did not like being assessed by friends, one student felt insufficient while assessing friends, 1 student found it difficult to use a rubric, and 2 students had problems with writing. Opinions of some students regarding this subject are as follows:

"….. *my friend criticized my essay a lot, which made me feel insufficient*." (Student A)

"….. *I found it difficult to use this tool since it was the first time I used it*." (Student B)

"…. It is very difficult for me to write, but assessing the work of others was fun. I had difficulty because I do not like writing." (Student C)

3.6. Findings Related to Students' Opinions Regarding the Effect of Self-Assessment on Writing Skills

Content analysis was performed to analyze the responses to two questions found on the interview form designed to reveal answers to the question "What are the opinions of the students regarding the effect of self-assessment practices on writing skills?" The 1st item of the interview was the question "Do you think that self-assessment practices carried out to improve your writing skills have contributed to improving your writing skills? Please explain," and findings related to the answers are illustrated below.

Students' opinions regarding the contribution of self-assessment to writing skills were reviewed, and it was revealed that these opinions could be brought together under two dimensions: cognitive and affective. The summary of the findings is illustrated in Table 9.

Table 9. *Opinions of the students in the 2nd experimental group regarding writing processes.*

Category	Code	Frequency
Cognitive	Quality of the essays written increased	10
	Identifying the shortcomings of one's own work	12
	Leading to contemplating more on one's own work	11
Affective	Positive emotions (Enjoying the process, finding it amusing, being happy, having a fruitful time)	11
	Decreased anxiety towards writing	4
	Increased motivation for writing	9

Among the participants, 10 of them stated that self-assessments during writing practices increased the quality of their writings. 12 of the students indicated that they also identified their shortcomings, and 11 students specified that thanks to self-assessment, they could contemplate more on their own work. Regarding to the affective characteristics, 11 of the students found it positive to assess themselves during writing practices while 4 students realized a decrease in the anxiety that they had towards writing practices. 9 students specified that their motivation to write increased. Opinions of some students regarding this subject are as follows:

"…*In the past, I used to complete my writing and not contemplate on what I had written. I did not know what to pay attention to. Contemplating on what I had written increased the quality of my writings*." (Student A)

"… *I was happy to find the opportunity to contemplate on my work. With more practices, I started making less mistakes in my writing*." (Student B)

"… *It contributed a lot. My motivation increased. In my opinion, if students practice more like this, the quality of our wok will increase because once we are done with something, we usually do not have the chance to contemplate on it*." (Student C)

The 2nd item on the interview form was "What were the things that gave you a difficult time in making self-assessment? Please explain." The findings indicated that the students had difficulty in assessment, rubric use, and writing skills. The opinions are illustrated in Table 10.

Table 10. *Opinions of the 2nd experimental group regarding the situations they had most difficulty in during self-assessment.*

Category	Code	Frequency
Assessment	Feeling insufficient in assessing oneself	2
Rubric	Finding it hard to use a rubric since it was the first time	1
Writing Skill	Failure in self-assessment due to having problems with writing	2

Among the participants, 2 of them expressed that they felt insufficient for self-assessment, 1 student found it difficult to use a rubric, and 2 students had problems with self-assessment due to not their dislike towards writing. Opinions of some students regarding this subject are as follows:

"… *I felt insufficient in assessing my own work. I was anxious about if I was assessing myself correctly*" (Student A)

"…. *There was detailed information on how to use a rubric but it took some time to get used to it*" (Student B)

"…. *I cannot write long because I do not like writing. So, there is not much to assess*" (Student C)

4. DISCUSSION and CONCLUSION

Prior to the writing practices on the grounds of peer- and self-assessment activities, findings obtained from the pre-test application of the students in the experimental and control groups suggested that students' writing skill was not adequate. This finding validates the findings of other studies related to writing skills in the literature (Çağımlar & Oğlazoğlu, 2002). This present state implies that sufficient importance is not given to the improvement of this skill in our country.

Another finding of the present study is that there is a significant difference between the pre-test and post-test scores of the experimental groups. As for the control group, there is not a significant difference between pre- and post-test scores. Almost all of the studies investigating the effect of peer- and self-assessment on writing skills show that peer- and self-assessment have a positive effect on writing skills in general (Andrade & Boulay, 2003; Andrade et al., 2010; Cömert & Kutlu, 2018; Javaherbashsh, 2010; Meihami & Varmaghani, 2013).

In the final part of the study, a significant difference was found in all groups when a comparison was made between the mean score of the post-test scores belonging to the control and experimental groups. Post-test scores of the 1st Experimental Group (peer assessment) were detected to be significantly higher than those of the 2nd Experimental Group (self-assessment) and Control Group (teacher assessment). This finding coincides with the findings of experimental studies demonstrating a more positive effect of peer-assessment in writing in the mother tongue when compared to traditional feedback techniques (Cho & Schunn, 2007; Richer, 1992; Topping, 2003). For instance, in a study by Richer (1992) conducted on university students, the researcher has investigated the influence of peer and teacher assessment on writing skill and found that the writing skill of the students receiving peer feedback is significantly better than that of the students receiving teacher feedback. Additionally, Cho & Schunn (2007) have revealed that students receiving feedback from six peers were more successful than the students getting teacher feedback in improving the writing practices they have carried out for the Scientific Research Methods course. In the present study, post-test scores of the 2nd Experimental Group (self-assessment) were detected to be significantly higher than those of the Control Group (teacher assessment). This finding is parallel with the finding of other experimental studies in which self-assessment approach has been compared with teacher assessment (Andrade & Boulay, 2003; Andrade et al., 2010). For instance, according to the findings of the study by Andrade et al., (2010) which investigates the effect of self and teacher assessment on the writing skills of junior year high school students. Feedback based on self-assessment using a rubric has been found to have a more positive influence on the improvement in writing skills when compared to teacher assessment. The present study also found out that writing skill post-test scores of the 1st Experimental Group in which peer-assessment was used were significantly higher than those of the 2nd Experimental Group in which self-assessment was used .In the literature, experimental studies questioning which assessment is more effective in enhancing writing skills in the mother tongue could not be found; however, there are studies reporting that peer feedback is more effective in English as a second language teaching compared to self-assessment (Conrad & Goldstein, 2009; Khonbi & Sadeghi, 2012; Nakanoshi, 2015). Peer-assessment approaches are relatively more common when compared to self-assessment approaches (Fallows & Chandramohan, 2001). This context may even have helped students gain more advantage from peer-assessment approaches in writing practices.

Qualitative data of the study supports the findings obtained from quantitative analysis. Findings related to qualitative data obtained from the students in the experimental groups assessing writing skills with peer and self-assessment approaches suggest that peer- and self-assessment approaches contributed to the cognitive-affective characteristics of the students, enabled them to see their shortcomings, and gave them the opportunity to contemplate on their own work.

The findings in this report are subject to at least three limitations. First, the study was limited to freshmen year high school students. Second, different teachers conducted the writing processes in the experimental and control groups. As in any educational study, teacher differences may affect the presentation of the method and the results. Therefore, teachers should be supported, and care should be given to construct the same educational practices in each classroom. However, anyone teacher cannot be the same, and differences between the teachers may affect educational outcomes. Research with more than one teacher is affected by this limitation. On the other hand, two experienced raters obtained the students' pre-test and post-test scores using a rubric, and reliability of the scores was satisfied through this way; however, it should be kept in mind that subjective judgements during the rating process have limitations for the reliability of the scores.

Some future recommendations can be made regarding the results of the present study. When writing skill is considered as a critically important skill for students, feedback depending on peer- and self-assessment can be provided as of primary school within the scope of writing lessons. Thus, while writing skills of the students improve, so do their interest and motivation. Using metacognitive skills including peer- and self-assessment in writing practices can also improve these skills in students and help them be aware of their writing skills and processes. Therefore, teachers can receive vocational training on how to perform peer- and self-assessment activities in classroom. Researchers can investigate how peer- and self-assessment influence students' writing skills in mother tongue in different levels of grades. Moreover, it is also important to investigate in what way students with different proficiency levels in terms of writing are affected by peer- and self-assessment practices.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number**: Ufuk University/ Social and Human Sciences Scientific Research and Publication Ethics Committee, 2020-86.

Orcid

Aslihan Erman Aslanoglu https://orcid.org/0000-0002-1364-7386

REFERENCES

Andrade, H.G. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(4), 2-22. https://l24.im/BPRSj

Andrade, H.L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of modal, cr iteria generation and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice, 27*(2), 3-13. https://doi.org/10.1111/j.1745 3992.2008.00118.x

Ateş, S. (2011). *Evaluation of fifth-grade turkish course learning and teaching process in terms of comprehension instruction* [Doctoral dissertation, Gazi University]. Yöktez. https://l24.im/6AY

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2003). *Assessment for learning.* Open University Press. http://www.mcgraw-hill.co.uk/html/0335212972.html

Boud, D. (1989). The role of self-assessment in students' grading. *Assessment in Higher Education, 14*(1), 20-30. https://doi.org/10.1080/0260293890140103

Boud, D., & Falchikov, N. (2007) *Rethinking assessment in higher education: Learning for the long term.* Routledge. https://doi.org/10.4324/9780203964309

Chamot, A. U. (2009). *The CALLA handbook: Implementing the cognitive academic language learning approach* (2nd ed.). Pearson. https://l24.im/81v

Cheng, W., & Warren, M. (1997). Having second thoughts: Students' perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*(2), 233-239. https://doi.org/10.1080/03075079712331381064

Cho, K., Schunn, C.D., & Charney, D. (2007). Commenting on writing: Typology and perceiv ed helpfulness of comments from novice peer reviewers and subject matter experts. *Written communication, 23*(3), 260-294. https://doi.org/10.1177%2F0741088306289261

Collins, J.L. (2000). Review of key concepts in strategic reading and writing instruction. In J.L. Collins (Ed.), *Cheektowaga-sloan handbook of practical reading and writing strategies* (pp. 5-10). Cheektowaga-Sloan Union Free School District.

Cömert, M. & Kutlu, Ö. (2018). The effect of self-assessment on achievement in writing in english. *Journal of Educational Sciences Research*, *8*(1), 107-118. https://dergipark.org.tr/tr/pub/ebader/issue/44691/555166

Creswell, J.W. (2005). *Educational research: Planning, conducting, and evaluating quantitati ve and qualitative research* (2nd ed.). Pearson.

Çağımlar, Z., & Oflazoğlu, A. (2002). Evaluation of written and oral expression skills (compo sition) in primary 5th grades in terms of teacher and students' opinions. *Cukurova University Faculty of Education Journal, 2*(23), 12-22. https://l24.im/v9E5eFq

Desoete, A., & Roeyers, H. (2002). Off-line metacognition-A domain-specific retardation in young children with learning disabilities. *Learning Disability Quarterly, 25*(2), 123-139. https://doi.org/10.2307%2F1511279

Earl, L., & Katz, S. (2006). *Rethinking classroom assessment with purpose in mind. Assessme nt for learning, assessment as learning, assessment of learning*. Western and Northern Canadian Protocol for Collaboration in Education (WNCP). https://l24.im/zTPWItn

Eckes, T. (2008). Raters types in writing performance assessment: A classification approach to rater variability. Language Testing, 25(2), 155-185. https://psycnet.apa.org/doi/10.1177/0265532207086780

Edwards, R., Ranson, S., & Strain, M. (2002). Reflexivity: towards a theory of lifelong learning. International Journal of Lifelong Education, 21(6), 525-536. https://doi.org/10.1080/02601370220000016749

Erman Aslanoğlu, A., Sata, M., & Karakaya, İ. (2020). Evaluation of university ÖĞRENCİs' rating behaviors in self and peer rating process via many facet rasch model. *Eurasian Journal of Educational Research, 20*(89), 25-46. https://doi.org/10.14689/ejer.2020.89.2

Falchikov, N., & Goldfinch, J. (2000). Students' peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322. https://doi.org/10.3102%2F00346543070003287

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education, 6*(2), 229-246. https://doi.org/10.1080/13562510120045212

Ferris, D.R. (1997). The influence of teacher commentary on students' revision. *TESOL Quar terly, 31*(2), 315-339. https://doi.org/10.2307/3588049

Gardner, D. (2000). Self-assessment for autonomous language learners. *Links and Letter, 7*(1), 49-60. https://l24.im/isyk

Hanrahan, S., & Isaacs, G. (2001). Assessing self and peer-assessment: The students' views. *Higher Education Research and Development, 20*(1), 53-70. https://doi.org/10.1080/07294360123776

Howell, D.C. (2002). *Statistical methods for psychology* (5th ed.). Duxbury.

Javaherbashsh, M. R. (2010). The impact of self-assessment on Iranian EFL learners' writing skill. *English Language Teaching, 3*(2), 213-218. https://doi.org/10.5539/elt.v3n2p213

Karatay, H. (2013). Süreç temelli yazma modelleri: 4+1 planlı yazma ve değerlendirme modeli [Process-based writing models: 4+1 planned writing and evaluation model]. In M. Özbay (Ed.), *Yazma eğitimi [Writing education]* (pp. 21-48). Pegem. https://l24.im/uRAJ4y

Khonbi, Z.A., & Sadeghi, K. (2012). The effect of assessment type (self vs. peer vs. teacher) on Iranian university EFL students' course achievement. *Language Testing in Asia, 2*(4), 47-74. https://doi.org/10.1186/2229-0443-2-4-47

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Lang uage Testing, 26*(20), 275-304. https://doi.org/10.1177/0265532208101008

Lam, R. (2010). The role of self-assessment in students’ writing portfolios: A classroom investigation. *TESL Reporter, 43*(2), 16-34. https://l24.im/1tiXK

Nielsen, K. (2021). Peer and self-assessment practices for writing across the curriculum: Learner-differentiated effects on writing achievement. *Educational Review, 73*(6), 753-774. https://doi.org/10.1080/00131911.2019.1695104

Kulm, G. (1994). *Mathematics assessment: What works in the classroom*. Jossey-Bass. https://l24.im/MEVdk

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme [Determining students’ success: Assessment based on performance and portfolio].* Seçkin. https://www.seckin.com.tr/kitap/532235955

Meihami, H., & Varmaghani, Z. (2013). The implementation of self-assessment in EFL writing classroom: An experimental study. *International Letters of Social and Humanistic Sciences, 9*(1), 39-48. https://doi.org/10.18052/www.scipress.com/ILSHS.9.39

Merriam, S.B. (2015). *Nitel araştırma: Desen ve uygulama için bir rehber [Qualitative research: a guide to design and implementation]* (S. Turan, Trans.). Nobel (Original work published 2013, 3th ed.).

Nakanoshi, C. (2015). The effects of different types of feedback on revision. *The Journal of Asia TEFL, 4(*4), 213-224. https://l24.im/f6YxtBQ

Miles, M.B., & Huberman, A.M. (1994). Qualitative data analysis (2nd ed.). Sage. https://psycnet.apa.org/record/1995-97407-000

Mistar, J. (2011). A study of the validity and reliability of self-assessment. *TEFLIN Journal, 22*(1), 45-58. http://dx.doi.org/10.15639/teflinjournal.v22i1/45-58

Moussaoui, S. (2012). An investigation of the effects of peer evalu¬ation in enhancing Algerian students’ writing autonomy and positive affect. *Procedia: Social and Behavioral Sciences, 69*(1), 1775-1784. https://doi.org/10.1016/j.sbspro.2012.12.127

Noonan, B., & Duncan, C. (2005). Peer and self-assessment in high schools. *Practical Assessment Research & Evaluation, 10*(17), 1-6. https://doi.org/10.7275/a166-vm41

Oscarson, A.D. (2009). *Self-assessment of writing in learning English as a foreign language: A study at the upper secondary school level.* Geson Hylte Tryck. https://files.eric.ed.gov/fulltext/ED505960.pdf

Pierce, L.V. (2003). *Assessing English language learners*. National Education Association. https://l24.im/l3vxESD

Popham, W.J. (2006). *Assessment for educational leaders.* Allyn & Bacon. https://l24.im/vrbpOo

Richer, D.L. (1992). *The effects of two feedback systems on first year college students’ writing proficiency* [Doctoral dissertation, University of Massachusetts Lowell]. Dissertation Abstracts International, 53, 2722. ProQuest. https://l24.im/Y4Sws2V

Ross, J.A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6*(1), 107-132. https://doi.org/10.1016/S1075-2935(99)00003-3

Ruegg, R. (2015). The relative effects of peer and teacher feedback on improvement in EFL students’ writing ability. *Linguistics and Education, 29*(1), 73-82. https://doi.org/10.1016/j.linged.2014.12.001

Sadler, P., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1-31. https://doi.org/10.1207/s15326977ea1101_1

Sperling, M., & Freedman, S.W. (2001). Research on writing. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.) (pp. 370-389). American Educational Research Association. https://l24.im/ie4D1

Strijbos, J.W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functi onal, and conceptual developments. Learning and Instruction, 20(4), 265-269. https://doi.org/10.1016/j.learninstruc.2009.08.002

Sun, C., & Feng, G. (2009). Process approach to teaching writing applied in different teaching models. English Language Teaching, 2(1), 150-155. https://doi.org/10.5539/elt.v2n1p150

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6th ed.). Pearson. https://l24.im/5nQpO

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessm ent: In search of qualities and standards* (pp. 55–87). Kluwer. https://doi.org/10.1007/0-306-48125-1_4

Uysal, K. (2008). Involving students' in the assessment process: Peer assessment and self-ass essment *[Master dissertation, Abant İzzet Baysal University].* https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Wang, J., & Luo, K. (2019). Evaluating rater judgments on ETIC advanced writing tasks: An application of generalizability theory and many-facets Rasch model. *Papers in Language Testing and Assessment, 8*(2), 91–116. https://l24.im/jJG

Weigle, S.C. (2002). *Assessing writing.* Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Yıldırım, A., & Şimşek, H. (2011). *Sosyal bilimlerde nitel araştırma yöntemleri (7. Baskı) [Q ualitative research in social sciences, 7th ed.].* Seçkin.

Young, J.E., & Jackman, M.G.A. (2014). Formative assessment in the Grenadian lower secondary school: Teachers' perceptions, attitudes and practices. *Assessment in Educati on: Principles, Policy & Practice, 21*(4), 398-411. https://doi.org/10.1080/0969594X.2014.919248

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 197–217

https://doi.org/10.21449/ijate.1111886

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Measurement-evaluation applications of context-based activities in hybrid learning environments

Ahmet Kumas[1,*]

[1]Usak University, Ulubey Vocational High School, Department of Medical Services and Techniques, Usak, Türkiye

ARTICLE HISTORY

Received: May 1, 2022
Revised: Oct. 17, 2022
Accepted: Oct. 17, 2022

Keywords:
Measurement-evaluation,
Context-based learning,
Hybrid learning,
Science education,
Covid-19.

Abstract: Students may be at a disadvantage when learning if they cannot follow lessons face to face due to such reasons as epidemics, disasters, transportation, or family. The main purpose of this study is to perform alternative measurement and evaluation practices in hybrid learning environments in a way that will make students in online physics lessons active participants in the process. The research uses the developmental, emancipatory, and critical action research models within the scope of the qualitative research method. The research was carried out over three weeks under the guidance of the researcher with 32 10th-graders at the school where the researcher taught physics for 12 years. Semi-structured interview forms, rubric forms, and documents were used as data collection tools. The interviews and documents were evaluated using content analysis, while the rubrics were evaluated using descriptive analysis. The students' active and decisive roles during the assessment and evaluation activities within the context-based learning activities regarding physics subjects as well as at the end of learning encouraged the students attending the lesson online and those attending in person to learn under the same conditions. In this context, activities in which students are a part of the learning and measurement-evaluation processes should be encouraged in online and hybrid-learning environments. Developing context-based activities with regard to experiments, analogy, and theoretical applications and developing qualified practices in which students will be active throughout the process under the guidance of action researchers will be beneficial for ensuring this.

1. INTRODUCTION

Alongside technology's presence having become felt intensely in all areas of social life in recent years, high-level studies have been carried out on the effective use of technological applications in health, education, economy, communication, defense, and transportation for increasing the quality of life (Fisher *et al*., 1996). With the COVID-19 pandemic having been the most important agenda item for all countries and societies of the world since 2019, important problems have emerged regarding teaching practices in educational environments during the pandemic (Tarkar, 2020). Economically developed countries have focused on technology-

*Corresponding Author: Ahmet Kumas ✉ ahmetkumas_61@hotmail.com Usak University, Ulubey Vocational High School, Department of Medical Services and Techniques, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

supported education by allocating high budgets to provide equal education practices without interruption to all individuals in their countries. In addition, these countries have ensured equal learning opportunities by including economically and socially disadvantaged student groups in their educational environments (Reuge *et al.,* 2021).

With the COVID-19 pandemic, face-to-face education was suspended for a long time in primary, secondary, and high schools in Türkiye and the world, and lessons started in alternative learning environments (Koçoglu & Tekdal, 2020). Teaching was provided in face-to-face and hybrid environments in accordance with course content. While educating in such new environments, new opportunities as well as problems emerged. Opportunities to apply these new experiences in teaching environments have emerged since the COVID-19 pandemic, one of which is the education and assessment-evaluation practices hybrid environments provide to students who are described as disadvantaged student groups and who have avoided face-to-face education in classes due to illness, the pandemic, or family reasons (Xie *et al*., 2020).

Hybrid education is defined as the situation of some students attending courses online while other students attend schools in-person. Hybrid education can advance verbal and visual communication by minimizing the communication distance between the instructor and the learner (Triyason *et al*., 2020). In hybrid learning, face-to-face or online monitoring of learning can be shaped according to the student's request, the teacher's planning, and the requirements of the environmental conditions (Potra *et al*., 2021). Some areas where student-centered and interactive applications are most needed in the hybrid teaching process within the scope of science are physics, chemistry, and biology courses. Due to the courses and subjects within the scope of science involving contents such as theories, experiments, observations, and applications, offering these courses simultaneously both face to face and to online students is very difficult. As a result, making qualified measurements and evaluations is also very difficult (Senel & Senel, 2021).

Society's needs and lifestyles have also started to change rapidly in parallel with the global technological developments. Accordingly, a need for change has emerged in the curricula, textbooks, and application activities of science courses that support technology in order to meet today's needs (Syafril *et al*., 2021). In this context, the physics curriculum and physics textbooks in Türkiye have been renewed, structured with context-based content, and updated in 2007, 2013, and 2018. The new curricula have been encouraged to present problems, concepts, and contexts of daily life effectively in the classroom environment, apply life-based assessment-evaluation practices throughout education, and give students active roles in all processes (Dicle Erdamar, 2019).

Teaching scientific knowledge in an interactive student-centered manner by associating it with life-based examples is called context-based learning, in which scientific concepts are associated and presented with meaningful events in students' lives (Hansman, 2001). Context-based learning can implement measurement and evaluation applications at every moment of the process. Using real-life scenarios while determining assessment-based problems encourages students to learn and results in them developing positive attitudes toward learning (Williams, 2008). In this sense, context-based problem-solving activities are the basic requirement of context-based learning. Since problems can involve more than one context, they should be prepared under the guidance of experts at a level that does not create misconceptions when determining their content (Yu *et al*., 2015). Context-based learning must include four stages in the learning outcomes; namely, (1) associating and integrating new concepts and information with the old information that has been learned, (2) creating the conceptual and theoretical infrastructure for the newly obtained information, (3) presenting the conceptually obtained and mentally modeled information by associating it with daily life problems, and (4) presenting the results in a report (Edelson *et al*., 1999).

Context-based measurements and evaluations structure information in depth and associate it with daily life problems in groups. While the teacher can prepare and present daily life scenarios, the students can also transform and present the previously learned information in scenarios using video and cinema (Avargil *et al.,* 2012) The transfer stage of learning prefers seeing student groups presenting videos and movies as daily life scenarios in the classroom environment through collaborations (Utami *et al*., 2016). Detective, action, military, and science fiction films are suggested as context tools that can be used for measurements and evaluations in context-based learning (Yu *et al*., 2015). The use of simulations as a tool for assessing subjects that can consume a lot of time abstractly, experimentally, or both makes students more interested in concepts and subjects. In addition, simulation scenarios help students learn abstract concepts in-depth and present their knowledge and skills more easily (De Jong & Van Joolingen, 1998).

Many studies are found in the literature on the effects of context-based teaching on learning, the positive-negative effects of hybrid learning, and student-centered measurement and evaluation practices in context-based learning. As a result of COVID-19, learning environments have moved out of the classroom all over the world, and online learning and hybrid learning environment (HLE) with active student participation have been developed through the effective and purposeful use of technology. One of the problems experienced in this process has been how to simultaneously perform a qualified evaluation of students (both online and in class) throughout the process (Makhachashvili, 2021). Context-based alternative measurement-evaluation practices are not found in the literature on online and HLEs.

Some problems have been encountered in the practice of HLE teaching, the most common of which are being able to simultaneously monitor the student groups participating in the lesson online and in class, accessing qualified materials that are able to attract the attention of both groups, keeping students' constant interest through measurement-evaluation activities in all learning processes, and designing learning environments in line with the teaching goals (Villegas-Ch *et al*., 2021). As a result of the effect of COVID-19 on students in all age groups, students who test positive for COVID-19 or who are suspected of having contracted the virus are not permitted to attend lessons in class for an extended time due to the quarantine conditions. Hybrid education applications appear as ideal environments for benefitting from teaching practices under the same conditions for students attending their courses in classroom environments (Benito *et al*., 2021). To eliminate the negative aspects of HLEs, emphasis should be placed on teaching that will attract the attention of high school student groups. In order to do this, technology-supported, student-centered, analogy-containing virtual laboratory applications and experiments that will enable students to develop positive attitudes toward science lessons should be used effectively (Dexter & Richardson, 2020). Considering that hybrid education applications will continue to increase after the Covid 19 epidemic, context-based measurement-evaluation applications are needed under the guidance of action researchers who personally face problems in learning environments and have researcher identities. Considering that the content of the lessons in hybrid education is provided with technology support to the student groups participating in the lessons online, teachers should use technological opportunities at an advanced level for educational purposes while teaching concepts, interacting with students, and making measurement-evaluation. In this context, this research study aims to develop alternative measurement-evaluation practices under the guidance of an action researcher in a way that will allow students attending online physics lessons in HLEs to be active participants in the process. In line with this purpose of the research, answers to the following questions are sought:

1. What student-centered measurement-evaluation methods exist in the context-based teaching process in HLEs, and how can they be applied in the relating stages?

2. Which student-centered measurement-evaluation methods can be applied during the stages of experiencing and applying in the context-based teaching process in HLEs, and how can they be applied?

3. Which student-centered measurement-evaluation methods can be applied in the relating stage of the context-based teaching process in HLEs, and how can they be applied?

4. What are students' views on the alternative measurement-evaluation methods used in context-based teaching in HLEs?

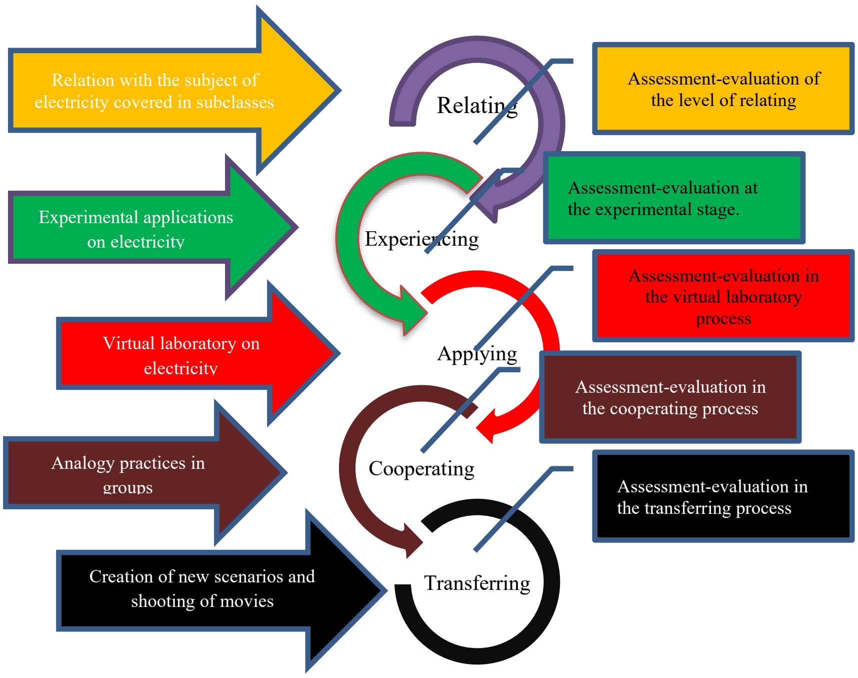
2. METHOD

2.1. General Background

This research uses the critical action research models within the scope of qualitative research. The main purpose of these models is to have students experience new knowledge, skills, and experiences under the guidance of the researcher and guide the development of the process in accordance with the learning objectives through the practitioner’s and students’ critical perspectives. Thus, the action researcher will see the deficient and clear aspects of their practices and have the opportunity to develop them. This approach aims to develop applications by considering students’ readiness levels for theoretical studies in particular (Yıldırım & Şimşek, 2016). The study prefers a hybrid learning environment and action research because the researcher taught physics, whose curriculum in Türkiye is context-based, for 12 years at the school where the application was made, and some students had been attending classes online and others in class due to COVID-19. The process was additionally carried out in HLEs using the developmental, emancipatory, and critical action research models by applying alternative and student-centered measurement and evaluation practices at every stage to encourage students to participate actively in the whole process.

This research adapts the REACT teaching strategy, provides the teaching of, and applies the measurements-evaluations of the 10th-grade electricity subject in terms of a context-based approach. The REACT teaching strategy has been structured by considering the stages from Crawford’s (2001) study. The implementation process is shown in Figure 1.

Figure 1. *The framework of context-based measurement-evaluation learning activity.*


Relation with the subject of electricity covered in subclasses
Relating
Assessment-evaluation of the level of relating
Experimental applications on electricity
Experiencing
Assessment-evaluation at the experimental stage.
Virtual laboratory on electricity
Applying
Assessment-evaluation in the virtual laboratory process
Analogy practices in groups
Cooperating
Assessment-evaluation in the cooperating process
Creation of new scenarios and shooting of movies
Transferring
Assessment-evaluation in the transferring process

During the relating phase, the students were shown the movie “The Current War” (Gomez-Rejon, 2017). Since science courses in Türkiye have a spiral structure, the subject of electricity had gradually been covered during the previous five years. In order to activate students’ pre-knowledge, the request was made to note the factors affecting the brightness of lamp bulbs; the factors affected by resistance, voltage, and current; and the situations involving electricity being converted to heat, light, or sound on the worksheet. Peer evaluations were also requested by taking into account the grades each group member received. After discussing the concepts in the movie in groups, each group was given three minutes to make a presentation in order to compare the common decisions of all groups. Differences of opinion between groups were resolved with group discussions. In the experience phase, the teacher provided the student groups with a context-based scenario. Experimental applications and measurements were made under this scenario, and the students were asked to fill in the relevant figures and tables. Peer evaluations were made according to the ability of each person in the group to fulfill their role and responsibilities within the group. Peer group evaluations were carried out based on the groups’ experimental practices. The experiencing phase lasted for one lesson. During the application phase, a text on conceptual change was handed out, and elimination of the students’ misconceptions that could have arisen over the previous years was ensured. In addition, the students were asked to fill in the analogy map and then discuss it as groups. The application phase lasted for one class hour. During the collaboration phase, the student groups were asked to design and write a movie scenario that included concepts related to electricity. The students were asked to design in a virtual laboratory environment a model electrical system that they used at home in daily life. During the transfer stage, the groups were asked to implement the script they had written as a short 10-minute film. The stages of collaboration and transfer lasted a week (i.e., two course hours in total).

2.2. Sample

The research was conducted with 32 students studying at Mehmet Akif Ersoy Anatolian High School in the 2022 spring semester in Araklı, Trabzon, Türkiye. The applications on the 10th-grade students were maintained over three weeks for six-lesson hours. The school accepted students who had an average academic grade of 89 or higher in the 2021 fall semester. The students’ academic achievement level at school was slightly above medium level. The researcher holding a master's and a doctorate in physics education had 12 years of experience in teaching physics at the school where the research was conducted. Interviews with students lasted between 17-23 minutes. The interviews were conducted with 12 male and 20 female students. As for the students’ 2021 fall semester academic averages, seven students were between 50-70, 19 were between 70-85, and six were between 85-100. Also, 12 of the students were boarding students and 20 were day students.

2.3. Data Collection

For the research ethics permission, permit number 2022-10 dated March 8, 2022 with registration number E-54749836-050.99-71646 was obtained from the Usak University Rectorate of Science and Engineering Sciences Scientific Research and Publication Ethics Committee.

The research data were obtained using a semi-structured interview (SSI) form, the documents the students filled out during the process, and a structured observation form. One of the researcher’s main goals was to obtain in-depth information from the students in the context-based activities in the HLE (Yin, 2009). During the evaluation of the measurement and evaluation practices within the scope of the action research, five interview questions were finalized to reflect the content of the research using the opinions of two academicians who are experts in their fields and of a psychological counselor guidance teacher. The third sub-goal of the research asked the students the following questions: What are your views on peer measurements and the peer group measurement practices in the teaching process? What are your views on the movie you watched at the beginning of the subject, on the scenario-based experiment, and on the simulation measurement? What are your opinions on the evaluation application? What are your opinions on the measurement of the activities where you wrote the script and shot short films? What are your opinions on the measurement applications regarding electricity?

Unstructured field study is a type of observation and was also used to arrive at the findings on how to apply measurement-evaluation methods for the first, second, and third sub-goals of the research. In unstructured observations, the researcher assumes the role of a participant observer (Yıldırım & Şimşek, 2016). The Mathematics and Science Classroom Observation Profile System (M-SCOPS) was developed by Stuessy et al. (2003); it was restructured within the scope of this research and turned into a draft form. With the observation form, the proficiency of the students at each stage was observed under five categories. Behaviors and practices were also noted that would support the research but were not found on the observation form. The SSI was used to compare the applications the teachers and students mentioned on the interview form and the applications within the scope of hybrid education.

For assessing the context-based activities in the hybrid learning process, the findings obtained from the peer and peer group measurements and evaluations of the students regarding all the processes as well as the teacher's findings obtained from the observation and document data of the students were evaluated with regard to their passing grades. The film, experiment, simulation, and new scenario stages were evaluated at 100 points each. Students noted their scientific knowledge and opinions about each stage on the worksheet. This paper was evaluated by the teacher within the scope of the document analysis. Observation findings were evaluated at a maximum of 100 points per stage. The students' evaluations of their peers and the peer group evaluations for the four stages were calculated at a maximum of 100 points each. The eight evaluations of the teacher and students for each student were collected, and the students' evaluation grades were obtained based on the total scores from the evaluations divided by 16.

2.4. Validity and Reliability

Achieving internal validity in action research is based on having the determined situations be consistent with reality and reflect the truth. In order to ensure the internal validity (credibility) of the research, one needs to explain the system of the evidence; provide diversity in the data, participant approval, and long-term interactions; and reveal the appropriate patterns and model (Yıldırım & Şimşek, 2016). External validity in qualitative research is based on being able to generalize the results obtained (Noble & Smith, 2015). In order to ensure validity and reliability within the scope of this research, data diversification was made through the interviews, the interview and document analyses, the presentation of the process for obtaining the research data alongside the evidence, and providing the participants' volunteer statements. Long-term interactions were additionally provided with the students as a result of the researcher having been a teacher at the boarding school as well as a physics teacher at the school for two years. Prior to the research, a broad literature study was conducted, and the developmental, emancipatory, and critical action research models were determined to be suitable to the nature of the research. In accordance with Miles and Huberman's (2015) agreement analysis formula for interview data (intercoder agreement = number of common opinions / [number of common opinions + number of different opinions]), the encoder similarity rate was calculated as 89% based on the results from the two expert examinations. This value shows the intercoder agreement to be high (Miles & Huberman, 1994). Interviews were conducted outside of class hours by informing the participants beforehand and obtaining the necessary permissions from them. During the interviews, the researcher did not interfere with the participants' views; also, probing questions were asked in places that went beyond the subject. Video recordings were taken with the permission of each participant.

2.5. Data Analysis

Content analysis was conducted using the appropriate themes, categories, and codes for the interview data. Content analysis involves comprehensively examining written statements that are similar in terms of meaning in order to ensure that readers and researchers can understand them in a way that creates integrity (Yıldırım & Şimşek, 2016). Codes with similar content are combined for ease of understanding. Descriptive analysis has been used to evaluate the observation findings. In descriptive analysis, direct statements are used to reflect the individuals' situations, views and the environment in which they are observed. The purpose of this type of analysis is to present the findings to the reader in an organized and interpreted form (Yıldırım & Şimşek, 2016). In this context, a four-stage descriptive analysis was adapted to the research, which involved deciding which template to follow, as well as the data processing and interpretation, and making sense of the data. While performing the descriptive analysis of the observation data, the scoring criteria of "No response/cannot be coded," "Alternative idea," and "Scientific idea" were used in the coding (Nassaji, 2015). Peer and peer-group evaluation forms were made based on Patri's (2002) study, with scores ranging from 1 to 5. While evaluating the scores of the students during the learning stages, the grading systems in their schools were taken as the basis. The scores in the student evaluation system were converted into a five-point system and an evaluation was made. Multiple linear regressions models were used to estimate beta coefficients and 95% confidence intervals. While interpreting the findings, students' opinions (direct quotations) were included to make the subject more understandable. The students have been encoded as S1, S2, …, S16 in order to preserve their anonymity.

3. FINDINGS

3.1. Findings Regarding Assessment-Evaluation in the Relating Stage

After watching the movie, which includes the concepts related to electricity, the students were asked to take notes in the relevant section of the worksheets, where which events and in which second of the movie the basic concepts related to electricity were used. Then, the students were asked to compare the answers within the group at the stage of group work. The data obtained from the students' evaluation of each other after the groups' common ideas were formed are shown in Table 1.

As seen in Table 1, in the relating phase, after watching the movie about electricity, the students evaluated their peers and received high scores in the themes of volunteering to work, sharing what they know, and working together. On the other hand, the students received low scores on such themes as duty responsibility and cooperation and also exhibited high-level behaviors in the codes of learning by taking notes with questions that developed their sense of curiosity by doing research voluntarily and interactively during the process of watching movies on the subject of electricity. The scores of the students who attended the course online and face to face are close to each other.

Table 1. *Peer measurement data at the relating stage.*

Theme	Category	Cods	*N*	$\overline{X}$
Participates in studies voluntarily	Face-to-face	Theoretical work, taking notes while watching the movie, interaction with the course content, curiosity	25	4.7
	Online		7	4.6
Shares what he knows with his friends	Face-to-face	Asking questions, interacting in intriguing places	25	4.5
	Online		7	4.5
Helps friends when needed	Face-to-face	Active role in the group, research when questioned	25	4.0
	Online		7	4.0
Gathers information from different sources	Face-to-face	Scientific resources, scientific content internet resources	25	4.4
	Online		7	4.2
Respects the opinions of his group mates	Face-to-face	Don't care even if they have different opinions, have the right to speak as much as necessary	25	4.3
	Online		7	4.4
Duty responsibility is at a high level	Face-to-face	Homogeneity in task sharing	25	3.9
	Online		7	4.1
Likes to work together	Face-to-face	Volunteering, Willingness for new knowledge	25	4.5
	Online		7	4.2
Contribution to the formation of the group idea	Face-to-face	Original ideas, contribution to group opinion	25	4.2
	Online		7	4.0
Total	Face-to-face		32	4.3
	Online			

After each student group wrote their common views on the worksheets, the groups made presentations and compared their views. The data obtained from the evaluations of the groups as a result of the presentations are shown in Table 2.

Table 2. *Peer group evaluation data in the relating phase.*

Theme	Category	Cods	*N*	$\overline{X}$
Presentation	Face-to-face	Time use, content, persuasion	6	4.3
	Online			
Accuracy of information	Face-to-face	Inclusivity, scientific		4.4
	Online			
Collaboration of group members	Face-to-face	Collaboration, research when questioned		4.6
	Online			
All group members fulfill individual responsibilities	Face-to-face	Involvement of the whole group, individual responsibility		3.7
	Online			
Interaction of group members	Face-to-face	Everyone has a say, everyone contributes		4.2
	Online			
Task sharing competence	Face-to-face	Homogeneity in task sharing		3.8
	Online			
Persuasion competence	Face-to-face	Scientific persuasion, collaborative persuasion		4.4
	Online			
Learning competence of group members	Face-to-face	Whole group learning, individual competence		4.7
	Online			
Total	Face-to-face		6	4.4
	Online			

As seen in Table 2, student groups scored high in learning competencies and group cooperation themes, and scored low in task sharing and fulfilling responsibilities. The findings obtained as a result of the document analysis regarding the process performances of the students in context-based activities in hybrid learning environments are shown in Table 3. As seen in Table 3, as a result of the document review, the students got advanced scores in electrical energy, brightness and current intensity, but low scores in potential difference.

Table 3. *Document review data on students' competencies in the process at the relating stage.*

Concepts	Events in the movie	N	$\overline{X}$
Electrical voltage	Burning of the lamps	32	3.2
Electrical current	Increasing or decreasing the brightness of the lamps		4.2
Resistance	Using lamps with different characteristics		4.0
Brightness	The increase in light intensity with the change of the characteristics of the generator and lamps		4.4
Electrical energy	Illumination of environments with light connected to electricity		4.1

3.2. Findings Regarding Measurement-Evaluation in Experiencing and Applying Stages

In the electrical circuit on the house presented as a model in the worksheet, the circuit elements were placed in series and parallel and the drawings were made in the figure. The data obtained from the students as a result of the document review are shown in Table 4.

Table 4. *Documentary findings on students' competencies in experiencing and applying stages.*

Concepts	Experiencing and Applying applications	f		$\overline{X}$	
		Experiment	Drawing	Experiment	Drawing
Electrical voltage	Parallel, serial	6	32	4.8	4.4
Current intensity	Association with voltage, branching, main branch			3.9	3.8
Resistance	Parallel, serial			4.0	4.3
Luminescence	Relationship with resistor, association with voltage			4.2	3.8
Electrical energy	Voltage, resistance, current intensity relationship			4.1	3.8

As can be seen in Table 4, in the experimental applications of experiencing and applying stages, electrical voltage and luminosity concepts received high scores. In the drawings, the electrical voltage and resistors exhibited high-level behaviors; scenario-based activities related to electricity were developed as well. By presenting the house model in the worksheet, the students were asked to demonstrate their electrical circuits experimentally by making use of the scenario. During the activity process, the students were evaluated by the teacher with a rubric form, and the data are shown in Table 5.

Table 5. *Evaluation of students with rubric form in experiencing and applying stages.*

Measurement	Factors	*N*	$\overline{X}$
Getting to know the tools	Ammeter, Voltmeter, Generator, Switch, Lamp	32	5.0
Associate concepts	Resistance, Ohm's law, R-i relationship		3.8
The experimental setup	Series circuits, Parallel circuits, Branching of current		4.4
Ability to operate	Electrical energy, current branching		4.6
Simulation setup	Series circuits, Parallel circuits, Branching of current		4.8
Ability to explain	To be able to explain Ohm's law, to associate theory with practice.		4.3

As can be seen in Table 5, students got high scores while creating simulation mechanisms during the experiencing and applying stages. In addition, the tools used in the experiments were successfully recognized by the students and the experimental setups could be operated successfully. In the category of associating concepts, it was revealed that they did not develop enough. In the Experiencing and Applying stages, after the groups' common ideas were formed, the students were asked to evaluate each other within the group. The data obtained from the students' evaluations of each other are shown in Table 6.

Table 6. *Evaluation of students with rubric form in experiencing and applying stages.*

Theme	Category	Cods	*N*	$\overline{X}$
Participates in studies voluntarily	Face-to-face	Theoretical work, taking notes while watching the movie, interaction with the course content, curiosity	26	4.6
	Online		5	4.6
Shares what he knows with his friends	Face-to-face	Asking questions, interacting in intriguing places	26	4.5
	Online		5	4.4
Helps friends when needed	Face-to-face	Active role in the group, research when questioned	26	4.1
	Online		5	4.1
Gathers information from different sources	Face-to-face	Scientific resources, scientific content internet resources	26	4.5
	Online		5	4.3
Respects the opinions of his group mates	Face-to-face	Don't care even if they have different opinions, have the right to speak as much as necessary	26	4.4
	Online		5	4.5
Duty responsibility is at a high level	Face-to-face	Homogeneity in task sharing	26	4.0
	Online		5	4.1
Likes to work together	Face-to-face	Volunteering, Willingness for new knowledge	26	4.3
	Online		5	4.2
Contribution to the formation of the group idea	Face-to-face	Original ideas, contribution to group opinion	26	4.2
	Online		5	4.2
Total	Face-to-face		26	4.3
	Online		5	4.3

As seen in Table 6, in the Experiencing and Applying stages, students evaluated their peers after experiment and simulation applications; they received high scores in the themes of voluntary participation in studies, sharing what they know, respecting the opinions of their groupmates, and collecting information from different sources. At this stage, the scores of students participating in the course online and face-to-face are close to each other.

3.3. Findings Regarding Measurement-Evaluation in Cooperation and Transferring Stages

The data obtained from the worksheet documents for analogy maps during the implementation process in the Cooperation and Transferring stages are shown in Table 7. As can be seen in Table 7, as a result of the evaluation of analogy map document data, individual student success was 65%, while group success was 96% as a result of the answers they created in interaction with each other. In the Cooperation and Transferring stages, the findings obtained with the help of the rubric form as a result of the groups developing film scenarios and shooting as short films are shown in Table 8.

Table 7. *Finding from the analogy map.*

Situations	Expected answers	*N*		*f*	
		Individual	Group	Individual	Group
Similar feature	Farmers	32	6	19	6
Compare	Comparable			28	6
Simulated feature 1	Electrical current			22	6
Simulated feature 2	Electrical voltage/Current			14	5
Total (%)				65	96

Table 8. *Scenario and short film evaluation findings of student groups.*

Concepts	Experiencing and applying applications	*N*		$\overline{x}$	
		Experiment	Drawing	Experiment	Drawing
Electrical voltage	Parallel, serial	6	6	5.0	5.0
Current intensity	Association with voltage, branching, main branch			4.9	4.9
Resistance	Parallel, serial			4.7	4.8
Luminescence	Relationship with resistor, association with voltage			5.0	5.0
Electrical energy	Voltage, resistance, current intensity relationship			5.0	4.8
Total		6	6	4.9	4.9

As it can be seen in Table 8, after the student groups structured the subjects and concepts related to electricity in-depth in the cooperation and transferring stages, they put forward applications by getting high scores as scenarios and films.

In the cooperation and transferring stages, after the groups' common ideas were formed, the students were asked to evaluate each other within the group. The data obtained from the students' evaluations of each other are shown in Table 9.

Table 9. *Peer evaluation data in cooperation and transferring stages.*

Theme	Category	Factors	*N*	$\overline{x}$
Participates in studies voluntarily	Face-to-face	Scenario creation stage, associating the scenario with electricity concepts, taking part in a short film	26	4.8
	Online		6	4.7
Shares what he knows with his friends	Face-to-face	Asking questions, interacting in intriguing places	26	4.7
	Online		6	4.7
Helps friends when needed	Face-to-face	Active role in the group, research when questioned	26	4.5
	Online		6	4.6
Gathers information from different sources	Face-to-face	Scientific resources, scientific content internet resources	26	4.7
	Online		6	4.9
Respects the opinions of his group mates	Face-to-face	Don't care even if they have different opinions, have the right to speak as much as necessary	26	4.5
	Online		6	4.5
Duty responsibility is at a high level	Face-to-face	Homogeneity in task sharing	26	4.6
	Online		6	4.7
likes to work together	Face-to-face	Volunteering, Willingness for new knowledge	26	4.8
	Online		6	4.7
Contribution to the formation of the group idea	Face-to-face	Original ideas, contribution to group opinion	26	4.5
	Online		6	4.6
Total	Face-to-face		26	4.8
	Online		6	4.7

As seen in Table 9, students exhibited high-level behaviors in all categories in cooperation and transferring stages. The peer measurement scores of the students who attended the course online and face-to-face are close to each other.

After each student group wrote their common views on the worksheets, the groups made presentations and compared their views. The data obtained from the evaluations of the groups after the presentations are shown in Table 10.

Table 10. *Peer group evaluation data at the cooperation and transferring stage.*

Theme	Category	Cods	Groups (*N*)	$\overline{x}$
Presentation	Face-to-face	Time use, content, persuasion	6	4.8
	Online			
Accuracy of information	Face-to-face	Inclusivity, scientific		4.9
	Online			
Collaboration of group members	Face-to-face	Collaboration, research when questioned		4.8
	Online			
All group members fulfil individual responsibilities	Face-to-face	Involvement of the whole group, individual responsibility		4.7
	Online			
Interaction of group members	Face-to-face	Everyone has a say, everyone contributes		4.7
	Online			
Task sharing competence	Face-to-face	Homogeneity in task sharing		4.8
	Online			
Persuasion competence	Face-to-face	Scientific persuasion, collaborative persuasion		4.6
	Online			
Learning competence of group members	Face-to-face	Whole group learning, individual competence		4.7
	Online			
Total	Face-to-face		6	4.8
	Online			

As seen in Table 10, student groups achieved high scores by exhibiting high-level behaviors in all categories in Cooperation and Transferring stages.

3.4. Student Views on Measurement-Evaluation Methods in Context-Based Teaching Process in Hybrid Learning Environments

In the context-based teaching process in the hybrid learning environment, the students' views on measurement and evaluation as a result of the teaching practices based on the REACT strategy are shown in Table 11.

In Table 11, when students' views on measurement-evaluation methods in context-based teaching processes in hybrid learning environments are examined, it is seen that there is an intensity in the positive theme. It is seen that the opinions of peer and peer group evaluations cause students to work harder and increase their motivation and success in cooperation. The use of analogy maps as a measurement-evaluation tool comes to the fore in the theme of positivity, which contributes to the structuring of students' knowledge permanently and entertainingly in interaction, and as negativity, it is complicated because it is a new situation.

Table 11. *Student opinions on measurement and evaluation in context-based teaching.*

Theme	Category	Cods	*f*
Positive	Peer evaluation	Objectivity	23
		Motivation	21
		Hard work	20
		Following closely	19
		Success	19
	Peer group evaluation	Success	18
		In-depth learning	12
		Attitude	10
		Research	10
		Equality	9
	Document analysis	Detailed information	13
		Evaluation by grade	12
		Learning all information	11
	Analogy map	Fun	14
		Permanent information	10
		Interaction	10
	Rubric form	Continuous motivation	13
		Keeping up with the lesson	12
		Interacting with the teacher	9
Negative	Peer evaluation	Close friend	23
		Inability to follow	15
		Privacy	11
		Duration	9
	Peer group evaluation	Impartiality	17
		Grade	11
		Duration	10
	Document analysis	Duration	14
		Cooperation	12
	Analogy map	First time event	13
		Complicated	10
	Rubric form	Inability to distinguish	12

As a result of the use of rubric forms as an alternative measurement-evaluation tool in the in-class interaction process in science, it is revealed that students provide long-term interaction and continuous motivation towards the lesson. Some of the students' views on measurement-evaluation methods in the context-based teaching process in hybrid learning environments are as follows:

> *S11: "We had some emotional difficulties in peer measurement in the first activities, we made more qualified measurements in the following activities, taking into account the measurement criteria set by our teacher. Thanks to the peer and peer group evaluation, we felt the obligation to work continuously and efficiently both individually and in the group. This has led to an increase in our individual and group success."*

> *S5: "Watching a movie at the beginning of the lesson both relaxed and motivated us and made us more interested in the subject. The movie was beautiful. Since we watched the*

movie for lesson purposes, taking notes all the time helped us remember the subject better and be more interested in the subject we were going to learn. As a result of these activities, it was quite easy to evaluate our group friends and other groups. I made a comparison with the results I found myself, I compared the compatible ones and those that were not. Apart from taking a little too much time, it was quite productive."

T:30 "I encountered the analogy map for the first time. It took quite a while to understand and interpret at first. I was able to understand how to do it by getting support from my teacher and my groupmates. Using analogies in the lessons gave me a different perspective on the subject. Our teacher's evaluation of us as a result of the analogy map caused us to be more careful and to behave carefully."

T:19 "While we were learning about electricity, we did many activities. As a result of these activities, it would not be efficient if we took notes with a single evaluation. In addition to individual and group evaluations in each activity process, our teacher's evaluation of our notes and our behavior in the group motivated us individually and as a group. On the negative side, it takes a lot of time to constantly evaluate our friends and other groups. When making a measurement, sometimes the fact that friends look at what we have written prevents us from being objective towards them."

Content analysis was conducted using the appropriate themes, categories, and codes for the interview data. Content analysis involves comprehensively examining written statements that are similar in terms of meaning in order to ensure that readers and researchers can understand them in a way that creates integrity (Yıldırım & Şimşek, 2016). Codes with similar content are combined for ease of understanding.

4. DISCUSSION and CONCLUSION

The research on HLEs was conducted by applying alternative measurement-evaluation practices in a way that would have the students attending the physics lessons online be as active as the students attending the lessons in person. In the relating phase of the context-based teaching practices in HLEs, effective communication was ensured between the students who participated in the lesson online and face-to-face; they were also ensured to share their knowledge in their interactions. In addition, the use of movies that would attract students' attention during the relating phase in the context-based teaching process increased the students' interest in physics subjects. As a result, their voluntary participation in the individual and group activities was also ensured. Due to hybrid learning being a process in which students in different learning environments are provided with interactive learning activities, students who attend the course online may be at a disadvantage with regard to their learning. Including context-based practices in learning activities as well as peer measurements during the assessments-evaluations encourages students to be active throughout the process and requires them to concretely present their contributions to the group work. Murray *et al*.'s (2012) research revealed positive developments to occur more with students' success and attitudes as the rate of interaction among the students who attend the course online increases.

Although online-supported context-based learning is not new, student-centered measurement-evaluation practices within the scope of context-based education in HLEs are lacking in the literature. Many teachers assume that they will apply context-based learning applications in online environments within the scope of technology-assisted teaching applications if it is needed in their daily lives. However, one of the essential stages of context-based learning applications is the use of measurement and evaluation activities that make students active throughout the process. The focus is not on the technology itself in online and HLEs but on the context-based activities they support and the measurement-evaluation practices that will make students active throughout the process. For example, Pathoni *et al*. (2021) revealed context-based measurement applications to be something students in physics lessons in online environments need, but these applications do not provide a type of teaching that will make students active. Similarly, Sulistiyono *et al*. (2021) expressed the need to enrich teacher guide materials in context-based applications in online environments; however, they did not present applications related to the content of measurement-evaluation applications that will keep interactions at high levels throughout the process.

The use of peer group measurements-evaluations as an assessment tool in the relating stage of the context-based teaching process in HLEs encourages all group members to construct the knowledge they have learned over the past years. This is because the other groups consider having even one person in a group not take on a role or fulfill their responsibility to be a negative aspect. In such a case, the scores of each member in that group will suffer. The use of peer and peer group measurements and evaluations during the relating stage contributes to effective learning in individual and collaborative environments. The reason for this is revealed as the evaluation of the teaching activities carried out during each stage of the process. The use of both peer and peer group assessment-evaluation tools encourages students who attend the class online as well as those who attend in person to engage with the course at higher levels. Freeman's (1995) research revealed peer and peer group assessments and evaluations to encourage students to be actively involved in the learning process.

As a result of the document review, assessing and evaluating students' learning processes in the relating and experiencing-applying stages in context-based applications contribute significantly to their learning outcomes. The reasons for this can be shown as learning each learning outcome and writing it down on worksheets alongside the justifications and then having the teacher evaluate these at the end of the lesson and give feedback to the students. In order for students to be successful as individuals on the document review, all group members must actively participate in the process in the group activities, which emphasizes that the program objectives were fully learned in group interactions at other stages. As a result of the use of rubric forms as a measurement tool in the relating phase, activities suitable can be planned for the level of students in order to eliminate their learning deficiencies by revealing the level at which students have learned the subject's preliminary information. Researchers are recommended to develop alternative measurement tools that can reveal students' readiness levels during the association phase. Corcoran *et al.*'s (2004) research revealed having students evaluate individually and in groups by applying the alternative assessment and evaluation practices while implementing activities in the process of learning in-depth knowledge of science concepts encourages students to learn.

The use of analogy maps as an assessment tool is not a frequently encountered situation in physics teaching. Within the scope of the results obtained from these research data, informing the students about analogy maps at the beginning of the subject would be beneficial. In addition, during the cooperative and transferring stages, having students measure and evaluate the learning process individually by using analogy maps and then finish it by measuring and evaluating the groups with the help of analogy maps would contribute to students' in-depth knowledge of the subject and its concepts. Other researchers are recommended to examine the effectiveness of analogy maps by using them in the relating and experiencing stages.

Including life-based practices is important while carrying out measurements and evaluations within the scope of context-based learning. In order to raise to higher levels high school students' interests and motivations as a generation intertwined with technology and to enable them to learn about subjects and concepts in a qualified way, having them write movie scenarios related to the subject and concepts within the scope of their interests and shoot these scenarios as short films contribute to the teaching objectives. Having them print out and implement the movie scenarios in the collaborative and transfer stages contributes to the realization of learning in a more qualified manner. The reason for this can be shown activities helping students realize how to transfer concepts to new situations after teaching them. Chase *et al*.'s (2019) research stated that transferring concepts newly learned in science to new situations is a high-level learning activity. Including practices that will appeal to students' interests and attitudes is encouraged so that transfer to new situations can occur.

In context-based learning, the peer measurement and peer group measurement scores in the relating stage were determined to be lower than the scores in the collaboration and transfer stages. The fact that students had continued their group work interactively for a long time shows this to lead to more successful results with regard to collaborations with the subjects to be learned. In this context, the inclusion of alternative measurement and evaluation practices in all processes while conducting teaching practices shows that students learn subjects and concepts in depth, and this leads them to apply the subjects and concepts to new situations.

Syafril *et al*. (2021) compiled research from different countries of the world over the last 10 years. In this context, hybrid learning environments in countries such as Taiwan, Belgium, Indonesia, England, and Germany were shown to reveal practice deficiencies to exist regarding practices that enable learning activities in the collaboration and transfer stages despite their contribution to students' problem-solving skills.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Usak University, 08/03/2022-2022/10.

Orcid

Ahmet Kumas https://orcid.org/0000-0002-2898-9477

REFERENCES

Avargil, S., Herscovitz, O., & Dori, Y.J. (2012). Teaching thinking skills in context-based learning: Teachers' challenges and assessment knowledge. *Journal of Science Education and Technology*, *21*(2), 207-225. https://doi.org/10.1007/s10956-011-9302-7

Benito, Á., Dogan Yenisey, K., Khanna, K., Masis, M.F., Monge, R.M., Tugtan, M.A., ... & Vig, R. (2021). Changes that should remain in higher education post COVID-19: A mixed-methods analysis of the experiences at three universities. *Higher Learning Research Communications*, *11*, 51-75. https://doi.org/10.18870/hlrc.v11i0.1195

Chase, C.C., Malkiewich, L., & S Kumar, A. (2019). Learning to notice science concepts in engineering activities and transfer situations. *Science Education*, *103*(2), 440-471. https://doi.org/440-471.10.1002/sce.21496

Corcoran, C.A., Dershimer, E.L., & Tichenor, M.S. (2004). A teacher's guide to alternative assessment: Taking the first steps. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *77*(5), 213-218. https://doi.org/10.1002/sce.21496

Crawford, M.L. (2001). *Teaching contextually: Research, rationale, and techniques for improving student motivation and achievement in mathematics and science.* CCI Publishing.

De Jong, T., & Van Joolingen, W.R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179-201.

Dexter, S., & Richardson, J.W. (2020). What does technology integration research tell us about the leadership of technology? *Journal of Research on Technology in Education*, *52*(1), 17-36. https://doi.org/10.1080/15391523.2019.1668316

Dicle Erdamar, I.Y. (2019). Lise Fizik dersi öğretim programının program geliştirme bağlamında analizi [Analysis of High School Physics Curriculum in The Context of

Program Development]. *Harran Education Journal, 4*(2), 29-44. http://dx.doi.org/10.22596/2019.0402.29.44

Edelson, D.C., Gordin, D.N., & Pea, R.D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences*, *8*(3-4), 391-450. https://doi.org/10.1080/10508406.1999.9672075

Fisher, C., Dwyer, D.C., & Yocam, K. (1996). *Education & technology: Reflections on computing in classrooms*. Jossey-Bass Publishers.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education*, *20*(3), 289-300.

Hansman, C.A. (2001). Context-based adult learning. *New directions for Adult and Continuing Education, 89*, 43-52.

Koçoglu, E., & Tekdal, D. (2020). Analysis of distance education activities conducted during COVID-19 pandemic. *Educational Research and Reviews*, *15*(9), 536-543. https://doi.org/10.5897/ERR2020.4033

Makhachashvili, R. (2021). Digital hybrid learning individual quality assessment in european and oriental languages programs: Student case study in Ukraine. In *14th International Conference on ICT, Society, and Human Beings, ICT 2021* (Vol. 14, No. 1, pp. 11-22). International Association for Development of the Information Society (IADIS).

Miles, M.B. & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook.* Sage.

Miles, M.B. & Huberman, A.M. (2015). *Nitel veri analizi: Genişletilmiş bir kaynak kitap* (Çev. Ed. S. Akbaba-Altun & A. Ersoy). Pegem Akademi.

Murray, M.C., Pérez, J., Geist, D., & Hedrick, A. (2012). Student interaction with online course content: Build it and they might come. *Journal of Information Technology Education Research*, *11*(1), 125-140.

Nassaji, H. (2015). Qualitative and descriptive research: Data type versus data analysis. *Language Teaching Research*, *19*(2), 129-132. https://doi.org/10.1177/1362168815572747

Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence-Based Nursing*, *18*(2), 34-35. http://dx.doi.org/10.1136/eb-2015-102054

Pathoni, H., Ashar, R., & Huda, N. (2021). Analysis student needs for the development of contextual-based STEM approach learning media in online learning: An evidence from Universities in Jambi, Indonesia. *International Journal on Research in STEM Education, 3*(1), 17-26.

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language testing*, *19*(2), 109-131. https://doi.org/10.1191/0265532202lt224oa

Potra, S., Pugna, A., Pop, M.D., Negrea, R., & Dungan, L. (2021). Facing COVID-19 challenges: 1st-year students' experience with the Romanian hybrid higher educational system. *International Journal of Environmental Research and Public Health*, *18*(6), 1-15. https://doi.org/10.3390/ijerph18063058

Reuge, N., Jenkins, R., Brossard, M., Soobrayan, B., Mizunoya, S., Ackers, J., ... & Taulo, W.G. (2021). Education response to COVID 19 pandemic, a special issue proposed by UNICEF: Editorial review. *International Journal of Educational Development*, *87*, 1-3. https://doi.org/10.1016/j.ijedudev.2021.102485

Senel, S., & Senel, H.C. (2021). Remote assessment in higher education during COVID-19 pandemic. *International Journal of Assessment Tools in Education*, *8*(2), 181-199. https://doi.org/10.21449/ijate.820140

Stuessy, C.L., Parrott, J.A. & Foster, A.S. (2003). Mathematics and science classroom observation profile system (M-SCOPS): Using classroom observation to analyze the how and what of mathematics. In *Annual Meeting of the School Science and Mathematics Association.*

Sulistiyono, E., Missriani, M., & Fitriani, Y. (2021). Constructivism and contextual based learning in improving Indonesian language learning outcomes in elementary school using online learning techniques in the middle of the Covid 19 pandemic. *JPGI (Jurnal Penelitian Guru Indonesia), 6*(1), 304-309. https://doi.org/10.29210/021037jpgi0005

Syafril, S., Latifah, S., Engkizar, E., Damri, D., Asril, Z., & Yaumas, N.E. (2021, February). Hybrid learning on problem-solving abiities in physics learning: A literature review. In *Journal of Physics: Conference Series* (Vol. 1796, No. 1, p. 012021). IOP Publishing. https://doi.org/10.1088/1742-6596/1796/1/01202

Tarkar, P. (2020). Impact of COVID-19 pandemic on education system. *International Journal of Advanced Science and Technology*, *29*(9), 3812-3814.

Triyason, T., Tassanaviboon, A., & Kanthamanon, P. (2020). Hybrid classroom: Designing for the new normal after COVID-19 pandemic. In *Proceedings of the 11th International Conference on Advances in Information Technology* (pp. 1-8). https://doi.org/10.1145/3406601.3406635

Utami, W.S., Ruja, I.N., & Utaya, S. (2016). React (relating, experiencing, applying, cooperative, transferring) strategy to develop geography skills. *Journal of Education and Practice*, *7*(17), 100-104.

Villegas-Ch, W., Palacios-Pacheco, X., Roman-Cañizares, M., & Luján-Mora, S. (2021). Analysis of educational data in the current state of university learning for the transition to a hybrid education model. *Applied Sciences*, *11*(5), 1-18. https://doi.org/10.3390/app11052068

Williams, P. (2008). Assessing context-based learning: Not only rigorous but also relevant. *Assessment & Evaluation in Higher Education*, *33*(4), 395-408. https://doi.org/10.1080/02602930701562890

Xie, X., Siau, K., & Nah, F.F.H. (2020). COVID-19 pandemic–online education in the new normal and the next normal. *Journal of Information Technology Case and Application Research*, *22*(3), 175-187. https://doi.org/10.1080/15228053.2020.1824884

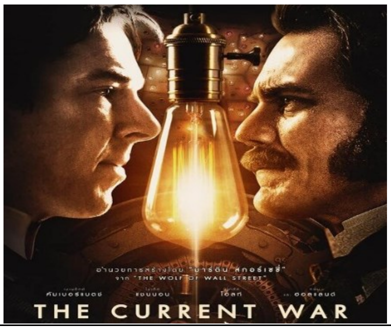
Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri [Qualitative research methods in the social sciences].* Seçkin.

Yin, R.K. (2009). *Case study research: Design and methods* (4th ed.). Sage.

Yu, K.C., Fan, S.C., & Lin, K.Y. (2015). Enhancing students' problem-solving skills through context-based learning. *International Journal of Science and Mathematics Education*, *13*(6), 1377-1401. https://doi.org/10.1007/s10763-014-9567-4

APPENDIX

The Current War is a 2017 American historical drama film inspired by the 19th century rivalry between Thomas Edison and George Westinghouse over the power distribution system in the United States. Directed by Alfonso Gomez-Rejon, the film was released in the United States on October 25, 2019. The film received generally mixed reviews, with praise for the actors' performances and engaging story, but criticism of the general execution. Take note of the following concepts while watching the movie "The Current War."



Concepts	Events in the movie	Time
Concepts		
Electrical voltage		
Electrical current		
Resistance		
Brightness		
Electrical energy		

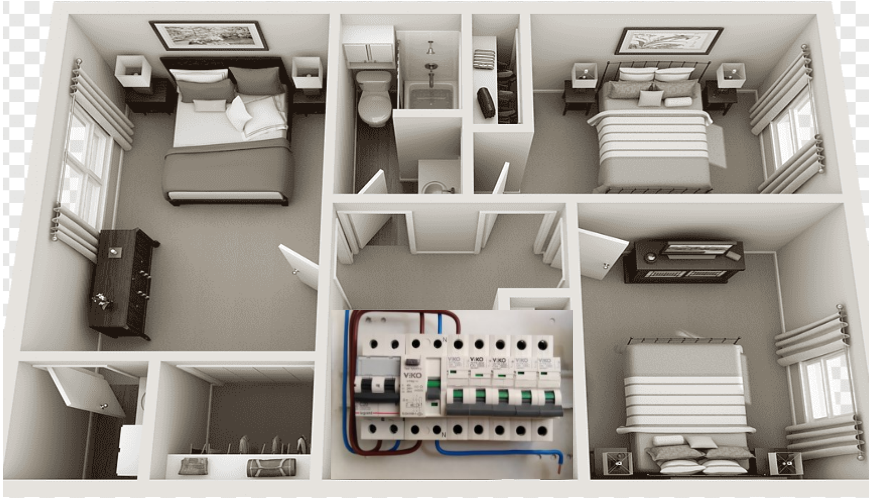
After watching the movie and completing the sections in the table, fill in the following peer evaluation form by giving points in the range of (1-5), taking into account your interaction as the group members in the process of filling out the joint group form. “1” is the lowest level, “5” is the highest level.

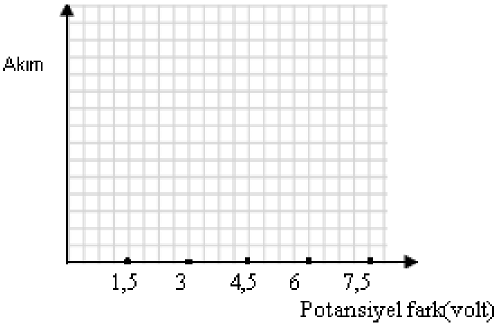
Criteria	1.My friend	2.My friend	3.My friend	4.My friend	5.My friend	According to me I
Participates in studies voluntarily						
Shares what he knows with his friends						
Helps friends when needed						
Gathers information from different sources						
Respects the opinions of his group mates						
Duty responsibility is at a high level						
Likes to work together						
Contribution to the formation of the group idea						

Evaluate other groups based on group presentations and discussions.

Criteria	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Presentation						
Accuracy of information						
Collaboration of group members						
All group members fulfill individual responsibilities						
Interaction of group members						
Task sharing competence						
Persuasion competence						
Learning competence of group members						

Anıl, who graduated from the university and became an electrical engineer, bought a land and built a small house on this land where he could rest with his family on weekends. Draw the project of the electrical lines and make the experimental application on the electrical circuit by drawing the project that can be done according to the principle of not affecting the other parts when there is a deterioration in one part of the house related to the lighting.





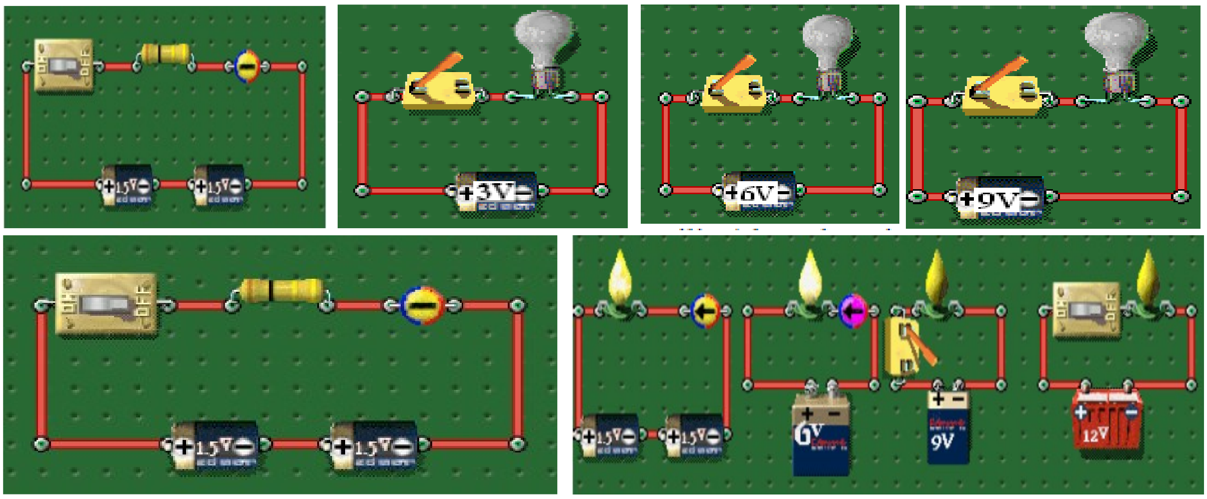
Ölçüm No	Potansiyel Farkı(volt)	Akım (amper)	Potansiyel farkı/akım
1	1.5		
2	3		
3	4.5		
4	6		

Are current and potential difference the same thing?

Some of the students studying in high schools think that the concepts of current intensity and potential difference are the same concepts, however, this is wrong information.

The phenomenon of moving +1-unit charge in conductive materials in a unit distance in the electric field with the help of electrical forces is called potential difference. The movement of electrons in the conductor by the half of the electric field strength is called electric current intensity. In this case, it turns out that electric potential difference and electric current intensities are different concepts.

Draw the electrical wiring of the house project using the Edmark simulation program.



A farmer living with his wife in the village makes a living by allocating five decares of land to himself and five decares of land to his wife. For each born child, he buys five acres of land from where he lives and increases his land gradually. After a certain period of time, something has caught the attention of the farmer, whose land has increased considerably. Although it increases its land so much, the total product increases, but the amount of product falling on itself does not change.

Similar feature	Compare	Simulated feature
Amount of Land	Comparable	Electrical voltage
Total Product	Comparable	Electrical current
Product amount per person	Comparable	Electrical voltage/Current
Farmers	Incomparable	Generator

Write a movie scenario in which the basic concepts of electricity are used practically, together with your group mates.

With your friends, shoot the movie that you have determined the scenario of as a short film so that all the group members will take an active role. Write the events in the movie in the table below.

Concepts	Events in the movie	Section time
Electrical voltage		
Current intensity		
Resistance		
Luminescence		
Electrical energy		

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 218–235

https://doi.org/10.21449/ijate.1089472

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Identifying the presence of context and item-writing flaws in practice items: The case of Turkish mathematics textbooks

Munevver Ilgun Dibek[1,*], Zerrin Toker[2]

[1]TED University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]TED University, Faculty of Education, Department of Mathematics and Science Education, Ankara, Türkiye

ARTICLE HISTORY

Received: Mar. 17, 2022
Revised: Oct. 14, 2022
Accepted: Nov. 04, 2022

Keywords:
Context,
Item writing rules,
Mathematics textbooks.

Abstract: This study seeks to ascertain the degree to which context-based items are offered in Turkish mathematics textbooks as well as the quality of the items in terms of item writing guidelines, whether or not they are given as traditional or context-based. A qualitative research approach is used in this study. The eighth-grade mathematics textbook used in public schools and a textbook used in certain private school chains constitute its sample. The practice items (i.e, exercises without solutions given) included in the textbooks were analyzed by performing document analysis. The results revealed that both textbooks contain several flawed items in terms of item writing rules, as well as having mainly non-contextual items.

1. INTRODUCTION

Ensuring that students gain the requisite knowledge and skills to satisfy the demands and expectations of contemporary society is one of the goals of education in schools. To what extent people are citizens who have gained the knowledge and skills required for both personal and social life depends on their level of mathematics literacy (Geiger *et al*., 2015). It is crucial to foster mathematical literacy in the mathematics classroom to attain the ultimate goal of education (Bolstad, 2020). The term "mathematics literacy," which is one of the competencies assessed in the Programme for International Student Assessment (PISA), is defined as follows: (i) understanding and defining the role of mathematics in real life; (ii) making decisions based on mathematics in constructive, associative, and reflective ways in life; and (iii) making it a lifestyle (OECD, 2009).

A strategy to strengthen students' mathematics literacy is to use situations from life outside of school, considering the mathematical needs of current living. According to Kaiser and Willander (2005), students should be given questions that incorporate real-world scenarios where mathematical models can be employed to increase their mathematical literacy; thus, they can formulate the issue, create a model, and mathematically assess their findings. Goos *et al*. (2012) suggested a model (see Figure 1) to describe the complicated nature of mathematics

*Corresponding Author: Munevver Ilgun Dibek ✉ munevver.ilgun@tedu.edu.tr TED University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

literacy in general and numeracy in particular. They claimed that mathematics literacy is a broad interpretation of numeracy.

Figure 1. *A model for mathematics literacy (Goos et al., 2012, p. 149).*

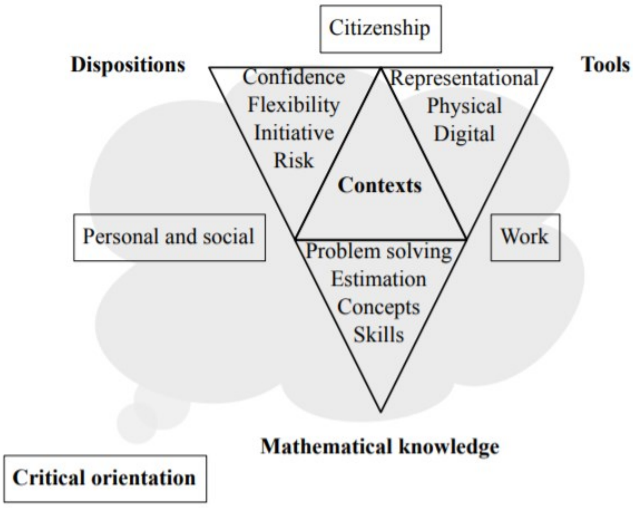


Figure 1 illustrates how literacy in many contexts is necessary for mathematical literacy. It is important to incorporate context into mathematics instruction and use context-based questions to increase students' mathematical literacy. Despite the importance of context in the development of students' mathematical literacy skills, students in diverse countries face difficulties when it comes to correctly answering these questions (Schwarzkopf, 2007; Verschaffel *et al*., 2000). A country where a majority of students had trouble responding to questions with context is Turkey. For example, the same problem is observed when the results regarding the mathematics literacy tests administered in different PISA cycles are examined. Although the average mathematics literacy scores of Turkish students increased slightly in each implementation year, the increase was not sufficient to exceed the OECD average. Specifically, in PISA 2018, Turkish students' mean score regarding mathematics literacy is 454 although the OECD average is 489. Similarly, although this increase brought Turkey to the forefront in some PISA cycles, it did not yield in great changes in its ranking in general (MoNE, 2019).

When given context-based questions or activities, students frequently struggle to discriminate between relevant and irrelevant information in the question as well as comprehend the nature of the problem and define the requisite steps to solve them (OECD, 2019a). Context-based questions are difficult for students to answer because they are solely used in evaluation procedures and do not have a place in the teaching process (Başaran, 2005; Fidan, 2018). In addition to the context, the role of the quality of items regarding compliance with item writing rules in teaching and learning is considered important. The high number of multiple-choice items in the books, where the writing process of both the item stem and the plausible distractors is difficult (Shin *et al*., 2019), necessitates revision of some items in Turkish textbooks according to the item writing principles (Kul *et al*., 2018; Simsek, 2016).

In the light of these issues, this study deals with revealing the situation regarding the extent to which students encounter such questions in the textbooks and the quality of the questions in the textbooks as one of the possible reasons for the difficulties experienced by Turkish students while solving context-based tasks and traditional items. Therefore, we analyzed practice items in textbooks in terms of context and quality regarding compliance with item writing rules. The approach adopted in the study is to examine what Turkish eighth grade mathematics textbooks offered to Turkish students regarding solving context-based items. Despite being conducted in a Turkish context, the study has the potential to contribute to the international literature by offering details on the connections between students' learning and a number of aspects of the textbook's practice items.

2. THEORETICAL BACKGROUND AND RESEARCH QUESTION

2.1. The Role of Context in Teaching and Learning

Context can be considered as real-world settings, imaginary situations, or the formal world of mathematics (Van den Heuvel-Panhuizen, 2005). The realization of learning depends on the use of context-based questions. Learning can occur effectively when students can relate to an idea and its applications to their own culture, family, friends, or their daily lives (Yam, 2005). At some point during the learning process, every student wonders, "Why do we need to learn this?" However, very few students are able to provide appropriate responses to the questions that arise when they attempt to make sense of what they are being required to perform (Krouse, 2016). Therefore, using contexts for the development of mathematical thinking contributes to an understanding of mathematical concepts and prevents or eliminates misconceptions by improving the students’ ability to use mathematics in various contexts of daily life. The use of daily life contexts as a didactic tool to support learning provides a meaningful basis for the concepts in the mathematics curriculum.

2.2. The Role of Quality of Items Regarding Compliance with Item Writing Rules in Teaching and Learning

Regardless of the psychological construct that is being tested, the method of creating the test items that will be used to measure it is crucial because the test items make up the structure of mental properties. Test items must therefore be described succinctly and clearly (Osterlind, 2002). Moreover, the high level of validity and reliability of the results obtained from the test depends on the quality of the items that make up the test. Specifically, the item must reflect the structure or content to be measured for the results from the test to provide valid interpretation (Peeters *et al*., 2013). If a strong link is not established between the test item and the psychological construct to be measured or its content, the item will lose its purpose and will not be different from a thought that circulates freely on a test page (Osterlind, 2002). Besides, the difficulty level of the test item increases due to item-writing flaws. In other words, construct-irrelevant variance is introduced to the results obtained from the test item; therefore, the reliability of the results to be obtained decreases (Downing, 2005). Hence, the interpretability of test results is closely associated with the quality of the item.

Certain technical features should be considered to ensure the high quality of the test item. For example, the use of the correct item format, level of complexity of the words used, use of a sufficient number of answer options, and absence of negative words are a few of these features. Every word is valuable in a test item. The test-taker should be able to understand the meaning of the item's stem and recognize the incorrect choices/distractors from the correct one (Osterlind, 2002). The way the items are built is crucial for the students, the researchers who will use the assessment results, and the program evaluators since the test items serve as the fundamental building blocks of the assessment tools. In this context, examining the existence of item-writing flaws in the textbooks to be used in the teaching and learning process will provide valuable information.

2.3. Context of the Study and the Case of Turkey

The teaching process has not regularly used context-based questions until now because textbooks do not contain enough context-based questions, and instructors could lack expertise in this subject and feel unqualified (Kayhan Altay *et al*., 2020). For instance, in a study conducted by Fidan (2018) teachers said that assessment questions from textbooks and context-based questions are incompatible. Similar results were found in the study of Kayhan Altay *et al*. (2020) who focus on the context and daily life in sixth-grade mathematics textbook. In studies focusing on a particular section of the textbooks, it is stressed that items presented directly, rather than through a mathematization context, come to the fore (e.g, Kar & Işık, 2015). In parallel with teachers' opinions, since statewide exams play a significant role in students' life, students state that they desire tests that reflect what is expected of them in their textbooks and lectures (Başaran, 2005). In Turkey, when compared with previous administrations, the recent statewide exam called the High School Entrance Exam (LGS in Turkish) includes many context-based questions (Güler & Ülger, 2018). However, recent research indicates that teachers have complained that the exam is incompatible with educational materials like textbooks (e.g., Korkmaz *et al*., 2020).

Students need to be familiar with tasks involving contexts within the teaching and learning process for them to be successful in answering such questions. Textbooks, that play an important role in the planning of teaching, are expected to include such tasks (Korkmaz *et al*., 2020). Given the strong relationship between textbooks and educational processes, it is crucial to understand how much opportunity for activities, items, and other contents—including context—are provided by textbooks, which support educational processes. To the best of our knowledge, while numerous studies (e.g., Hadar, 2017; Törnroos, 2005; Wijaya *et al*., 2015) have examined mathematics textbooks in relation to learning opportunities and students' mathematics achievement, no study has looked at the items in mathematics textbooks in two dimensions, such as context and the quality of item regarding compliance with item writing rules. Examining to what extent and how such questions are included in the textbooks currently in use will contribute to understanding the difficulties experienced by students.

2.4. Research Purpose

This study attempts to investigate the practice items in Turkish eighth grade mathematics textbooks, which are utilized as main course materials by teachers, in terms of context and their compliance with the item writing principles. Within this context, this study seeks answers to the following research questions:

(1) To what extent do Turkish eighth grade mathematics textbooks offer context-based practice items?

(2) What are the item-writing flaws of practice items in Turkish eighth grade mathematics textbooks?

3. METHOD

3.1. Research Design

The present study aims to examine several aspects of the practice items included in Turkish eighth grade mathematics textbooks, such as context and quality regarding item writing rules. In this regard, a document analysis is used in this study. It is a systematic procedure for reviewing or evaluating documents including text and images that the researcher did not interfere with (Bowen, 2009).

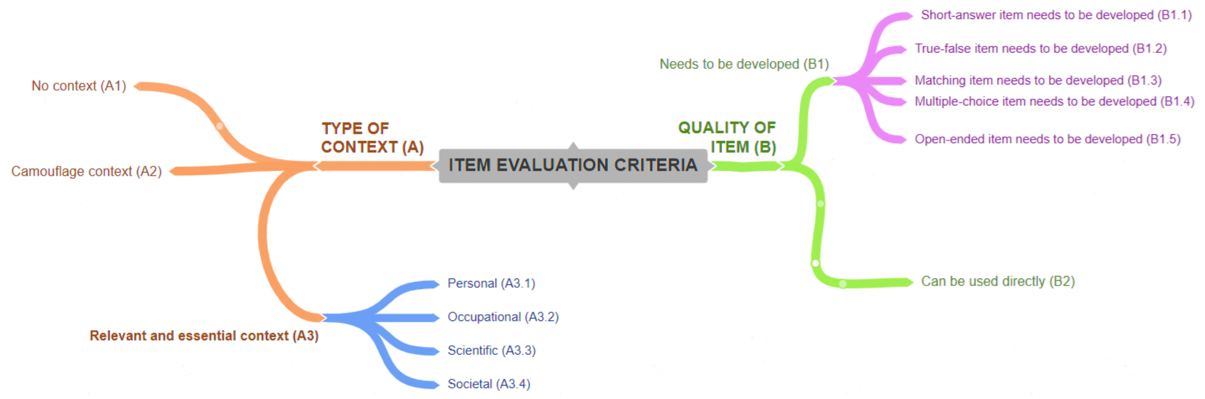
3.2. Sample

The eighth-grade mathematics textbooks, used by public schools and one of the private school chains, constitute the sample of the study. These two textbooks (hereafter referred to as Book 1 and Book 2) have been selected using the purposive sampling method that enables researchers to select their sample according to predefined criteria (Fraenkel *et al*., 2012). The reason for choosing Book 1, approved by MoNE, is that this book is used as the principal course resource in all schools, while Book 2 was chosen to increase the representativeness of the eighth grade mathematics textbooks used in private schools. More precisely, a private school, where the number of students in the 8th grade level is higher than other private schools, uses Book 2. Moreover, some of the other private schools use Book 2 as a supplementary material in the mathematics course at the 8th grade level. Eight grade level students were chosen because, according to different PISA cycles that are pioneering applications where mostly context-based items are used, grade 8 can be considered a relevant grade level to prepare students to be able to solve context-based tasks (Wijaya *et al*., 2015). Also, 8th grade Turkish students attend centralized exams, indicating the tendency to include context-based items.

3.3. Data Collection and Analysis

Data collection and analysis were performed by using a two-dimensional framework given in Figure 2.

Figure 2. *Two-dimensional framework.*


ITEM EVALUATION CRITERIA
TYPE OF CONTEXT (A)
No context (A1)
Camouflage context (A2)
Relevant and essential context (A3)
Personal (A3.1)
Occupational (A3.2)
Scientific (A3.3)
Societal (A3.4)
QUALITY OF ITEM (B)
Needs to be developed (B1)
Short-answer item needs to be developed (B1.1)
True-false item needs to be developed (B1.2)
Matching item needs to be developed (B1.3)
Multiple-choice item needs to be developed (B1.4)
Open-ended item needs to be developed (B1.5)
Can be used directly (B2)

Context Analysis Form (CAF) was utilized to provide an answer to the first research question regarding the context of the practice items. In addition, Checklists for Evaluating Item Quality (CEIQs) were employed to address the second research question about item-writing flaws in practice items. By using these tools, data collection and analysis were performed simultaneously.

3.3.1. *Context analysis form (CAF)*

One of the tools utilized to collect data for the study was the CAF, which was used to assess the context-related aspects of the items from the textbooks that were the subject of the current investigation. The Wijaya *et al*. (2015) classification, which is more appropriate for real-world circumstances and 21st century abilities, is the basis for the subcategories of the CAF. They were coded as no context (A1), camouflage context (A2), and relevant and essential context (A3). The framework of PISA (OECD, 2019b) was taken as the basis for the items that were determined to be context-based. Accordingly, personal, occupational, societal and scientific contexts are coded as A3.1, A3.2, A3.3, and A3.4, respectively. Explanations related to each code and category of CAF are provided in Appendix 1.

3.3.2. *Checklists for evaluating item quality (CEIQs)*

CEIQs were employed as additional data collecting tools to get information regarding the second dimension depicted in the framework shown in Figure 2. More specifically, different Checklists for Evaluating Item Quality (CEIQs) (see Appendix 2) were created by considering the recommendations made by Miller *et al*. (2013) to determine the quality related conformity with item writing rules of the items. These checklists were used to determine whether the item violates any item writing guidelines and, if so, what specific violations it may contain. At this point, it has been decided whether the item will be used directly, based on the rules in the relevant checklist, depending on the type of item (open ended, multiple choice, true-false, etc). The item that does not have the item-writing flaws indicated in the relevant checklist is defined as “the item that can be used directly” by giving the B2 code.

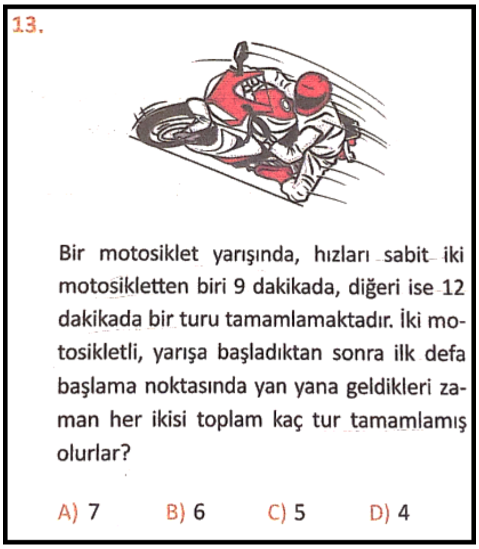
3.3.3. *Training of the coders*

Two researchers, one with expertise in measurement and evaluation and the other in mathematics education, carried out the process of assessing the practice questions in the chosen books in terms of several dimensions and assigning codes to them in this study. These coders have knowledge of the dimensions addressed by the current study. Specifically, throughout their doctoral studies, these researchers completed a number of graduate courses about test items and item structures. They currently teach undergraduate-level courses focusing on these topics and various taxonomies for the classification of learning outcomes. A third coder—a pre-service teacher—was brought in when the two researchers could not agree on a particular item. It was ensured that the chosen preservice teacher had sufficient understanding of the various item types and how to write an appropriate item in accordance with item-writing guidelines. The pre-service teacher received in-depth instruction from the two researchers on the aspects of analysis and coding of sample items prior to the use of coding. Each category in the data gathering tools was examined individually during this training, and what was meant to be communicated was discussed. Consequently, it was ensured that each coder assigned identical meanings to each code.

3.3.4. *Coding procedure*

Each item received a location code in the book and a dimension code that covered the two dimensions of the analysis demonstrated in Figure 2 throughout the coding process. For example, the dimension of "context" received the letter "A," whereas the dimension of "item quality" received the letter "B." Additionally, sub-codes were given to the items to precisely specify the subcategory they belong to. Figure 2, that demonstrates the framework used in the present study's coding process, contains more details. Moreover, in cases where a poor item in the textbook exists, possible item-writing flaws that an item possesses have been previously listed in CEIQs, and codes have been assigned to each item-writing flaw to indicate the kind of item-writing flaw an item possesses. Since each rule included in each CEIQ is stated as a question, the negative answer to each question considered that the particular item has an item-writing flaw, and was coded as B1. Further code was assigned to specify which item-writing flaw an item had. On the other hand, items with the positive answer to these questions are coded as B2. More than one code was given to the item when the item included more than one problem regarding item quality. The number of items in each category was counted at the end. A sample of codes given to an item is shown in Figure 3.

Figure 3. *The sample of codes given to an item.*



13.

Bir motosiklet yarışında, hızları sabit iki motosikletten biri 9 dakikada, diğeri ise 12 dakikada bir turu tamamlamaktadır. İki motosikletli, yarışa başladıktan sonra ilk defa başlama noktasında yan yana geldikleri zaman her ikisi toplam kaç tur tamamlamış olurlar?

A) 7 B) 6 C) 5 D) 4

This question asks how many laps both motorcycle racers completed when they first came together at the starting point after completing a lap in 9 minutes and 12 minutes respectively. The location code assigned to the item shown in Figure 3 was "B2, U1, P38, I13," and the dimension code for this item was "A2, B1.4.2." First of all, if we interpret the location code, this is a question from the book used in certain private schools that we consider as Book 2 (B2). Moreover, this is the thirteenth item (I13) on page 38 (P38) in unit 1 (U1). To continue with the dimension code, the camouflage context (A2) is used in this item. In addition, the item has a problem in terms of being a qualified item. The negative answer to rule B1.4.2 (see Appendix 2) indicated that the stem of the item presents an unnecessary element. More specifically, the use of the image in this item is not necessary to solve the problem.

In addition to the authors of the current study, one measurement and evaluation specialist and one mathematics educator were consulted regarding the dimensions and definitions in the data collection tools developed or adapted in order to establish the validity of the results obtained from various data collection tools, such as CAF and CEIQs. In response to their suggestions, the data collection tools were changed. Additionally, the coders conducted pilot coding using all of the data collection tools before the researchers coded every task in the two textbooks to ensure that they comprehended each criterion in the same manner. This strengthened the validity of the conclusions drawn from the measurement results. Within this context, as in similar studies (e.g., Wiijaya *et al*., 2015), 15% of the items included in each textbook selected within the scope of the research were coded independently by all the coders. Items to be coded by all the coders were randomly selected. Interrater reliability was calculated for each dimension of the analysis to determine the reliability of the results obtained from this coding procedure. For this, "the agreement percentage formula" developed by Miles and Huberman (1994, p. 64) was used. Accordingly, the formula is as follows:

$$\text{Agreement percentage} = \frac{(\text{the number of agreement})}{(\text{the number of agreement} + \text{the number of disagreement})} \times 100$$

The scorer agreement coefficients of context dimension and item quality during the pilot coding procedure were found to be .95 and .90, respectively. The raters' coding is consistent because the coefficient is larger than .90 (Miles & Huberman, 1994). The items that the coders were not in agreement about were returned and the coding for them was repeated until agreement was achieved. The frequency and percentage values for the number of items gathered under each dimension were then presented following the final item coding.

4. FINDINGS

4.1. Context Dimension

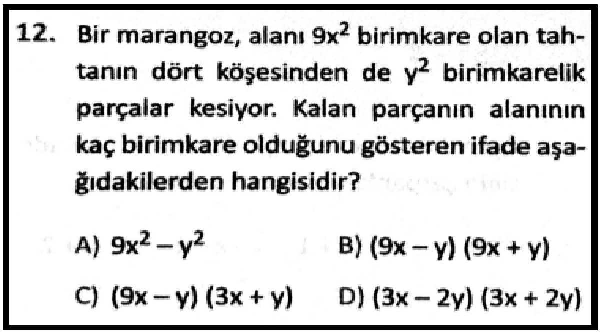
Table 1 displays the findings of the analysis of the context dimension of the items found in Books 1 and 2.

Table 1. *Results of the analyses of items in terms of context dimension.*

		Book 1											Book 2	
Context		Multiple-Choice		Open-Ended		Short Answer		Matching		T-F Items			Multiple-Choice	
		f	%	*f*	%	*f*	%	*f*	%	*f*	%	Total	*f*	%
A1		131	74.70	9	81.80	16	76.20	5	83.30	22	78.60	183	464	80.00
A2		35	16.30	-	-	2	9.50	1	16.70	4	14.30	42	55	9.50
A3	A3.1	4	2.20	-	-	1	4.80	-	-	2	7.10	7	21	3.60
	A3.2	6	4.50	1	9.10	1	4.80	-	-	-	-	8	6	1.00
	A3.3	-	-	-	-	-	-	-	-	-	-	-	13	2.20
	A3.4	2	2.20	1	9.10	1	4.80	-	-	-	-	4	21	3.60
Total		178	100	11	100	21	100	6	100	28	100	244	580	100

Table 1 demonstrates that when the context of the items in the two books is compared, the majority of the items in Book 1 ($f = 183$) do not contain any contextual components, while some of them use camouflage context ($f = 42$), and a small number of items include occupational, personal, and scientific context elements. A similar pattern is noticed for Book 2. A few items ($f = 55$) in Book 2 have camouflage context, while the majority of items ($f = 464$) do not use context. Similar to Book 1, Book 2 has a small number of items ($f = 61$) with relevant and essential contexts. It was also discovered that there are more items with personal and scientific context than any occupational or societal context. Figures 4 and 5 show examples of contexts related to camouflage and personal context, respectively.

Figure 4. *Example for camouflage context (coded as A2).*



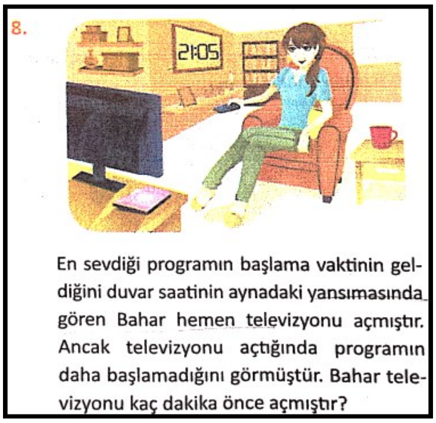
12. Bir marangoz, alanı $9x^2$ birimkare olan tahtanın dört köşesinden de y^2 birimkarelik parçalar kesiyor. Kalan parçanın alanının kaç birimkare olduğunu gösteren ifade aşağıdakilerden hangisidir?

A) $9x^2 - y^2$ B) $(9x - y)(9x + y)$

C) $(9x - y)(3x + y)$ D) $(3x - 2y)(3x + 2y)$

In the item displayed in Figure 4, it is stated that a carpenter cuts sections of y^2 square units from each of the board's four corners, with an area of 9x2. The students are then asked to calculate the area of the piece that is left over. Due to the fact that it does not just refer to mathematical objects, symbols, or structures, this item is classified as having "camouflage context." On the other hand, the context is not necessary and the operations needed to solve the problems are already obvious; the answer is simply obtained by adding the numbers provided in the item.

Figure 5. *Example of personal context falling under the category of relevant and essential context (coded as A3.1).*



8. En sevdiği programın başlama vaktinin geldiğini duvar saatinin aynadaki yansımasında gören Bahar hemen televizyonu açmıştır. Ancak televizyonu açtığında programın daha başlamadığını görmüştür. Bahar televizyonu kaç dakika önce açmıştır?

In this item, it is said that a person who noticed her favorite television program through the reflection of a wall clock in a mirror realized that the program had not yet started when she turned on the television. The students were asked to calculate how many minutes earlier she might have turned on the television. Given that context is necessary to comprehend the issue and find a solution, this item is placed under the category of "relevant and essential context," and categorized under "personal" because the context relates to personal life.

4.2. Quality Dimension regarding Compliance with Rules of Item Writing

Table 2 displays the findings of the study of items from Books 1 and 2 in relation to the item quality dimension.

Table 2. *Analysis of items in terms of the quality dimension.*

Book		Book 1											Book 2	
Item Type		Multiple-choice		Open-ended		Short Answer Completion		Matching		T-F Items		Total	Multiple-choice	
		f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	*f*	%
Quality	B1	31	17.40	3	39.40	4	19.00	6	100	0	-	44	190	32.80
	B2	147	82.60	8	72.70	17	81.00	0	-	28	100	200	390	67.20

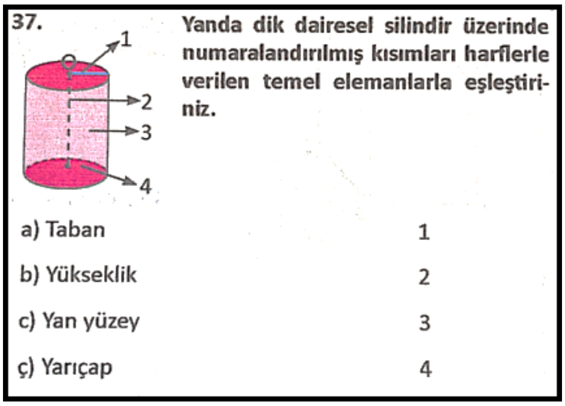
When the items selected within the scope of the research were examined in terms of their compliance with the principles of item writing and their quality, it was found that the number of items written by considering these principles (coded as B2) was higher than the number of items in which these principles were neglected (coded as B1). However, it should be noted that there are still a sizable number of items created that do not consider these principles. The various types of items, such as short answers, matching, and open-ended questions, are provided in Table 3 to highlight common item-writing flaws in Book 1, as Book 2 only contains multiple-choice questions.

Table 3. *Frequently-made mistakes in writing qualified items included in Book 1.*

Item Type	Criteria List	Total Number of Items Reviewed	*f*	*(%)*
Short-answer	Do the answers blank place at the end of the items?	21	2	9.50
	Do the items contain any clues?		2	9.50
Matching	Do the responses rank alphabetically or numerically?	6	2	33.30
	Do the directions specify the number of times each response may be used?		4	66.70
Open-ended	Does the material to be interpreted contain some novelty to require interpretation?	11	1	9.10
	Does each question specify the expected response?		2	18.20

Table 3 shows that out of the 21 short-answer questions, 4 contained some specific item-writing flaws. The short answer questions in Book 1 specifically had item-writing flaws in that the blanks were not at the end of the items and some clues would reveal the solution within the item. All matching type items found in Book 1 have errors according to item-writing rules. Common item-writing flaws observed in the matching type items are that there is no information on how many times the expressions/numbers can be used in the response column, and that these expressions/numbers are not in alphabetical or numerical order. Among 11 open-ended items, the neglected item writing principle is that the expected answer from the student should be made explicit. An example of an item-writing flaw in a matching question is provided in Figure 6.

Figure 6. *Example of a mistake in a writing matching item (coded as B1.3.4 and B 1.3.6).*



37. Yanda dik dairesel silindir üzerinde numaralandırılmış kısımları harflerle verilen temel elemanlarla eşleştiriniz.

a) Taban	1
b) Yükseklik	2
c) Yan yüzey	3
ç) Yarıçap	4

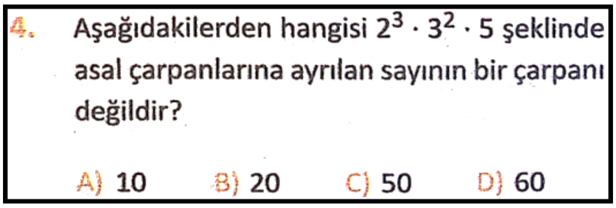
The students were required to match the numbers of a right circular cylinder with the cylinder's fundamental components in this item. This item's quality was judged to contain two item-writing flaws. The first one is that each response is not allowed to be used more than once in the item's instructions. Another is that the total number of premises in the premise column and the total number of responses in the response column are the same. However, more statements in the response column are needed. Otherwise, even though they are unfamiliar with the concept needed to match the final premise and response, the students are still able to accomplish it. Table 4 lists the common item-writing flaws that were made when creating multiple-choice items for the two books analyzed within the scope of the research.

Table 4. *Frequently-made mistakes in the creation of multiple-choice items.*

Criteria List for Multiple-Choice Item	Book 1		Book 2	
	f	%	*f*	%
Is each item stem meaningful?	6	3.40	7	1.20
Do the item stems contain irrelevant material?	1	.60	9	1.60
If used, has negative wording been given special emphasis (for example, capitalization)?	-	-	75	12.90
Is there grammatical consistency between the alternatives and the item stem?	-	-	1	.20
Are the alternative answers brief and free of unnecessary words?	-	-	5	.90
Are the length and form of the alternatives similar?	4	2.20	15	2.60
Are the distractors plausible to low achievers?	10	5.60	21	3.60
Do the items contain any verbal clues to the answer?	-	-	2	.30
Do the verbal alternatives rank alphabetically?	1	.60	-	-
Do the numerical alternatives rank numerically?	9	5.10	55	9.50

Table 4 demonstrates that failing to make the distractors plausible enough was the most frequently observed item-writing flaw in Book 1. Another typical one was that the response options that were numerical were not presented in a sequential order. On the other hand, failing to emphasize the negative statements at the stem of the multiple-choice questions was the item-writing flaw that was observed frequently in Book 2. The distractors were not written in a numerical order, another common item-writing flaw in the items in this book, similar to the case in Book 1's items. Figures 7 and 8 illustrate examples of item-writing flaws observed when creating multiple-choice questions.

Figure 7. *Example of a mistake in writing multiple-choice item (coded as B1.4.3).*

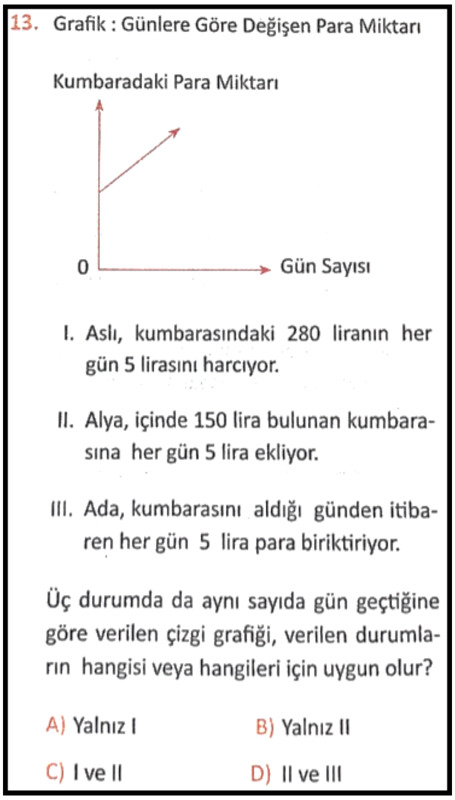


4. Aşağıdakilerden hangisi $2^3 \cdot 3^2 \cdot 5$ şeklinde asal çarpanlarına ayrılan sayının bir çarpanı değildir?

A) 10 B) 20 C) 50 D) 60

“Which of the numbers provided in the response options is the prime factorized number?” is the question in this item. This item violates the rules for item writing because the negative phrase "it is not" (in Turkish, "değildir") was not highlighted or stressed. Another common item-writing flaw in the item writing approach is demonstrated in Figure 8 for another item.

Figure 8. *Example of an Item-Writing Flaw in the Multiple-Choice Item (coded as B1.4.7).*



13. Grafik : Günlere Göre Değişen Para Miktarı

Kumbaradaki Para Miktarı
0
Gün Sayısı

I. Aslı, kumbarasındaki 280 liranın her gün 5 lirasını harcıyor.

II. Alya, içinde 150 lira bulunan kumbarasına her gün 5 lira ekliyor.

III. Ada, kumbarasını aldığı günden itibaren her gün 5 lira para biriktiriyor.

Üç durumda da aynı sayıda gün geçtiğine göre verilen çizgi grafiği, verilen durumların hangisi veya hangileri için uygun olur?

A) Yalnız I B) Yalnız II

C) I ve II D) II ve III

In the item displayed in Figure 8, it is asked for which of the cases listed in I, II and III can the graph showing the amount of money in the penny bank and the number of days be created. In Case I, it is stated that Aslı spends 5 liras every day of her 280 liras in the penny bank. In Case II, it is indicated that Alya adds 5 liras every day to her penny bank, which currently holds 150 liras. In the last case, Case III, it is stated that Ada has saved 5 liras every day since the day she received her penny bank. The distractors of this item are not plausible enough, because a student who knows that this graph cannot be drawn for Case I can directly rule out options A and C.

5. DISCUSSION and CONCLUSION

The present study demonstrates that Turkish eighth-grade mathematics textbooks rarely include context-based items. Most of the items in these textbooks is non-contextual and does not require mathematization. In other words, this study shows that the items in the eighth-grade mathematics textbooks, commonly-used in Turkey, are insufficient in making connections to real-world situations in terms of personal, scientific, occupational, and social aspects. The results of two studies—one by Kayhan Altay *et al.* (2020), that investigated the contexts used for real-life connections in mathematics textbooks for sixth graders and found that more than half of the tasks presented in the textbook are not related to real life, and another by Kar & Işık (2015), that examined Turkish mathematics textbooks in a more specific area, concentrating on addition and subtraction operations with integers—support this conclusion. This situation with Turkish textbooks is also observed in the textbooks of a few other countries that fall behind OECD average like Turkey in PISA, where context plays an important role in the measurement of literacy. For example, Indonesia shows similar patterns in terms of mathematics literacy performance in PISA 2018 (OECD, 2019a) and mathematics textbooks. It appears that the results of the current study are consistent with those of Wijaya *et al*. (2015), who looked at the learning opportunities provided by Indonesian textbooks for completing context-based mathematical activities. One reason for this situation might be that sufficient information about curriculum change must be given for the existing curriculum framework to be implemented, (Rea-Dickson & Germanie, 2001).

The results of this study show that multiple-choice items make up the majority of the material covered in Turkish textbooks. This result is in line with the results of the study conducted by Kul *et al*., (2018) which analyzed the item types in Turkish and Canadian textbooks and discovered that multiple-choice items made up a higher percentage of the items in Turkish textbooks. Multiple-choice items are more prevalent than other item types in the eighth-grade mathematics textbooks used in Turkey, which may be explained by the fact that these types of items also appear in the middle to high school transition exam. The 8th grade level, the level covered by the mathematics textbook under investigation in this study, is the stage between secondary school and high school. Students take a centralized test at this transitional level. They are exposed to questions that are similar to the item types in this exam during the learning and teaching process to succeed in this high stakes exam. In other words, this exam system, where significant decisions are made depending on the results, also impacts the teaching process (Kahraman, 2014). Consequently, the course textbooks now contain more multiple-choice questions.

When the frequencies of the item-writing flaws in multiple choice were compared for both books, it was concluded that Book 1 had fewer item-writing flaws than Book 2. Since Book 1 was approved by MoNE, both field experts and assessment and evaluation experts took part in the item writing process in Book 1. Therefore, the item writing process could have been conducted more meticulously, and the relevant item redactions could have been made. Accordingly, there may have been a decrease in the number of item-writing flaws related to multiple-choice items. Additionally, in terms of the type of the most commonly observed item-writing flaws in constructing the multiple-choice items in eighth-grade textbooks addressed in the present study, this study shows that negative statements in the stem of the item are not emphasized, and plausible distractors are not developed. The learner might not pay attention if the negative term at the stem of the multiple choice item is not highlighted. Even though the student is aware of the right answer, they may still respond incorrectly since they failed to notice the negative word. However, the primary goal of multiple-choice questions is to discover whether students have acquired the idea being measured, not to gauge how attentive they are (Chiavaroli, 2017). Additionally, asking students to identify the incorrect options is not a preferred method in teaching. Just because someone is aware of the incorrect options does not imply that they are also aware of the solution (Burton *et al*., 1991).

This study's conclusion is consistent with the results of the study conducted by Simsek (2016), who compared the items created by teachers and trainers and found that almost 60% of them need improvement. The two most frequently observed item-writing flaws were the use of implausible distractors and the use of negative items without emphasizing the negative features of the items. The use of distractors like this (i.e., using implausible distractors) causes the question with more response options to function as an item with fewer response options, increasing the possibility of getting the right answer just by chance (Royal & Stockdale, 2017); even if the students do not know the answer to the question, it causes them to eliminate distractors without prompting them to think and directly turn to the right answer. This reduces the item's ability to discriminate (Rush *et al*., 2016). In other words, the item will no longer be sufficient to distinguish between students who met the required learning goals and those who did not. Since creating plausible distractors and producing a high-quality multiple-choice test item stem are challenging tasks that require time and expertise, it may be understandable to use a lot of multiple-choice items with problematic distractors (Shin *et al*., 2019). For instance, in a test with 100 multiple-choice questions, each with five response alternatives, 400 distractors should be prepared along with 100 item stems and 100 right answers (Gierl *et al*., 2017). So even if it is not ideal, it is fairly obvious that writing illogical distractors is an often made blunder.

The findings of this study offer important insights into how context-based textbooks are currently written, as well as an understanding of the qualities of good context-based items to educational politicians who direct item writers and textbooks writers. Consequently, this study might be able to provide information that can be used in textbook preparation. More specifically, it is suggested that, in light of the findings of the present study, mathematics textbooks should include more context-based materials and students should be required to employ mathematization for these questions. To put it another way, more relevant and essential contexts should be used in the eighth grade mathematics textbooks. Additionally, whether in traditional form or a context-based form, the items' quality in terms of conformity with item writing rules should take precedence. The item cannot assess the material in a valid and reliable manner if the item writing principles are ignored.

The study, even if its primary focus is on the analysis of the practice problems in the Turkish eighth grade mathematics textbooks, also has the potential to provide a framework for increasing practitioners' knowledge of selecting qualified items. Teachers can choose from the pre-existing items, make necessary modifications, or use the forms as a checklist to create new items using the present study's forms. Along with the quantitative and qualitative results of the study, the implementation process can therefore aid future practices.

When conclusions are drawn from the findings of the present study, the following limitations need to be considered, since they also point to future possible research trajectories. First, it is incorrect to just attribute the low achievement of Turkish students, particularly in large scale assessments like PISA, to the inadequateness of the textbooks used by that age group in Turkish schools. As previously mentioned, different teachers differentiate their instructions even while using the same items. The fact that the items in the books meet the criteria for the dimensions considered in the context of this study does not, therefore, ensure the quality of the instruction. Future research should look into how much teachers use these textbooks and particularly items in those books in their lessons. Second, this study is limited only to mathematics textbooks. As context-based items are also featured in other subjects on national central exams and in large-scale assessments, textbooks of other courses, such as Turkish, Science, and Social Studies, could also be analyzed in the framework of the criteria stated in this study. Third, because the age group for the test, eighth graders, is the only one included in this study, additional research may be conducted with students of other grade or age levels. Lastly, although this study reflects the situation regarding 8th grade mathematics textbooks in Turkey, its results may also be useful for the 8th grade students in other countries below the OECD average in terms of mathematics performance in large scale assessments. International comparative studies might be carried out by identifying and selecting the textbooks of such countries to generalize.

Acknowledgments

This research was supported by the TEDU IRF Project (T-20-B2010-90025).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Authors conducted the whole research process including conceptualization, research design, literature review, data collection and processing, interpretation, writing and critical review together.

Orcid

Munevver Ilgun Dibek https://orcid.org/0000-0002-7098-0118
Zerrin Toker https://orcid.org/0000-0001-9660-0403

REFERENCES

Başaran, S. (2005). *Diğer ülkelerde lise bitirme sınavları ve Türk eğitim sistemi için lise bitirme sınavı önerisi [High school leaving exams in other countries and high school leaving exam recommendation for the Turkish education system]*. MEB Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

Bolstad, O.H. (2020). Secondary teachers' operationalisation of mathematical literacy. European *Journal of Science and Mathematics Education, 8*(3), 115-135. https://doi.org/10.30935/scimath/9551

Bowen, G.A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Burton, S.J., Sudweeks, R.R., Merrill, P.F., & Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. Bringham Young University Testing Services and the Department of Instructional Science. http://testing.byu.edu/info/handbooks/betteritems.pdf

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research & Evaluation, 22*(3), 1-14. https://doi.org/10.7275/ca7y-mm27

Downing, S.M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences, 10,* 133-143. https://doi.org/10.1007/s10459-004-4019-5

Fidan, M. (2018). Ortaokul öğrencilerinin Türkçe ders kitaplarının tasarımına yönelik görüşlerinin analizi [Analysis of middle school students' views on the design of Turkish textbooks.]. *Bayterek International Journal of Academic Research,1*(2), 178–189.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw Hill.

Geiger, V., Goos, M., & Forgasz, H. (2015). A rich interpretation of numeracy for the 21st century: A survey of the state of the field. *ZDM Mathematics Education, 47*(4), 531–548. https://doi.org/10.1007/s11858-015-0708-1

Gierl, M.J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Goos, M., Geiger, V., & Dole, S. (2012). Auditing the numeracy demands of the middle years curriculum. *PNA, 6*(4), 147-158. https://doi.org/10.30827/pna.v6i4.6138

Güler, H.K. & Ülger, B. (2018). PISA, TIMSS ve TEOG sınavlarının temele aldığı öğrenme kuramları [Learning theories based on PISA, TIMSS and TEOG exams]. In S. Çepni (Ed.), *PISA ve TIMSS mantığını ve sorularını anlama* (ss.111-153). Pegem A Yayıncılık.

Hadar, L.L.(2017). Opportunities to learn: Mathematics textbooks and students'achievements, *Studies in Educational Evaluation, 55,* 153-166. http://dx.doi.org/10.1016/j.stueduc.2017.10.002

Kahraman, İ. (2014). Merkezi ortak sınav uygulamasının etkilerine ilişkin öğretmen görüşleri [The effect of common implementation that related to teachers' opinion]. *Tunceli Üniversitesi Sosyal Bilimler Dergisi, 2*(4), 53-74.

Kaiser, G., & Willander, T. (2005). Development of mathematical literacy: results of an empirical study. *Teaching Mathematics and Its Applications, 24*(2-3), 48-60. https://doi.org/10.1093/teamat/hri016

Kar, T. & Işık, C. (2015). Comparison of Turkish and American seventh grade mathematics textbooks in terms of addition and subtraction operations with integers. *Education and Science, 40*(177), 75-92. https://doi.org/10.15390/EB.2015.2897

Kayhan Altay, M., Kurt Erhan, G. & Batı, E. (2020). Contexts used for real life connections in mathematics textbook for 6th graders. *Elementary Education Online, 19*(1), 310-323. https://doi.org/10.17051/ilkonline.2020.656880

Korkmaz, E., Tutak, T., & İlhan, A. (2020). Ortaokul matematik ders kitaplarının matematik öğretmenleri tarafından değerlendirilmesi [Evaluation of secondary school mathematics textbooks by mathematics teachers]. *Avrupa Bilim ve Teknoloji Dergisi, 18,* 118-128. https://doi.org/10.31590/ejosat.667689

Krouse, S. (2016, Jan 8). Why do we need to learn this. *Medium*. https://medium.com/@stevekrouse/why-do-we-need-to-learn-this-3ba1d42bd08a.

Kul, Ü., Sevimli, E., & Aksu, Z. (2018). A comparison of mathematics questions in Turkish and Canadian school textbooks in terms of synthesized taxonomy. *Turkish Journal of Education, 7*(3), 136-155. https://doi.org/10.19128/turje.395162

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook.* Sage.

Miller, D.M., Linn, R.L, & Gronlund, N.E. (2013). *Measurement and assessment in teaching* (11th ed.). Pearson Education, Inc.

MoNE (Ministry of National Education) (2019). *PISA 2018 ulusal ön raporu [PISA 2018 Preliminary National Report].* Eğitim Analiz ve Değerlendirme Raporları Serisi,10.

OECD (2009). *Learning mathematics for life: A view perspective from PISA* OECD Publishing.

OECD (2019a). *PISA 2018 results (Volume I): What students know and can do.* PISA OECD Publishing. https://doi.org/10.1787/5f07c754-en

OECD, (2019b). PISA 2018 assessment and analytical framework. PISA OECD Publishing. https://doi.org/10.1787/b25efab8-en

Osterlind, S.J. (2002). What is constructing test items? In S. J. Osterlind (Ed.), *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (pp. 1–16), Springer. https://doi.org/10.1007/0-306-47535-9_1

Peeters, M.J., Beltyukova, S.A., & Martin, B.A. (2013). Educational testing and validity of conclusions in the scholarship of teaching and learning. *American Journal of Pharmaceutical Education, 77*(9), 1-9. https://doi.org/10.5688/ajpe779186

Rahimah, D. & Visnovska, J. (2021). Analysis of mathematics textbook use: An argument for combining horizontal, vertical, and contextual analyses. *Journal of Physics: Conference Series, 1731*, 1-5. https://doi.org/10.1088/1742-6596/1731/1/01204

Rea-Dickson P. & Germania, K. (2001). Evaluating curriculum change. In D. Hall & A. Hewimng (Eds.), *Innovation in English language teaching: A reader*. British Library Catalogue.

Royal, K.D. & Stockdale, M.R. (2017). The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *Journal of Advances in Medical Education & Professionalism, 5*(2), 84-89.

Rush, B.R., Rankin, D.C & White, B.J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education, 16*(250), 1-10. https://doi.org/10.1186/s12909-016-0773-3

Schwarzkopf, R. (2007). Elementary modeling in mathematics lessons: The interplay between real-world knowledge and mathematics structures. In W. Blum, P. L. Galbraith, H.W. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education: The 14th ICMI study* (pp. 209–216). Springer.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. Frontiers in Psychology, 10, 1-14. https://doi.org/10.3389/fpsyg.2019.00825

Simsek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education, 4*(4), 477-489.

Törnroos, J. (2005). Mathematics textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation, 31*, 315-327. https://doi.org/10.1016/j.stueduc.2005.11.005

Valverde, G., Bianchi, L, Wolfe, R., Schmidt, W. & Houang, R. (2002). *According to the book: Using TIMSS to investigate the translation of policy into practice through the world of textbooks*. Kluwer Academic Publishers.

Van den Heuvel-Panhuizen, M. (2005). The role of context in assessment problems in mathematics. *For the Learning of Mathematics, 25*(2), 2-23.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Swets & Zeitlinger.

Wijaya, A., van den Heuvel-Panhuizen, M., & Doorman, M. (2015). Opportunity-to-learn context-based tasks provided by mathematics textbooks. *Educational Studies in Mathematics, 89*(1), 41-65. https://doi.org/10.1007/s10649-015-9595-1

Yam, H. (2005). What is contextual learning and teaching in physics? Retrieved from http://www.phy.cuhk.edu.hk/contextual/approach/tem/brief_e.html

APPENDIX

Appendix 1. Context Analysis Form (CAF)

Sub-category (Code)	Explanation	
No context (A1)	Contains only mathematical symbols or structures	
Camouflage context (A2)	Daily life experiences and reasoning are not required. The mathematical operations required to give answer to the problems are already clear. The results can be found by combining the numbers given in the question text.	
Relevant and essential context (A3)	To provide answer to problem, common sense of reasoning within the context is necessary. The mathematical operation necessary for solving the problem is not obvious Mathematical modeling is necessary.	The item is included in the '*personal*' category if the item is related to students' families, their lives, such as shopping, games, personal life and so on (A3.1). The item is included in the '*occupational*' category if the item is related to the job/profession such as measuring, architecture, job-related decision-making and so on (A3.2). The item is included in the '*societal*' category if the item focuses on community perspectives, such as public transport, government, public policies and so on (A3.3). The item is included in the '*scientific*' category if the item context is related to science and technology, such as the weather, medicine, ecology and so on (A3.4).

Appendix 2. Checklists for Evaluating Item Quality (CEIQs)

Criteria list for short-answer items (B1.1)

Criteria
1.Can the items be answered with a number, symbol, word, or brief phrase? (B1.1.1)
2.Has textbook language been avoided? (B1.1.2)
3.Are the answer blanks equal in length? (B1.1.3)
4. Do the answers blank place at the end of the items? (B1.1.4)
5.Has the degree of precision been indicated for numerical answers? (B1.1.5)
6.Have the units been indicated when numerical answers are expressed in units? (B1.1.6)
7.Have the items been phrased so as to minimize spelling errors? (B1.1.7)
8. Do the items contain any clues? (B1.1.8)

Criteria list for true-false items (B1.2)

Criteria
1.Can each statement be clearly judged true or false? (B1.2.1)
2.Have specific determiners (e.g., usually, always) been avoided? (B1.2.2)
3.Have negative statements (especially double negative) been avoided? (B1.2.3)
4.Have the items been stated in simple, clear language? (B1.2.4)
5.Are the true and false items approximately equal in length? (B1.2.5)
6.Is there an approximately equal number of true and false items? (B1.2.6)
7.Has a detectable pattern of answers (e.g., T, F, T, F) been avoided? (B1.2.7)

Criteria list for matching items (B1.3)

Criteria
1.Is the material in the two lists homogeneous? (B1.3.1)
2. Do the responses rank alphabetically or numerically? (B1.3.2)
3.Do the directions indicate the basis for matching? (B1.3.3)
4. Do the directions specify the number of times each response may be used? (B1.3.4)
5.Is all of each matching item on the same page? (B1.3.5)
6.Is the list of responses longer or shorter than the list of premises? (B1.3.6)

Criteria list for multiple-choice items (B1.4)

Criteria
1. Is each item stem meaningful? (B1.4.1)
2. Do the item stems contain irrelevant material? (B1.4.2)
3. If used, has negative wording been given special emphasis (e.g., capitalized)? (B1.4.3)
4. Are there any grammatical consistency between the alternatives and the item stem? (B1.4.4)
5. Are the alternatives answers brief and free of unnecessary words? (B1.4.5)
6. Do the length and form of the alternatives similar? (B1.4.6)
7. Are the distracters plausible to low achievers? (B1.4.7)
8. Do the items contain any verbal clues to the answer? (B1.4.8)
9. Do the verbal alternatives rank alphabetically? (B1.4.9)
10. Do the numerical alternatives rank numerically? (B1.4.10)
11. Have none of the above and all of the above been avoided? (B1.4.11)

Criteria list for open-ended items (B1.5)

Criteria
1.Is the material to be interpreted appropriate to the students reading level? (B1.5.1)
2. Have pictorial materials been used whenever appropriate? (B1.5.2)
3. Does the material to be interpreted contain some novelty (to require interpretation? (B1.5.3)
4. Are the test items based directly on the introductory material (cannot be answered without it)? (B1.5.4)
5. Are the questions designed to measure higher-level learning outcomes? (B1.5.5)
6. Does each question specify the response expected? (B1.5.6)

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 236–257

https://doi.org/10.21449/ijate.1121828

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

The effect of positive error climate on affective domains in mathematics teaching

Merve Ozkaya[1,*], Senem Kalac[2], Alper Cihan Konyalioglu[3]

[1]Ataturk University, Faculty of Education, Department of Mathematics Education, Erzurum, Türkiye
[2]Ministry of National Education, Van, Türkiye
[3]Ataturk University, Faculty of Education, Department of Mathematics Education, Erzurum, Türkiye

ARTICLE HISTORY

Received: May 26, 2022
Revised: Oct. 16, 2022
Accepted: Oct. 25, 2022

Keywords:
Positive error climate,
Attitudes towards mathematics,
Error orientation,
Middle school students.

Abstract: The aim of this study is to investigate the effect of positive error climate in classrooms on middle school students' error orientations and attitudes towards mathematics. The data of the research were collected in the 2021-2022 academic year. The participants of the study consisted of 44 students in two 6th grade classes in a middle school in the city of Van, Türkiye. Quasi-experimental design was used in the research and the pre and post scores of the experimental and comparison groups were compared before and after the study. The data obtained using the "Mathematics Attitude Scale" and the "Error Climate Scale" were analyzed to examine the differences within and between groups. As a result of the findings, it was seen that the positive error climate in the experimental group made a positive significant difference both on the attitude towards mathematics and on error orientations of the students. No significant change was observed at the end of the study in the comparison group in which a neutral error climate was applied. The interviews with the course teacher who carried out the application and the observations made in the classroom reinforced the positive effect of the application. Positive error climate can be seen as a part of formative assessment as it has a corrective effect on teaching in the process.

1. INTRODUCTION

The classroom is the main environment where learning and teaching activities take place while each class has its own classroom climate that changes depending on the in-class variables. Classroom climate can be defined as consisting of mutual relations and communication between teachers and students (Akınoğlu, 2004; Kalaç & Özkaya, 2021). Similarly, in the classroom, the class has its own attitudes, behaviors, and perceptions towards errors. This situation, called the error climate, is likely to turn into a positive error climate in the classrooms where errors are considered as an integral part of the learning process (O'Dell, 2015; Stuer et al., 2013). In the related literature, it has been found that error-based learning studies applied in classrooms

*Corresponding Author: Merve Ozkaya ✉ mdurkaya@atauni.edu.tr Ataturk University, Faculty of Education, Department of Mathematics Education, Erzurum, Türkiye.

e-ISSN: 2148-7456 /© **IJATE** 2022

generally give positive results in affective terms on students and teachers (Akkuşçi, 2019; Gedik, 2014; Heinze & Reiss, 2007; O'Dell, 2015; Özkaya, 2015; Soncini et al., 2021).

In order to increase the quality of learning that will take place in the classroom, teachers are expected to catch all kinds of clues that occur in students and give the most appropriate feedback (National Council of Teachers of Mathematics [NCTM], 2000). This is possible by correctly evaluating the changes in student behavior while in-class assessments being an important part of the teaching process. Such evaluations not only show the teachers what the students have learned in the lesson, but also provide feedback that shows whether the program applied functions effectively or not. Instead of ignoring errors and failures, accepting them and including them in the education process are integral parts of the evaluation process (McMillan, 2015).

Students' perceptions of the classroom assessment atmosphere are a significant predictor of their attitudes towards school (İlhan, 2017). While a positive classroom climate affects students' attitudes towards school, according to Kohen (2006), positive classroom climates are more effective on student success and performance than a negative classroom climate. A positive climate in the classroom also positively affects the quality of learning. The way to create a positive classroom climate is to use errors in the classroom. The purpose of the error in the classroom varies as Heinze (2005) states that errors can be used as teaching tools. Likewise, errors can act as a springboard in education, revealing some hidden points in teaching and contributing to teaching (Borasi, 1986; 1994). Borasi (1988) stated that in the practices she made with her students by using professional mistakes, the students gained benefits in the field of mathematics by understanding and perceiving the nature of mathematics.

Many studies dealing with errors have been related to mathematics courses (Borasi, 1988; Bray & Santagata, 2013; Heinze & Reiss, 2007; Özkaya, 2015; Palkki & Hastö, 2018; Rach et al, 2013). While the mathematics lesson is seen as a difficult lesson for which students develop negative attitudes since primary education, it is also seen by teachers as a lesson difficult to teach as students have a low interest in such a lesson (Avcı et al., 2011; Öcalan, 2004). Furthermore, not only students but also teachers have different attitudes towards mathematics. According to Trisha (1999), teachers' attitudes towards mathematics can also affect students. Attitude is a learned tendency to react positively or negatively to a particular object, situation, institution, concept, or other person (Tezbaşaran, 1997). The results of the teacher's attitude towards mathematics and the results of the attitude towards errors show similarities. As a matter of fact, the attitudes of the students towards errors in the classroom are determined by the attitude of the teacher towards errors. It has been observed that the same attitude develops in students in studies where the teacher is moderate towards errors and sees them as learning opportunities (Borasi, 1988; Bray, 2011; Heinze & Reiss, 2007; Tulis, 2013). On the contrary, a strict teacher's attitude towards errors reduces the possibility of learning from errors (Oser & Spychiger, 2005). Showing students how to improve their learning by using errors in the learning process is one of the components that increase student motivation (McMillan & Workman, 1998).

It is also important to note that there used to be a negative view of errors in teaching. This understanding of error, which was accepted before the constructivist approach, was also used in mathematics teaching. According to this understanding, error is a situation that should be avoided. However, errors are an important tool used to identify students' learning difficulties and provide important information about students' thinking processes (Baştürk, 2014; Borasi, 1996). In addition to this diagnostic feature of errors, errors can be turned into an opportunity in the classroom (Guzmán-Muñoz et al., 2009).

Error in mathematics teaching is the misuse and conclusion of mathematical expressions and ideas (Erbaş et al., 2010). From another perspective Borasi (1988) describes the use of errors in teaching as a springboard. According to Borasi (1988), errors save students from regarding mathematics as unnecessary and allow teachers to use errors as a teaching tool in the curriculum. Borasi (1988) also stated that errors in teaching are not adequately examined, and with her work, she showed that the conscious use of errors in teaching enriches teaching since students not only have the opportunity to learn mathematical concepts more deeply, but also increase their interest and curiosity towards mathematics (Borasi, 1986; 1989). In this way, in classroom atmospheres where a positive error climate is created, both students and teachers have a positive perception of errors. Students not afraid of making errors can turn this situation into a positive one (Guzmán-Muñoz et al., 2009; Heinze, 2005; Heinze & Reiss, 2007).

The effect of using errors in teaching on secondary and high school students was investigated by Heinze and Reiss (2007) and their study showed that although there was no cognitive difference between the two groups, it was determined that the students in the experimental group were positively affected. Akkuşçi (2019) obtained similar results in his study and found that there was an increase in students' critical thinking skills although he did not find a difference between the academic achievements of the students in the quasi-experimental stage of his study. Error-based practices had positive effects not only on students but also on teachers. In Gedik’s (2014) and Özkaya’s (2015) studies, it was also found that the affective effects of the error practices in the classroom were more than the cognitive effects on the teachers, and that these practices provided the teachers with the ability to conduct research and critical thinking affectively. In the study of Oser and Spychiger (2005), it was seen that students were affected by their teachers in their attitudes towards errors as the teacher's view and attitude towards errors cause the student to have the same point of view.

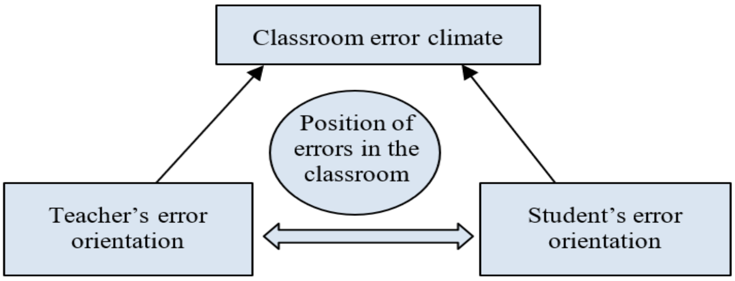
In their quasi-experimental study with middle and high school students Rach et al. (2013) found similar results like those in Heinze and Reiss’s (2007) study. Rach et al. (2013) investigated students’ attitudes towards errors and also whether students saw errors as an opportunity for learning. In their research, it was observed that the students in the experimental group were more courageous in making errors than the control group were although there was no significant difference between the two groups in terms of student attitude towards errors. It has been observed that while students are dealing with errors, learning processes are positively supported in a learning atmosphere that is moderate against errors. According to Rach et al. (2013) understanding of errors is necessary to distinguish between right processes or phenomena and wrong environment.

For errors to be effective in teaching, corrective feedback must be followed. Huelser and Metcalfe (2012) stated that generating an error serves more reminder than presenting the answer at the point of reaching correct answers, as long as it follows corrective feedback. With feedback, individuals not only get the right answer, but also increase their analysis and explanation abilities, thus in this way, the amount of learning from errors increases (Metcalfe, 2017). Accordingly, a positive classroom error climate is observed in environments where corrective and remedial feedback is given to errors.

Classroom error climate means how errors are used and evaluated in the classroom (O'Dell, 2015; Steuer et al., 2013). A positive error climate is observed in classroom atmospheres where errors are used as a part of the learning process in the classroom and errors are viewed positively. In such an atmosphere students can realize their misconceptions and start their learning process. These arrangements within the classroom express a positive culture of error (O'Dell, 2015; Oser & Spychiger, 2005). The classroom error climate is determined by the attitudes, behaviors, and perceptions of teachers and students towards errors.

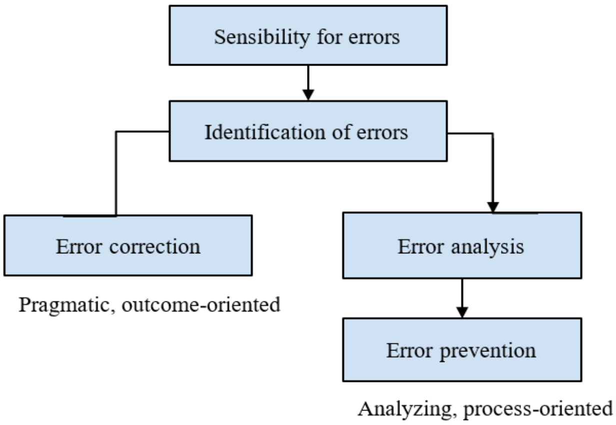
Error orientation refers to the way teachers understand, react, and use student errors in learning (O'Dell, 2015). Considering error orientation as part of the classroom error climate it can also be defined as the attitude of teachers and students towards errors, whether to use errors actively in the learning environment or not, the attitude towards making errors, and the accompanying behaviors (Kalaç & Özkaya, 2021). According to O'Dell (2015), a positive error orientation, which sees errors as learning opportunities rather than punishments, reduces negative academic motivation and can help improve students' perceptions, self-efficacy, and future goals. The classroom error climate is modeled in Figure 1.

Figure 1. *Components of the classroom error climate (Kalac & Ozkaya, 2021).*


Classroom error climate
Position of errors in the classroom
Teacher's error orientation
Student's error orientation

There are two ways to evaluate errors in the classroom, one of which is result-based aimed at correcting errors directly and the other is process-based including analyzing errors and preventing error precaution. According to Heinze (2005), teachers look at errors negatively because they disrupt the process in the classroom and mostly refer to direct intervention to errors made in the classroom. In the related studies, it is seen that there is mostly teacher intervention to the errors and the answer is given directly to the student (Son, 2013; Son & Sinclair, 2010). Rach et al. (2013) modeled the role of errors in the learning process as in Figure 2.

Figure 2. *Model of the role of errors in the learning process.*



There are eight dimensions of the classroom error climate in the "Perceived Error Climate Scale" created by Steuer et al. (2013). One of the dimensions of the perceived error climate in the classroom is error tolerance by teacher. Cultural beliefs and teaching practices shape teachers' reactions to errors (Santagata, 2004). Another dimension is irrelevance errors to assessment, which is about whether student errors adversely affect their performance and grade evaluation. Teacher support following errors made in the classroom is also an important dimension. This sub-dimension expresses the teacher's patience, explanations, and assistance in the face of student mistakes. When the studies involving interventional approaches to errors are examined, it has been observed that most of the errors made are not ignored (Didiş et al., 2016; Didiş-Kabar & Amaç, 2018; Son, 2013; Türkdoğan & Baki, 2012). Analyzing the data they obtained from 44 pre-service teachers through teaching scenarios, Didiş-Kabar and Amaç (2018) revealed that pre-service teachers had interventions such as recognizing the error, explaining the question, and lecturing. Analysis of errors and functionality of errors for learning sub-dimensions express how errors are handled in the classroom and their status in learning processes. The sub-dimensions of Absence of negative classmate reactions to errors and Absence of negative teacher reactions to errors refer to verbal and nonverbal reactions to student mistakes. Taking the error risk, the other sub-dimension of the classroom error climate, expresses the student's courage to make errors without being sure of her answer. According to Steuer et al. (2013), although the sub-dimensions of the perceived error climate are distinguishable, they are closely related sub-dimensions.

Discussing the errors made in the classroom and using them in teaching constitute an important part of the classroom error climate (Steuer et al*.*, 2013). This way of teaching positively affects student achievement (Barbieri & Booth, 2020; Heinze & Reiss, 2007; Rittle-Johnson & Star, 2009; Yıldırım, 2019) as if the error climate in the classroom is transformed into a positive one, students' perceptions of coping with their errors in a reliable and supportive learning environment will increase (Soncini et al*.*, 2021).

Students operate a verbal or nonverbal reasoning process in the lessons. One of the courses in which reasoning processes are most intense is mathematics. Students can make errors and these errors often put the student in a negative situation. It is thought that turning this situation into a positive one can contribute both cognitively and affectively to the students in many courses, especially in mathematics. In order to reveal whether this mentioned purpose can be realized or not, decimal notation has been chosen. Although the learning outcomes of decimal notation are seen in the fifth and sixth grades in the curriculum, this subject is related to most concepts in mathematics (for example percentages, rational numbers, length, and liquid measures). Apart from its importance, students have difficulty in understanding decimal notation and they make a lot of errors (Haser & Ubuz, 2000; Kaya, 2015; Yenil, 2020). Using errors in the mathematics teaching process within the framework of a planned learning process can create a positive error climate in the classroom. It is thought that a positive error climate may also affect attitude, which is another affective condition such as motivation. In this context, the aim of this specific research is to reveal the effect of the positive error climate in mathematics lessons on students' error orientations and attitudes towards mathematics.

To this end, the problem statement of the research is "Does the positive error climate created in mathematics lessons make a significant difference in students' error orientations and attitudes towards mathematics?" The research questions generated for this problem statement are as follows:

- Does the positive error climate created in the classroom make a significant difference on students' error orientations?
- Does the positive error climate created in the classroom make a significant difference on students' attitudes towards mathematics?
- What are the changes observed in students in the classroom where the positive error climate is created?
- What are the views of the teacher who performed the application in the process?

2. METHOD

In the research, a quasi-experimental design with nonequivalent pretest-posttest comparison group from quantitative approaches was decided as a research method in order to compare a positive error climate against a neutral error climate and examine the effects of a positive error climate. The quasi-experimental design is one of the research designs used to explore cause-effect relationships between variables. In this design, the groups available are randomly assigned as comparison and experimental groups. Designs with nonequivalent pretest-posttest comparison groups assign the groups randomly because it is not possible to randomly assign the participants and such designs are widely used in the field of education (Fraenkel et al., 2012; McMillan & Schumacher, 2010).

2.1. Participants

The participants of the research consisted of two 6th grade classes and 44 students studying in those classes in a middle school in Van, Türkiye in the 2021-2022 academic year and they were selected by purposive sampling. Purposive sampling occurs when the researcher makes a judgment about which participants should be selected in order to provide the best information that will serve his/her purpose (McMillan & Schumacher, 2010). The 6A class was assigned as the experiment and the 6B class as the comparison group by random assignment. The distribution of the samples by group and gender is given in Table 1.

Table 1. *Distribution of students in the sample by group and gender.*

Groups	Gender		Total
	Female	Male	
Experimental Group	9	11	20
Comparison Group	9	15	24
Total	18	26	44

It is known that these groups had similar success averages according to the mathematics score averages found the previous year. Neither of the groups learned in a positive error climate before the application. On the other hand, when the situation of the students in the sample of the research is evaluated in terms of socio-economic status, it is known that the students came from families with a medium socio-economic status. Students were coded S1, S2…, S44. The mathematics teacher who performed the application throughout the process had ten years of experience at the time of data collection and taught at every grade level.

2.2. Data Collection Tools

In the research, "Error Climate Scale", "Mathematics Attitude Scale", in-class observations, and interview forms were used.

2.2.1. *Error climate scale (ECS)*

This scale was developed to measure the perceived error climate in the classroom by Steuer et al. (2013). The scale is a 5-point Likert type scale, where 1 indicates that participants strongly disagree with the statement, while 5 indicates strongly agree. It was adapted into Turkish by Kalaç et al. (2022). As a result of the adaptation study of the classroom ECS, the scale consisted of 27 items and 7 factors for the Turkish sample; namely, A1; Irrelevance of errors for assessment, A2; Teacher support after errors, A3; Absence of negative teacher reactions to errors, A4; Absence of negative classmate reactions to errors, A5; Taking error risks, A6; Analysis of errors, A7; Functionality of errors for learning. While the Cronbach Alpha internal consistency value was .86 for the general scale, it was found between .73 and .89 for the sub-factors. The answers given by the students to the items are scored between 1 and 5 and the error orientation score of the student is determined. With the data obtained from the students before the application, the Cronbach Alpha reliability coefficient for the ECS was found to be .81.

2.2.2. *Mathematics attitude scale (MAS)*

The scale was prepared by Önal (2013) to measure the mathematics attitudes of middle school students. Validity and reliability studies of the scale consisting of 22 items and 4 sub-dimensions (B1: interest, B2: anxiety, B3: necessity, B4: study) were conducted. The internal consistency coefficient for the whole scale was found to be .90. The internal consistency coefficient of the factors that form the scale varied between .69 and .89. The answers given by the students to the items were scored between 1 and 5 so that the student's mathematics attitude scores could be determined. With the data obtained from the students before the application, the Cronbach Alpha reliability coefficient for the MAS was found to be .83.

2.2.3. *Observation and interview*

Teaching was carried out in the classroom under the guidance of the Positive Error Climate Framework program. Observations and interviews as data collections methods were used to illustrate the situation in the classroom. Thus, the findings obtained from the MAS and ECS were also supported. In the experimental group, an observation form was prepared to follow the process of the activities prepared with the teacher within the scope of the Positive Error Climate Framework program. With this form, which was prepared as unstructured, the important points about the errors in the classroom were recorded. During the process, an interview was held to reveal the teacher's thoughts on the positive error climate. The interview questions prepared were examined by two researchers as experts in their fields. With the common opinions of these experts, the interview questions took their final form.

In the pre-interview and post-interview, the teacher was asked such questions as "What is your view on students' errors?", "What do you think about using error examples in the lesson?", "What is the situation of students’ fear of making mistakes in the lesson?", "If you evaluate the two classes together, is there a difference between the error orientations? If so, what is the relationship?", "How is the error tolerance of the students towards each other?" Thus, the teacher's views on the change in the process were also taken.

2.3. Data Analysis

As a result of the application, there may be differences in the pre-test and post-test scale scores within and between the groups. In order to decide the significance of this difference, the paired t-test among the parametric tests and the independent sample t-test between groups were used. While the paired-sample t-test is used to decide whether the mean score difference that may occur within the group is significant or not, the independent sample t-test is used to decide whether the difference between the means of two independent groups is significant or not. Both tests are expected to satisfy the assumption of normal distribution. Otherwise, non-parametric equivalents of these tests, Wilcoxon test and Mann-Whitney U test can be used (Büyüköztürk, 2020; Özdamar, 2018). Likewise, these tests were used in the sub-dimensions of the scales, depending on the condition of providing the assumption of normal distribution. In order to control the assumption of normality, Shapiro-Wilk normality analysis and skewness- kurtosis values of the data were examined according to the groups to be compared.

ECS and MAS were applied to both classes before and after the application. The scores obtained from the ECS were evaluated as error orientation scores, and the scores obtained from the MAS were evaluated as mathematics attitude scores. Negatively worded items were reverse coded before the analysis. While calculating the total scale scores, the scores given by the participants to each item were added. While calculating the scores in the sub-factors, the scores in the items related to each sub-factor were summed and the total factor scores were calculated, and the analyzes were made on these scores. Paired sample t-test and Wilcoxon signed-rank analysis, which is the nonparametric equivalent of this test, were used to understand whether the differences within the group were significant or not. In order to understand whether the differences between the groups were significant, the independent sample t-test and Mann Whitney U test, which is the nonparametric equivalent of this test, were used. To calculate the effect size (d) of the differences found to be significant in the paired t-test analyses, the t value obtained was calculated by dividing the square root of the total number of participants. To calculate the effect size (d) of the differences found to be significant in the paired t-test analyses, the t value obtained was calculated by dividing the square root of the total number of participants. Likewise, the *z*-value obtained was divided by the square root of the total number of the participants in order to calculate the effect size (*d*) of the differences found to be significant in the Wilcoxon Signed Rank Test analysis. The effect size found (*d*) .2, .5 and .8 are interpreted as small, medium and large effects, respectively (Büyüköztürk, 2020; Cohen, 1988; Özdamar, 2018).

The data obtained from the observations and interviews were analyzed descriptively. Descriptive analysis details the obtained data by quoting directly. By reducing the data, it allows the subject to be presented and defined in a regular way (Ekiz, 2009).

2.4. Description of the Learning Environment on Positive Error Climate in the Classroom

In our study, a Positive Error Climate Framework program was prepared by associating the framework program suggested by Bray (2011) with the classroom error climate components of Steuer et al. (2013). In the research, it has been determined that although teachers look at mistakes positively, they are hesitant at the point of use in the classroom, and it has been revealed that teachers do not know how to use mistakes as an opportunity to teach (Özkaya & Konyalıoğlu, 2019; Palkki & Hastö, 2018). Bray (2011) presents a framework program that details how lessons are designed and implemented in order to take advantage of the teaching potential of errors in mathematics education. The program steps are given as follows:

i. Choosing mathematical tasks for their potential to elicit students' misunderstandings.

ii. Planning lessons by evaluating how mistakes can be used to improve students' mathematical understanding.

iii. Developing a plan for including mistakes in class discussions.

iv. Involving all students in the class to analyze and review errors so that students are elucidated on fundamental mathematical concepts.

In order to use this framework program suggested by Bray (2011), first of all, teachers should know the error climate of the classroom and be willing to turn it into a positive one.

In the research conducted using a quasi-experimental design, one of the classes was assigned as the experimental group and the other as the comparison group with purposive sampling. During the study, which lasted for six weeks and twelve hours, the class selected as the comparison group was given the learning outcomes-based instruction in mathematics lessons during the research. In the class assigned as the experimental group, the lessons were taught by creating a positive error climate in the classroom in addition to learning outcomes-based instruction.

The teacher who would carry out the application was informed about how to create a positive error climate in the classroom. The Positive Error Climate Framework plan was made together with the teacher. In this plan, it was stated how the teacher should give feedback to the errors, how she would motivate the students, and also how she would carry out the process. “The Positive Error Climate Framework” plan is given in Appendix at the end of the article. An interview was held with the course teacher about the error orientations of the classes. ECS and MAS pre-tests were applied to both groups before the application. In-class observations were used to observe the progress of the process.

A positive error climate activity was held every week in the classroom where a positive error climate was created. The teacher encouraged the students who did not attend the lessons and obtained answers from them about the question/subject. The answers of the students were discussed in the class and the teacher gave feedback to the students who gave wrong answers such as" Why did you think that way, let's think about the answer together, well done, you caught a very good point, you have revealed a general error made on this subject, thank you..." in order to motivate the students. During the lesson, the teacher tried to show the students that she was tolerant of errors with her actions and words.

At the end of the subject, the teacher added incorrect questions/phrases to the evaluation exam. Students were asked to write why the statements they thought were errors. The teacher solved the evaluation questions in the classroom with the students. Those students who did not want to go to the blackboard in the classroom and who were behind the class academically in mathematics lessons were encouraged to attend the lesson. At the end of each answer, the teacher asked the students to explain their answers and ECS and MAS were applied to both groups again as a post-test. The course teacher was interviewed about the process, as well.

3. RESULTS

In this section, first the pre-test, post-test ECS, and MAS analyzes of the experimental and comparison groups are given, and then, one of the observations made in the classroom and the teacher's views are presented.

3.1. Analysis of the ECS and MAS

Both the whole scale and the sub-factors were examined separately to see whether the changes in the pre-test and post-test scores in the ECS and MAS made a significant difference. Table 2 shows the statistics of the difference in the total scores of the groups in the ECS and MAS and the results of the Shapiro-Wilk normality test.

Table 2. *Group statistics and Shapiro-Wilk normality test.*

Groups	Scales	*N*	*D*	*SD*	SC	KC	*p*
Experimental	ECS	20	11.10	12.22	-.834	-.278	.038
	MAS	20	5.25	8.42	-.850	-.542	.011
Comparison	ECS	24	-.62	15.13	-.695	1.14	.283
	MAS	24	1.33	11.40	1.54	4.56	.004

N: Number of students D: Difference score, SD: standard deviation, SC: Skewness coefficient, KC: Kurtosis coefficient, p: Significance value

In Table 2, it was seen that the difference between ECS and MAS of the experimental group and the MAS pre-test and post-test total score of the comparison group did not show a normal distribution (*p*<.05). In the comparison group, although the difference between the pre-test and post-test difference scores of ECS showed a normal distribution according to the *p* value (*p*>.05), it was observed that the kurtosis value deviated from the normal distribution outside the range of (-1, 1) excessively. The main thing in the investigation of normality is that the data do not deviate excessively from the normal distribution (Büyüköztürk, 2020). In this case, it can be said that the difference between the total scores does not fit the normal distribution for both groups. Wilcoxon signed-rank test, which is the nonparametric equivalent of the paired t-test, was used to decide whether the difference between the pre-test and post-test scores of the groups at the end of the application was significant or not (see Büyüköztürk, 2020; Özdamar, 2018). Wilcoxon signed-rank test analysis results for two groups are given in Table 3.

As can be seen in Table 3, a significant difference was found between the pre and post test scores of the experimental group's mathematics attitude scores (*z*=-2.32; *p*=.020). Similarly, a significant difference was found between the error orientation scores of the group (*z*=-3.24; *p*=.001). When Table 3 is examined in both scales, it is seen that this difference is in favor of the post-test as positive rank totals are larger than the negative ones. The effect size of the difference in MAS was calculated as *d*=.51 and the effect size of the difference in ECS as *d*=.72. It can be said that the effect sizes found for both differences have medium effect. As a result of the post-tests of the experimental group, it was observed that there was a positive change in both their attitudes towards mathematics and their error orientation and this change was significant according to the Wilcoxon signed-rank test analysis results. When Table 3 is examined, it is seen that there is no significant difference between the pre-test and post-test mathematics attitude scores of the comparison group (*z*=-1.134; *p*=.257). Similarly, there is no significant difference between the error orientation scores of the group (*z*=-.237; *p*=.813). According to Wilcoxon test results, no difference was found between the pre-test and post-test scores in both error orientation and attitudes towards mathematics of the comparison group.

Table 3. *Wilcoxon test results of the groups' scores before and after the application.*

Group	Scales	Pre-test/Post-test	*N*	Mean Rank	Sum of Ranks	*z*	*p*
Experimental	MAS	Negative Ranks	4	3.88	15.50	-2.32	.020*
		Positive Ranks	10	8.95	89.50		
		Ties	6	0	0		
	ECS	Negative Ranks	3	2.67	8.00	-3.244	.001*
		Positive Ranks	14	10.36	145.00		
		Ties	3	0	0		
Comparison	MAS	Negative Ranks	6	9.92	59.50	-1.134	.257
		Positive Ranks	12	9.29	111.50		
		Ties	6	0	0		
	ECS	Negative Ranks	9	7.94	71.50	-.237	.813
		Positive Ranks	8	10.19	81.50		
		Ties	7	0	0		

*p<.05

The normal distribution of the data was investigated to see if there was a significant difference between the post-test scores of the experimental and comparison groups. Descriptive statistics of scales and Shapiro-Wilk normality analyzes are given in Table 4.

Table 4. *Group statistics and Shapiro-Wilk test for ECS and MAS post tests.*

	Group	*N*	M	*SD*	Shapiro-Wilk	
					df	*p*
ECS Post-test	Experimental	20	105.40	15.83	42	.009
	Comparison	24	95.83	26.41		
MAS Post-test	Experimental	20	81.15	12.55	42	.54
	Comparison	24	68.04	17.74		

N: Number of students, M: Mean: SD: standard deviation, df: Degree of freedom, p: Significance value

When Table 4 is examined, it is seen that the ECS post-test does not comply with the normal distribution, while the MAS post-test complies with the normal distribution. Whether the differences between the groups caused a significant difference in the change in scale scores was analyzed with the independent sample t-test for normal distribution and Mann-Whitney U test for non-normal distribution.

According to the independent sample t-test results, there is a significant difference between the experimental and comparison groups' post-test scores in MAS [$t_{(42)}$=2.77, *p*<.05]. The effect size of the difference between the groups in the MAS post-test was calculated as *d*=.83. The difference between the experimental and comparison groups' MAS post-test scores is a large difference, which can be considered significant. Mann-Whitney U analysis was performed to examine the difference between groups for ECS post-test scores that did not fit the normal distribution. The results of the analysis ECS post-test show that the difference in scores between the groups was found to be insignificant (*U*=175; *p* =.12; *z* =-1.53). The effect size of the difference between the groups in the ECS post-tests was calculated as *d*=.23. This difference shows that experimental and comparison groups' ECS post-tests scores is a small difference which can be considered as non-significant.

ECS consisted of seven sub-dimensions and MAS consisted of four sub-dimensions, and as a result of the pre-test and post-test, changes occurred between the total scores of these sub-dimensions. In order to see whether these changes create a significant difference, first of all, group statistics and the normal distribution of total score differences were examined. The descriptive statistics of the total score differences of the factors and the significance values obtained as a result of the Shapiro-Wilk normality analysis are given in Table 5.

Table 5. *Total score difference statistics of factors and Shapiro-Wilk normality test.*

Group	Scales	Factors	*N*	D	*SD*	*p*
Experimental	ECS	A1	20	.80	4.443	.244
		A2	20	.20	4.372	<.001
		A3	20	3.15	5.214	<.001
		A4	20	2.30	4.053	.017
		A5	20	2.40	3.101	.012
		A6	20	1.95	2.999	.027
		A7	20	.30	2.494	.503
	MAS	B1	20	3.75	6.455	.310
		B2	20	1.10	3.291	.042
		B3	20	-.70	3.614	.531
		B4	20	1.10	1.682	.001
Comparison	ECS	A1	24	.125	5.407	.157
		A2	24	-.166	3.963	.092
		A3	24	-1.62	6.212	.001
		A4	24	-.291	2.475	.053
		A5	24	.958	2.710	.026
		A6	24	-.416	2.339	<.001
		A7	24	.791	4.117	.001
	MAS	B1	24	1.04	5.287	.020
		B2	24	.5417	5.815	<.001
		B3	24	-.6667	4.039	.017
		B4	24	.4167	2.244	.001

N: Number of students D: Difference score, SD: standard deviation, p: Significance value

When Table 5 is examined, it is seen that the total score differences of the A1 (Irrelevance of errors for assessment) and A7 (Functionality of errors for learning) factors in the ECS test in the experimental group conform to the normal distribution (*p*>.05), while the total score differences of the other factors do not fit the normal distribution according to the results of Shapiro-Wilk normality analysis (*p*<.05). In the experimental group, it was also observed that the total score differences of the B1 (interest) and B3 (necessity) factors in the MAS test conformed to the normal distribution. In the comparison group, A1 (Irrelevance of errors for assessment), A2 (Teacher support after errors) and A4 (Absence of negative classmate reactions to errors) factor total score differences were in normal distribution, while total score differences of all sub-factors in the MAS scale did not comply with the normal distribution. In order to decide whether the difference between the factor total score difference obtained as a result of the pre-test and post-test is significant, the paired t-test for the factors satisfying the normality condition and the Wilcoxon signed-rank test were used to determine whether the difference between the total score differences of the factors that did not show normal distribution was significant or not.

Paired-sample t-test analyzes of A1, A7 and B1, B3 sub-factors satisfying the normality condition in the experimental group and A1, A2 and A4 factors in the comparison group providing the normal distribution condition were made and only the B1 factor in the experimental group was determined to have a significant difference in favor of the post-test [$t_{(19)}$=2.59, p<.05]. The effect size of this significance value was calculated as d=.57 and it was determined that the difference between them created a medium effect.

Wilcoxon signed-rank test was used for other factors that did not meet the normal distribution condition given in Table 6. The test results are given in Table 6.

Table 6. *Wilcoxon signed-row analysis results.*

Groups	Factors	Pre-test/Post-test	*N*	Mean Rank	Sum of Ranks	*z*	*p*	*d*
Experimental	A2	Negative Ranks	4	4.13	16.50	-.424	.671	-
		Positive Ranks	3	3.83	11.50			
		Ties	13					
	A3	Negative Ranks	12	8.25	99.00	-2.937	.003*	.66
		Positive Ranks	2	3.00	6.00			
		Ties	6					
	A4	Negative Ranks	13	8.00	104.00	-2.519	.012*	.56
		Positive Ranks	2	8.00	16.00			
		Ties	5					
	A5	Negative Ranks	13	8.69	113.00	-3.035	.002*	.68
		Positive Ranks	2	3.50	7.00			
		Ties	5					
	A6	Negative Ranks	11	9.64	106.00	-2.625	.009*	.59
		Positive Ranks	4	3.50	14.00			
		Ties	5					
	B2	Negative Ranks	10	6.75	67.50	-1.544	.123	-
		Positive Ranks	3	7.83	23.50			
		Ties	7					
	B4	Negative Ranks	9	5.94	53.50	-2.714	.007*	.60
		Positive Ranks	1	1.50	1.50			
		Ties	10					

*p<.05

When Table 6 is examined, a significant difference was found in the A3, A4, A5, and A6 factors of the ECS test and the B4 factor of the MAS test in the experimental group compared to the Wilcoxon signed-row test (p<.05). When the effect size of these differences is examined, it is seen that they have a medium effect (.5<d<.8). No significant difference was observed between the pre-test and post-test for any of the factors in the comparison group according to the Wilcoxon signed-rank analysis.

3.2. In-class Observations and the Course Teacher's Views

As a result of the observations made in the classroom, it was observed that the students who did not want to attend the lesson or remained silent because they did not trust their answers at the beginning of the research increased their participation in the lesson at the end of the application and did not hesitate to answer even if their answers were wrong. At the end of the application, a decrease was observed in the behavior of the students who made fun of their friends who gave wrong answers to the questions in the lesson.

During the application, the teacher, who made the process evaluation at the end of the subject, gave midterm exams to the students. She added one incorrect statement/question to the exams she prepared. Emphasizing several times before and during the exam, she said, "Please write down why they are wrong in front of the statements that you think are wrong and do not leave them blank". At the end of the exam, she evaluated the questions in class with the students and solved the assessment questions in the classroom together with the students. An example of an evaluation of the third week of the positive error climate is given as follows:

Question: Write a suitable number according to the expressions given in the blanks below:

• Greater than 8, less than 9……….

• Greater than 5, less than 5.1…………

• Greater than 2.5, less than 2.45 ………

• Greater than 0.32, less than 0.33…………

Yes children, most of you left question 3 blank. Let's examine together. Anyone wants to answer? (The teacher makes a promise to a student in the classroom who does not attend much.)

S1: It can be 2.44.

Teacher: Good answer. Why did you think like that?

S1: Because it is one step away from 2.45.

S2: But the number he said is less than 2.5. No way.

Teacher: So what could this number be? Or is it just one?

S3: No. The numbers are endless.

Teacher: Okay, then say one of those numbers, (the teacher picks up a student who doesn't raise a finger in the lesson) S4, which one of these numbers do you think is bigger?

S4: 2.45.,

Teacher: Why do you think so?

S4: Because the 45 is greater than 5.

Class: No, it's not. We can add as many zeros as we want to the end of the number after the comma. So it's not 5, it's 50 actually.

S5: We put a zero at the end of 2,5, it becomes 2.50. Then it becomes 2.50 > 2.45.

Teacher: Well done children, your friend S4 caught a very fine and important mistake. This is one of the most common mistakes made. Thanks for your friend pointing out this common mistake. S1-S4, did you understand the mistake? (Students say they understand and once again state why their answer is wrong in their own words.)

Teacher: Then what is the answer to this question, guys?

S3: There is no answer. There are no numbers in this range (Classmates confirm the answer).

As seen above, an evaluation exam was conducted at the end of the topic related to the order of decimal numbers, belonging to the third week of the positive error climate. The teacher solved the evaluation questions in the classroom together with the students. Those students who did not want to stand at the blackboard in the classroom and who were behind the class academically in mathematics lessons attended the lesson and misunderstandings in the students were revealed. At the end of each answer, the teacher asked the students to explain their answers. Misconceptions in students were both revealed and corrective feedback was given.

In the interview with the course teacher before the application, the teacher stated that she did not look positively towards making errors intentionally during the lesson. Although she did not have a negative attitude towards the students who made errors, the students kept silent in order not to give wrong answers. She also stated that the students often made fun of each other when they made errors in the class. At the end of the application, the course teacher said that she was very satisfied with the process since the process contributed positively to the students who did not attend the lesson much, and that if it was planned in this way, the students could benefit from their errors in the lessons.

4. DISCUSSION and CONCLUSION

As a result of the findings, it was observed that there was an increase in the error orientation and mathematics attitude scores of both the experimental and comparison groups. In the analyzes made to decide whether this increase had a significant effect or not, it was seen that only the increase in the experimental group was significant. When the error orientation and mathematics attitude scores of the experimental group before and after the application were compared according to the Wilcoxon signed-rank analysis, it was observed that there was a significant difference between the scores (p<.05). According to the research findings, it can be said that the positive error climate in the classroom has a positive effect on both students' attitudes towards mathematics and their error orientation. Independent sample t-test and Mann-Whitney U test analyzes were performed to see if there was a significant difference in error orientation and mathematics attitude post-test scores between the experimental and comparison groups. While there is a significant difference between the post-test scores of the experimental and comparison groups for MAS, there is no significant difference in ECS. Although there is no significant difference between the post-test scores of the two groups for ECS, the mean score of the experimental group (M=105.40) is higher than the mean score of the comparison group (M=95.83).

Paired sample t-test and Wilcoxon signed-rank test analyzes were applied to the sub-dimensions to decide whether the scores given by the experimental and comparison groups to the sub-dimensions of the ECS and MAS scales before and after the application made a significant difference or not. As a result of the analysis, it was seen that only the differences in the experimental group were in favor of the post-test. According to the paired t-test and Wilcoxon signed-rank test analyzes performed in the experimental group, a significant difference was found in favor of the post-test in the A3, A4, A5, and A6 sub-factors of the ECS (p<.05). Similarly, significant differences were found in favor of post-test in B1 and B4 sub-factors of MAS (p<.05). Thus, it can be said that the positive error climate application in the experimental group also gave positive results in the ECS and MAS sub-factors. During this application, it may be expected that there will be a change in A3 and A4 sub-factors since the teacher creates a positive error climate in the classroom since these two sub-factors include the positive behavior of the teacher against errors and support against errors. A5 and A6 sub-factors are related to the learners. With the positive error climate in the classroom, the students became able to take the risk of making errors and started not to react negatively to their friends who made errors in the classroom. The significant change in the B1 sub-factor indicates that positive classroom application increases the interest in the lesson; however, application has a medium effect on these significant changes. This effect is thought to increase with a longer application.

Error-based learning and teaching studies applied in the classrooms in the relevant literature generally leave a positive impression on students and teachers (Akkuşçi, 2019; Bray & Santagata, 2013; Gedik, 2014; Heinze & Reiss, 2007; O'Dell, 2015; Özkaya, 2015; Soncini et al., 2021). The results of this study show similarity to other studies in the literature at this point.

In studies conducted with teachers, it has been observed that they generally have positive beliefs about using errors in teaching, yet they are distant about making use of errors in the lesson (Ingram et al., 2015; Palkki & Hastö, 2018). The reason for this is the thought that the errors used will become widespread. A number of researchers show that one of the ways to prevent this is to intervene directly (Heinze, 2005; Özkaya, 2015; Santaga, 2005; Türkdoğan & Baki, 2012). When the views of the teachers before and after our specific research are examined, it can be seen that the course teacher had the same thought at the beginning of the research. When the study of Bray and Santagata (2013) was examined, it was determined that the teachers managed the learning process from errors better in the lessons in which lesson plans containing errors were applied. At the end of our research, the teacher who applied the positive error climate stated that she could also benefit from a planned error climate management process in the lessons.

According to research, teachers' attitudes towards errors in the classroom determine students' attitudes towards errors and mistakes. Teachers' tolerant attitude to errors and using them as teaching tools in the classroom cause students to adopt the same attitude (Bray, 2011; Heinze & Reiss, 2007; Tulis, 2013). Actually, it was seen that the students' error orientation and mathematics attitude scores increased positively as a result of the teacher's positive attitude towards errors in the experimental group in which a positive error climate was carried out.

If a teacher who includes errors in the learning process in the classroom knows how to benefit from errors and s/he draws a well line in teaching, he or she can benefit from errors as a teaching tool (Akpınar & Akdoğan, 2010; Heinze & Reiss, 2007). To such an end, the first step is to motivate the student. According to Tulis (2013), an error leads to an affective reaction and a good regulation process is required to turn this reaction into a positive one. When encouraging feedback is given to the student who is afraid of making errors, it has been observed that the student is more willing to participate in the lesson. In the same way, it was determined that there was a decrease in the behaviors of students who made fun of their friends' wrong answers during the process.

Mathematics curriculum in Türkiye expects teachers to evaluate students holistically and multidimensionally (MoNE, 2018). The quality of evaluations is determined by the methods and feedback used in the process (Bray, 2011). The purpose of evaluation is not only to grade the student but also to contribute to the improvement of the course. A positive error climate is one of the ways that can be used to improve the lesson while a positive error climate gives feedback to students' errors correctly and helps evaluate students for learning. Huelser and Metcalfe (2012) emphasized the importance of corrective feedback and stated that the correct feedback is more effective in remembering the answer than explaining the truth directly. With feedback and correct guidance, individuals not only get the right answer, but also increase their ability to analyze and explain; thus, they both begin to query and increase the amount of learning from errors (Karadağ, 2004; Metcalfe, 2017). In our specific study the teacher asked the student to explain his/her answer and made the class think about it, which contributed to the students' ability to explain the error and express why it was wrong, and which also contributed to their comprehension skills rather than memorizing the answer. Baki (2015) divides the evaluation approaches into three as diagnostic assessment, formative assessment, and complementary assessment. Formative assessment is the type of assessment that occurs in the process. In this respect, it can be said that the positive error climate is also a part of the formative assessment. Of course, seeing incorrect answers on the exam paper during the complementary assessment phase is annoying and causes a drop in the student's grade. However, until the complementary evaluation stage, managing errors in a positive way in the classroom and including them in the teaching process in a planned way will provide students with positive affective characteristics.

5. RECOMMENDATION

In this study, in which the effect of positive error climate on affective characteristics was examined, positive significant differences were found in the experimental group. ECS and MAS were used to measure the effectiveness of the positive error climate applied in the classroom. In future studies, positive error climate can be examined in more detail by increasing the number of practice lesson hours, teachers, and classes using various scales. The application was limited to the mathematics course. The effect of the positive error climate in other courses may be the subject of further research.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Atatürk University, 10.12.2021-2021/13/12

Authorship Contribution Statement

Merve Ozkaya: Investigation, Methodology, Materials, Data collection, Analysis, Literature Review, Writing. **Senem Kalac**: Investigation, Data collection, Analysis, Writing. **Alper Cihan Konyalioglu**: Methodology, Supervision and Validation.

Orcid

Merve Ozkaya https://orcid.org/0000-0002-0436-4931
Senem Kalac https://orcid.org/0000-0003-1636-977X
Alper Cihan Konyalioglu https://orcid.org/0000-0002-6009-4251

REFERENCES

Akınoğlu, O. (2004). Sınıfta grup etkileşimi [Group interaction in the classroom]. In Z. Kaya (Ed.), *Sınıf yönetimi* [*Class managemen*t] (pp. 111-130). Pegem Akademi.

Akkuşci, Y.E. (2019). *Investigation of the effectiveness of using mistake – handling activity applications in classroom in mathematics teaching* [Master's dissertation, Ataturk Unıversıty]. YÖK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Akpınar, B., & Akdoğan, S. (2010). Negative knowledge concept: Learning from mistakes and failures. *The Western Anatolia Journal of Educational Sciences, 1*(1), 14-22. https://dergipark.org.tr/en/pub/baebd/issue/3342/46246

Avcı, E., Coşkuntuncel, O., & İnandı, Y. (2011). Attitudes of twelfth grade students towards mathematics. *Mersin University Journal of the Faculty of Education, 7*(1), 50-58.

Baki, A. (2015). *Kuramdan uygulamaya matematik eğitimi [Mathematics education from theory to practice]* (6th ed.). Harf Eğitim Yayıncılığı.

Barbieri C.A., & Booth J.L. (2020). Mistakes on display: Incorrect examples refine equation solving and algebraic feature knowledge. *Applied Cognitive Psychology, 34*(1)*,* 862-878. https://doi.org/10.1002/acp.3663

Baştürk, S. (2014). Matematik öğretiminde öğrenci hatasının yeri: Hata ve engel kavramı [The place of student error in mathematics teaching: The concept of error and obstacle]. *Bilim ve Aklın Aydınlığında Eğitim*, *166*(1), 14-23. https://www.researchgate.net/publication/312295181

Borasi, R. (1986). *On the educational roles of mathematical errors: Beyond diagnosis and remediation* [Doctoral Dissertation, State University of New York]. Ataturk University Libraries. https://www.proquest.com/pagepdf/303527036?accountid=8403

Borasi, R. (1988, April). *Towards a reconceptualization of the role of errors in education: The need for new metaphors* [Conference session]. Annual Meeting of the American Educational Research Association. https://eric.ed.gov/?id=ED295969

Borasi, R. (1989, March). *Students' constructive uses of mathematical errors: A taxonomy* [Conference session]. Annual Meeting of the American Educational Research Association, https://eric.ed.gov/?id=ED309069

Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": A teaching experiment. *Journal for Research in Mathematics Education, 25*(21), 166-208. https://doi.org/10.5951/jresematheduc.25.2.0166

Borasi, R. (1996). *Reconceiving mathematics instruction: A focus on errors*. Ablex Publishing Corporation.

Bray, W.S. (2011). A collective case study of the influence of teachers' beliefs and knowledge on error-handling practices during class discussion of mathematics. *Journal for Research in Mathematics education, 42*(1), 2-38. https://doi.org/10.5951/jresematheduc.42.1.0002

Bray, W., & Santagata, R. (2013, January 24-26). *Developing teaching capacity for making productive use of mathematical errors* [Conference session]. Association of Mathematics Teacher Educators Annual Meeting, Orlando Florida, USA.

Büyüköztürk, Ş. (2020). *Sosyal bilimler için veri analizi el kitabı; İstatistik, araştırma deseni SPSS uygulamaları ve yorum [Manual of data analysis for social sciences; Statistics, research design SPSS applications and interpretation]* (28th ed.). Pegem Akademi.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

Didiş, M. G., Erbaş, A. K., & Çetinkaya, B. (2016). Matematik öğretmen adaylarının öğrenci hatalarına yönelik pedagojik yaklaşımlarının matematiksel modelleme bağlamında incelenmesi [Investigating prospective mathematics teachers' pedagogical approaches in response to students' errors in the context of mathematical modeling activities]. *Elementary Education Online, 15*(4), 1367-1384. https://doi.org/10.17051/io.2016.75429

Didiş-Kabar, M.G., & Amaç, R. (2018). Investigating pre-service middle-school mathematics teachers' knowledge of student and instructional strategies: An algebra case. *Bolu Abant İzzet Baysal Journal of Education, 18*(1), 157-185. https://doi.org/10.17240/aibuefd.2018..-359810

Ekiz, D. (2009). *Bilimsel araştırma yöntemleri [Scientific research methods]* (2nd ed.). Anı Yayıncılık.

Erbaş, A.K., Çetinkaya, B., & Ersoy, Y. (2009). Student difficulties and misconceptions in solving simple linear equations. *Education and Science, 34*(152), 44-59. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/7

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw Hill.

Gedik, S. D. (2014). *Effect of mistake-handling activities to mathematics content knowledge development process* [Doctoral dissertation, Ataturk University]. YÖK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Guzmán-Muñoz, F.J., Gruber, H., Heid, H., Lorenzer, K., Bauer, J., Heinze, A., Tuerling, J.M., Nägele, C., Baumgartner, A., Anja, M., Seifried, J., Link, M., Gartmeier, M., & Wuttke, E. (2009, September 25-30). Learning from errors [Conference session]. ECER, Vienna, Austria. https://eera-ecer.de/ecer-programmes/conference/2/contribution/2488/

Haser, Ç., & Ubuz, B. (2000, September 6-8). *İlköğretim 5. sınıf öğrencilerinin kesirler konusunda kavramsal anlama ve işlem yapma becerileri [Elementary school 5th grade students' conceptual understanding and processing skills about fractions].* IV. Congress of Science Education, Ankara, Turkey.

Heinze, A. (2005). *Mistake-handling activities in German mathematics classroom*. In H. L. Chick & J. L. Vincent (Eds.), Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education (Vol. 3, pp. 105- 112). PME.

Heinze, A., & Reiss, K. (2007). *Mistake-handling activities in the mathematics classroom: Effects of an in-service teacher training on students' performance in geometry*. In J.-H. Woo, H.-C. Lew, K.-S. Park, & D.-Y. Seo (Eds.), Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education (Vol. 3, pp. 9-16). PME.

Huelser, B.J., & Metcalfe, J., (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40,* 514-527. https://sci-hub.mksa.top/10.3758/s13421-011-0167-z

İlhan, M. (2017). The predictive role of students' perceptions of classroom assessment environment on their attitudes towards school. *Kastamonu Education Journal, 25*(1), 111-128. https://dergipark.org.tr/tr/pub/kefdergi/issue/27737/308251

Ingram, J., Pitt, A., & Baldry, F. (2015). Handling errors as they arise in whole- class interactions. *Research in Mathematics Education, 17*(3), 183-197. https://doi.org/10.1080/14794802.2015.1098562

Kalaç, S., & Özkaya., M (2021). Sınıf içi olumlu hata iklimi uygulamaları [Positive error climate practices in the classroom]. In A. Kızılkaya Namlı (Ed.), *Eğitimin kavramsal temelleri 4: Yöntem ve stratejiler [Conceptual basics of education 4: Methods and strategies]* (pp. 231-244). Efe Akademi.

Kalaç, S., Özkaya, M., & Konyalıoğlu, A.C. (2022). *The adaptation of the perceived error climate scale into Turkish*. *Educational Academic Research, 44*(1), 100-109. https://doi.org/10.54614/AUJKKEF.2022.11-22

Karadağ, Z. (2004, November 24-26). *Hatalardan öğrenme yönteminin bilgisayar destekli matematik 664 öğretiminde uygulanması (Koordinat düzlemi ve simetri konusu) [Application of learning from mistakes method in computer assisted mathematics teaching (Coordinate plane and symmetry issue)].* IETC 4th International Educational Technology Conference, Sakarya, Turkey. https://www.iet-c.net/publication_folder/ietc/ietc2004.pdf

Kaya, R. (2015). *Investigation of the 6th grade students' misconceptions about representations of the decimal numbers* [Master's thesis, Uşak University]. YÖK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Kohen, L. (2006). *Student and teacher expectations in creating a suitable classroom environment for effective classroom management, an application from universities in Istanbul* [Master's thesis, Yeditepe University]. YÖK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

McMillan, J.H. (2015). *Sınıf içi değerlendirme, etkili ölçütlere dayalı etkili bir öğretim için ilke ve uygulamalar* (A. Arı, Trans.). Eğitim Yayınevi (Original work published in 2014).

McMillan, J.H., & Schumacher, S. (2010). *Research in education: Evidence-based inquiry* (7. Ed.). Pearson.

McMillan, J.H., & Workman, D.J. (1998). *Classroom assessment and grading practices: A review of the literature. Metropolitan Educational Research Consortium* (ERIC Document Reproduction Service No. ED453263). https://files.eric.ed.gov/fulltext/ED453263.pdf

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*(1), 465-489. https://doi.org/10.1146/annurev-psych-010416-044022

National Council of Teachers of Mathematics [NCTM]. (2000). *Principles and standards for school mathematics.* NCTM.

Öcalan, T. (2004). *İlköğretimde matematik öğretimi [Mathematics teaching in primary education*]. Yeryüzü Yayınevi.

O'Dell, S. (2015). *Classroom error climate: Teacher professional development to improve student motivation* [Doctoral thesis, University of Central Florida]. Electronic Theses and Dissertations. https://stars.library.ucf.edu/etd/704/

Önal, N. (2013). A study on the development of a middle school students' attitudes towards mathematics scale. *Elementary Education Online, 12*(4), 938-948. http://ilkogretim-online.org.tr

Oser, F., & Spychiger, M.B. (2005). *Lernen ist schmerzhaft: Zur theorie des negativen wissens und zur praxis der fehlerkultur.* Beltz-Pädagogik.

Özdamar, K. (2018). *Eğitim, sağlık ve sosyal bilimler için SPSS uygulamalı temel istatistik [SPSS applied basic statistics for education, health and social sciences].* Nisan kitapevi.

Özkaya, M. (2015). *A study on the impact of mistake-handling activities on mathematics* 697 *teachers' professional development* [Doctoral thesis, Ataturk University]. National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Özkaya, M., & Konyalıoğlu, A.C. (2019). Mistake handling activities in the development of middle school mathematics teachers' subject matter knowledge: Addition operation with fractions. *Journal of Bayburt Education Faculty, 14*(27), 23-52. https://doi.org/10.35675/befdergi.475076

Palkki, R., & Hastö, P. (2018). Mathematics teachers' reasons to use (or not) intentional errors. *Teaching Mathematics and Computer Science, 6*(2), 263-282. http://www.problemsolving.fi/pp/intentionalErrors.pdf

Rach, S., Ufer, S., & Heinze, A. (2013). Learning from errors: Effects of teachers training on students' attitudes towards and their individual use of errors, *PNA, 8*(1), 21-30. http://hdl.handle.net/11162/101713

Rittle-Johnson, B., & Star, J.R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology, 101*(3), 529–544. https://doi.org/10.1037/a0014224

Santagata, R. (2004). "Are you joking or are you sleeping?" Cultural beliefs and practices in Italian and U.S. teachers' mistake-handling strategies. *Linguistics and Education, 15*, 141–164. https://doi.org/10.1016/j.linged.2004.12.002

Son, J.W. (2013). How preservice teachers interpret and respond to student errors: Ratio and proportion in similar rectangles. *Educational Studies in Mathematics, 84*(1), 49–70. https://doi.org/10.1007/s10649-013-9475-5

Son, J.W., & Sinclair, N. (2010). How preservice teachers interpret and respond to student geometric errors. *School Science and Mathematics, 110*(1), 31-46. https://doi.org/10.1111/j.1949-8594.2009.00005.x

Soncini, A., Matteucci, M.C., & Butera, F. (2021). Error handling in the classroom: An experimental study of teachers' strategies to foster positive error climate. *European Journal of Psychology of Education, 36*(3), 719-738. https://doi.org/10.1007/s10212-020-00494-1

Steuer, G., Rosentritt-Brunn, G., & Dresel, M. (2013). Dealing with errors in mathematics classrooms: Structure and relevance of perceived error climate. *Contemporary Educational Psychology, 38,* 196–210. https://doi.org/10.1016/j.cedpsych.2013.03.002

Tezbaşaran, A. (1997). *Likert tipi ölçek geliştirme kılavuzu* [*Likert type scale development guide*]. Türk Psikologlar Derneği.

Trisha, M. (1999). Changing student attitudes towards mathematics. *Primary Educator, 5*(4), 2–6.

Tulis, M. (2013). Error management behavior in classrooms: Teachers' responses to student mistakes. Teaching and Teacher Education, 33, 56-68. https://doi.org/10.1016/j.tate.2013.02.003

Türkdoğan, A., & Baki, A. (2012). Primary school second grade mathematic teachers' feedback strategies to students' mistakes. *Ankara University Journal of Faculty of Educational Sciences, 45*(2), 157-182. https://doi.org/10.1501/Egifak_0000001258

Turkish Ministry of National Education [MoNE]. (2018). *Matematik dersi öğretim programı (İlkokul ve ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. sınıflar) [Mathematics lesson curriculum (1, 2, 3, 4, 5, 6, 7 and 8th grades)]*. MoNE.

Yenil, T. (2020). *The correction of 6th-grade students' misconceptions on decimal notation with digital concept cartoons designed according to the 5E model* [Master's dissertation, Bartın University]. YÖK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Yıldırım, İ. (2019). *The effect of erroneous solution method on the achievement of some istatistical concepts of 7th grade students* [Master's dissertation, Adiyaman University]. YÖK National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

APPENDIX

POSITIVE EROR CLIMATE FRAMEWORK PROGRAM

The teacher expresses her/his tolerance towards mistakes verbally and in behavior.

The feedback that can be given as follows:

Verbal feedback:

- Answer even if you think you are wrong.
- Errors are ways that are not right, the m ore wrong ways we eliminate, the better.
- All mistakes are ways that will bring us closer to the truth.
- Making mistakes and giving wrong answers are inevitable in the classroom environment.
- You are a student, of course you will make mistakes to find the truth, do not hesitate.
- Do not think that I will be angry with you if you make a mistake.
- You are all classmates, let's try to learn a lesson instead of laughing or getting angry at wrong answers.

Behavioral feedback:

- S/he encourages students with low attendance and who are behind the class academically to get up and respond to the lesson.
- S/he encourages the student, who is hesitant and does not want to get up, to participate in the lesson and encourages them to respond.
- S/he asks students to answer even if they are wrong.
- S/he asks the students who make mistakes why they think that way without getting angry.
- Be tolerant towards student mistakes.

Associated sub-dimensions:

✓ *Errors tolerance by the Teacher*

✓ *Absence of negative teacher reactions to error.*

The teacher is tolerant of the student who makes an error or gives an incorrect answer, thanks him/her for the error s/he finds and turns students' attention to that error.

***The feedback that can be given as follows*:**

- Why did you think like that?
- Your friend gave a very nice answer.
- Shall we think together?
- Your friend has mentioned a very good mistake, let's be careful about it.
- Thanks for your friend's reply.
- Well done, you have caught a very important point -to class- do you think your friend's answer is correct?
- If it's wrong, let's think about why it's wrong.
- You gave a very good answer. Thank you.

Associated sub-dimensions:

✓ *Functionality of errors for learning*

✓ *Analysis of errors*

✓ *Taking the error risk*

✓ *Absence of negative teacher reactions to errors*

✓ *Teacher support following errors*

✓ *Irrelevance of errors for assessment*

Instead of giving the answer directly, the teacher gives clues to the students. Discusses the given answers in class. Draws students' attention to the given answer.

The feedback on this issue is as follows:

- S/he does not directly say that the mistake made is wrong. Or s/he does not give the correct answer directly to the student.
- S/he asks questions that will help the student find the right answer.
- S/he draws the attention of the students in the class to the mistake made.
- S/he involves the whole class in the process.
- S/he explains the importance of the mistake made by the student.
- S/he gives corrective feedback to the student.

- S/he discusses the student's mistake in class.
- S/he allows students who gave wrong answers to express the correct answer in their own words.

Associated sub-dimensions:

✓ *Functionality of errors for learning*
✓ *Analysis of errors*
✓ *Taking the error risk*
✓ *Absence of negative classmate reactions to errors*
✓ *Absence of negative teacher reactions to errors*
✓ *Teacher support following errors*
✓ *Irrelevance of errors for assessment*

The teacher encourages the student, who is shy and does not want to attend to lesson. S/he enables them to participate in the lesson and promotes them to respond.

The feedback that can be given as follows:

- It does not directly say that the answer given is wrong.
- Asks the students why they gave such an answer.
- Asks the class for the student's answer.
- Makes the students think about their errors.
- S/he thanks the student for the point s/he caught.

Associated sub-dimensions:

✓ *Functionality of errors for learning*
✓ *Analysis of errors*
✓ *Taking the error risk*
✓ *Absence of negative classmate reactions to errors*
✓ *Absence of negative teacher reactions to errors*
✓ *Teacher support following errors*
✓ *Irrelevance of errors for assessment*

After the teacher decides that he has solved enough examples at the end of the subject, he gives an incorrect statement about the subject or makes an incorrect solution and waits for the students to catch the mistake. Ask students to express both the incorrect statement/solution and the correct statement/solution in their own sentences.

The feedback that can be given as follows:

- Let's examine the given statement/solution/question.
- Do you think it is true?
- If it's wrong, why is it wrong.
- If true, why is it true?

Associated sub-dimensions:

✓ *Functionality of errors for learning*
✓ *Analysis of errors*
✓ *Taking the error risk*

At the end of the subject, the teacher exams the students, the exam is not for scoring. Puts a wrong example in the exam. Waits for the student to realize the error. At the end of the exam, he/she solves the questions in detail in the class.

Associated sub-dimensions:

✓ *Analysis of errors*
✓ *Taking the error risk*
✓ *Irrelevance of errors for assessment*

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 258–282

https://doi.org/10.21449/ijate.1116273

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Examination of map reading skills with orienteering activity: An example of Many Facet Rasch Model

Seyma Uyar[1,*], **Onur Yayla**[2], **Hidayet Zunber**[3]

[1]Mehmet Akif Ersoy University, Faculty of Education, Department of Educational Sciences, Burdur, Türkiye
[2]Mehmet Akif Ersoy University, Faculty of Education, Department of Turkish and Social Sciences Education, Burdur, Türkiye
[3]Mehmet Akif Ersoy University, Institute of Educational Sciences, Burdur, Türkiye

ARTICLE HISTORY

Received: May. 13, 2022
Revised: Nov. 3, 2022
Accepted: Nov. 7, 2022

Keywords:
Many Facet Rasch Model,
Rater reliability,
Rubric,
Spatial thinking,
Orienteering.

Abstract: The purpose of the current study is to examine the map reading skills of Social Studies pre-service teachers with orienteering, which is an activity-based and more active practice. To this end, a total of 10 students attending the Department of Social Studies Teaching in the Education Faculty of Burdur Mehmet Akif Ersoy University and taking the course of Map Skills and Applications were selected. An analytical rubric consisting of four criteria and scored in four categories was used to collect data in the study. The content validity of the developed rubric was calculated with the Davis Technique and it was thought that sufficient evidence was obtained for the content validity. During the orienteering activity, the map reading skills of the students were scored by 5 raters with this rubric in terms of four criteria, direction/location, recognizing signs/symbols, using landforms and managing time. They were examined with the many-facet Rasch model (MFRM). Map reading skills were evaluated according to the severity/leniency of the raters and the difficulty of the students in exhibiting the behavior. The results of the analysis showed that the agreement between the raters was found to be good. It was also concluded that the most difficult skill is determining direction/location and the easiest skill is using landforms.

1. INTRODUCTION

It is very important for students to gain map reading skills in terms of making sense of the space (Kızılçaoğlu, 2007). This is because space refers to places where people conduct activities and gain experiences. It has different meanings according to the way it is perceived and evaluated by the individuals living in it (Tümertekin *et al*., 2019, p.49). One of the indispensable indicators of perceiving the space and constructing it in the best way is the skill of reading a map (Kızılçaoğlu, 2007). In this sense, spatial perception is an important step for individuals in the concrete interpretation and evaluation processes and is presented as a skill in the educational environment. Subjects related to the perception of space in the elementary education period are generally included in the scope of social studies courses (Sönmez, 2010). Given that the

*Corresponding Author: Seyma Uyar ✉ syuksel@mehmetakif.edu.tr Burdur Mehmet Akif Ersoy University, Faculty of Education, Department of Educational Sciences, Burdur, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

attitudes and behaviours acquired in this period will be the basis for students throughout their life, it can be seen that social studies courses have an important area of influence in the development of spatial perception (Öcal, 2007). In order for students to acquire basic information such as determining routes and directions and locations, comprehending the geographical information in the place where they are located, and adapting to the place they live in, their spatial perceptions should be improved (Safi, 2010).

To make sense of and use spatial perception the most basic and most used tools are maps (Ertuğrul, 2008). Therefore, it is very important to understand and interpret maps as educational materials (Dong *et al*., 2018). In the social studies course; materials such as maps, graphics and tables are expected to be used in terms of using, organizing and developing the information. In order to understand the given information easily, students should be able to read maps, graphics and tables (Kıroğlu, 2006; Pala & Başıbüyük, 2019). A map is a tool that is frequently used not only in teaching the subjects in the social studies course, but also in daily life (Abbak, 2021). For example, the coordinate system is the basis of navigation used by many groups of people such as travelers, hikers, and mountaineers, to reach their destinations. In addition, the excess of opportunities provided by unmanned aerial vehicles and satellites and the transition from paper maps to maps with digital content (Carbonell-Carrera & Bermejo Asensio, 2016; Carbonell-Carrera *et al*., 2017) has increased the inclination of the cognitive field experts interested in geography, psychology and spatial thinking (Bednarz *et al*., 2006; Newcombe *et al*., 2013). In today's world, it is more important for individuals to know how to do it than to know everything. For this reason, it is expected from schools to raise the number of individuals who enjoy learning, creating, producing, thinking critically, and making connections between events (Pala & Başıbüyük, 2019. The importance of this issue is noticed with the intensity of the studies on map skills in the field of social studies (Aksoy & Ünlü, 2012; Aktürk *et al*., 2013; Alım & Girgin, 2012; Bahar *et al*., 2010; Buğdaycı & Bildirici, 2009; Darakçı, 2014; Demirci *et al*., 2013; Güneş & Öztürk Demirbaş, 2020; İncekara *et al.,* 2008; Kaymakçı, 2015; Kızılçaoğlu, 2007; Kızılçaoğlu & Ünlü, 2008; Koç & Karatekin, 2016; Koç *et al*., 2017; Özcan & Uzun, 2016; Sönmez & Aksoy, 2012; Sönmez & Aksoy, 2013; Taş, 2006; Taşlı *et al*., 2007).

When the studies are examined, it can be said that the mapping skills of the students are at a moderate level, therefore, there is a need for applications and materials that can improve their mapping skills. Akengin *et al*. (2016), in a study in which Social Studies teachers' opinions were taken, stated that teachers use narrative and question-answer methods to improve their mapping skills in lessons. In the study, he stated that in the 21st century, in accordance with the constructivist approach, more active learning methods should be included in which more students will be active. Orienteering can be considered one of these applications because orienteering, due to its nature, has the potential to provide map reading skills within the scope of geographical skills while having fun and racing (Arıkan & Aladağ, 2019).

1.1. Map and Map Skills

Maps, graphs and diagrams take an active role in the design and presentation of learning in today's education system. (Schnotz & Kulhavy, 1994). In addition, it is seen that maps are also used in the visual presentation of certain elements and data in many fields such as industry, politics, tourism, agriculture, etc. (Sarıgül, 2021). Due to the wide coverage of the maps and the differentiation in the way each area of expertise uses maps, it makes it very difficult to come up with a common definition of the map. According to ICA (International Cartographic Association), a map is a symbolized representation of geographical reality, representing selected features or characteristics resulting from the creative effort of its author’s execution of choices, and is designed for use when spatial relationships are of primary relevance (URL 1).

Map skills are classified with different terms in the literature (Borich & Bauman, 1972; Carbonell-Carrera & Medler, 2017; Stumpf & Eliot, 1999;). With the use of maps and different forms of cartographic representation and geo-referenced information, spatial orientation has become the most widely used one of these terms (Carbonell-Carrera & Medler, 2017). Spatial orientation is defined as the ability to navigate physically or mentally (Carbonell-Carrera & Medler, 2017; Maier, 1996;). Another commonly used term is spatial thinking (Atayeter *et al*., 2018; Bednarz, 2001; Gersmehl & Gersmehl, 2007; Jo, 2011; Lee, 2005; Sönmez, 2019, p. 219; Şanlı, 2021; Şanlı & Sezer, 2019). The report prepared by the National Research Council (NRC) has made an important contribution to the formation of the theoretical background of this subject. In this report, spatial thinking is defined as "a skill consisting of spatial concepts, representation tools and cognitive processes" (Şanlı, 2021). Spatial concepts refer to the terminology used for the description, perception and association of objects (Jo, 2007; Jo & Bednarz, 2014a; 2014b). The concepts frequently used in this terminology are "location, map, region, distribution, information, scale, navigation, symbology, coordinate, distance, area, direction, geographic data, overlay, buffer, contour, aspect" (Huynh & Sharpe, 2013; Şanlı, 2019; 2020; Ünlü & Yıldırım, 2017). Sönmez (2010) classified mapping skills under six sub-headings from concrete skills to abstract skills. These are the ability to understand and interpret symbols, to read and interpret maps, find directions, determine location coordinates, to use scales and measure distances.

1.1.1. *The ability to understand and recognize symbols*

Maps contain a whole consisting of points, lines and symbols (Sönmez, 2010). Various colors and symbols are also used while creating maps (Ünlü, 2021, p. 388). This whole consisting of colors and symbols is called the symbology of the map (Wiegand, 2006, p. 10). Abstract thinking and generalizations must be made in order to make sense of symbols by individuals (Bednarz *et al*., 2006). Thus, individuals can interpret the information encoded on the map in a whole sense in the context of events, facts and features (Ünlü, 2021, p. 388).

1.1.2. *Map reading and interpretation skills*

Map reading is the process of obtaining simple information from the map as a result of a complex process such as getting information from the map and using the map as a result of this mental process by using map skills at the same time (Sönmez, 2010; Ünlü, 2021, p. 388; Wiegand, 2006, p. 10). Map reading skills are seen as a cognitive process that enables the interpretation of information in the mind by including the information on the map with psychological processes, interests, purposes, abilities and external factors in the process (Koláčný, 1969; Ooms *et al*., 2016).

1.1.3. *The ability to find direction*

While traveling on a little-known or unknown route, the desire to seek in the process between the start and the destination is the ability to find direction (Golledge, *et al*., 2000; Wiegand; 2006, p. 19). It is seen that the wayfinding process occurs in three stages: cognitive mapping, wayfinding planning and movement (Chen & Stanney, 1999). The way-finding process, which occurs as a result of these three stages, is realized by the accumulation of geographical knowledge in the immediate environment of individuals and by systematic knowledge production (Murakoshi, 1997).

1.1.4. *Determination of coordinate position skill*

Location is the holistic evaluation of latitude, longitude, parallel, meridian and equator points together with the numerical and angular value components given on the maps. They are the processes of making inferences by associating the current location on the map and its immediate surroundings (Çepni, 2019, p. 367; Sönmez, 2010; Ünlü, 2021, p. 388; Wiegand, 2006, p. 150).

1.1.5. *The ability to use scale*

Maps are tools that systematically represent the distances between different spaces (Bartz, 1970). The ability to use scales is one of the most important sub-dimensions of the map skill that guides map reading in reaching the right distance and understanding spatial relationships as a result of the ratio of the real distance in the world and the distance on the map (Meyer, 1973; Ünlü, 2021, p. 388; Wiegand, 2006, p. 10).

1.1.6. *Distance measurement skill*

It is the reduction of the distance values given between two or more points by the map ratio, converting or proportioning them to the actual distance with the help of calculations using the map scale (Demiralp, 2006; Ünlü, 2021, p. 388).

1.2. Orienteering

Orienteering is defined as a sports activity in which individuals interpret the cartographic symbols given for a particular terrain, and during this interpretation, skills such as spatial perception, environmental cognition, analytical thinking and critical understanding are used in an integrated manner (Wilson, 2017). Orienteering is actually a branch of sport but it can also be considered an educational game to be used in educational activities. Orienteering not only makes it possible for students to have a good and productive time, but also enables them to develop their geographical skills (Candan, 2019). According to Baitan (2022), it has been emphasized that the use of maps and compasses, map perception and map comprehension skills are more developed in individuals dealing with orienteering from a young age. Orienteering activities, which are effective tools in out-of-school learning environments and map studies (Adams, 1972), also provide students with environments of learning by doing and experiencing, provide the opportunity to achieve objectives set for geography subjects in an enjoyable way and make permanent learning more effective (Candan, 2019).

1.3. The Purpose and importance of the research

It is seen that studies have been carried out on many subjects such as map skills, location analysis, and spatial perception in higher education (Balcı, 2015; Koç & Karatekin, 2016; Özcan & Uzun 2017). In addition, in the literature, it is seen that map reading skills have been assessed mostly by using interview methods (Akkuş & Kuzey, 2018; Balcı, 2015), self-efficacy scales (Özcan & Uzun, 2017), achievement tests (Arıkan & Aladağ, 2019; Koç & Bulut, 2014; Koç & Karatekin, 2016; Sönmez & Aksoy, 2012). In international studies (Atit *et al*., 2016; Ooms *et al*., 2016), it was seen that map skills were examined with optional tests. However, no application has been found in higher education in which map skills activities are evaluated based on performance and scored with rubrics in out-of-school environments. In performance evaluation, the student is expected to create an answer, put forward a product or perform an activity, rather than choosing from predetermined options (Darling-Hammond *et al*., 2010). Since orienteering is defined as a sport that requires finding the targets marked on the map of the same terrain in the shortest possible time using the map and compass in unknown terrain, it can be considered to be related to the concept of performance. For this reason, raters need to make quick decisions in real time in the evaluation of performance. Unless the measurement tools used during the evaluation are objective, specific and reliable, the evaluation and interpretation of the performance remain essentially subjective (Carlin & Louis, 2008). In case of differences in the value judgments of raters, it is inevitable that unreliable scores will emerge in the scores (Baird *et al*., 2013). In order to eliminate this limitation of classical approaches, the researchers suggested using the Many-Facet Rasch Model [MFRM] in cases where there is more than one rater. MFRM is also considered to be a more powerful psychometric model than Classical Test Theory in terms of features such as determining the interactions between different error sources (Haiyang, 2010) and taking into account more than one error source at the same time. It also provides information at the individual level rather than the group level for raters or students (Barkaoui, 2008).

In the current study, it is aimed to evaluate the map skills of the students studying in the Social Studies Teaching Undergraduate Program during the orienteering practice The skills were scored with a rubric and analysed with MFRM. In this context, the consistency, severity and leniency of more than one rater and the skills that students had difficulty in reading maps were examined. The limited number of studies worldwide, especially at the higher education level (Ooms *et al*., 2016), makes this study important. It is thought that this activity and performance-based study will contribute to the field in terms of providing measurement and evaluation opportunities in out-of-school environments. In addition, it is anticipated that the study will attract attention, since no study has been found on the use of MFRM in the evaluation of performance in the field of Social Studies. For this purpose, the questions to be answered in the research are as follows:

1) Which skills are difficult and easy for students in terms of map reading with oriente-ering?
2) What is the severity and leniency behavior of the raters in the evaluation of map rea-ding skills with orienteering?
3) What is the central tendency behaviour of the raters in the criteria taken into consi-deration for map reading skills?
4) What is the biased behavior of the raters?

2. METHOD

In this study, it was aimed to examine the map reading skills of Social Studies pre-service teachers with the Many-Facet Rasch Model. For this purpose, the skills in which the students had difficulty and the scoring behaviors of the raters were examined. Therefore, this is a descriptive study in which the existing situation is tried to be described (Büyüköztürk *et al*., 2019; Karasar, 2005).

2.1. Study Group

The study group of the current research is comprised of a total of 10 students attending the Department of Social Science Education in the Education Faculty of Burdur Mehmet Akif Ersoy University and taking the course of Map Skills and Applications. The ethical committee approval was obtained from the Non-interventional Clinical Research Ethics Committee at Burdur Mehmet Akif Ersoy University (GO 02/2022/472). The participants were randomly selected from among the students who take the Map Knowledge and Applications course in the Department of Social Science Education. Of the participating 10 students, 5 (50%) are females and 5 (50%) are males. In this study, orienteering activities were conducted in an area of approximately 5.7 hectares, where landforms were densely located, rather than a school garden or classroom due to the age level of the participants. The study was limited to 10 students because the area was large, and it took a long time to complete the track for each student and to prepare the next student. In the current study, 5 raters were included to evaluate map reading skills with orienteering activities. The raters are Social Science Education and Geography Department instructors and an orienteering specialist.

2.2. Data Collection Tool

In the current study, an analytic rubric was used. The use of an analytical rubric is recommended where the attribute to be measured can be broken down into components. In map reading, individuals are expected to be able to interpret, analyze and evaluate by establishing a relationship with the place on the map based on the signs (legends) and symbols on the map (Sönmez, 2010, p.105). This skill is important for individuals to perceive the space and establish a space-event connection (Akengin *et al*., 2016). At the same time, the speed in understanding the map can be accepted as an indicator of the development of this skill. Speed is an important factor in map reading skills (Lobben, 2007). The studies on speed in map reading skills (Lobben, 2007) show that its effect on spatial orientation and positioning is also examined by making evaluations on eye movement measurements for the development of speed (Dong *et al*., 2018). In this respect, map literacy has a structure suitable for division into components. Therefore, it was found to be suitable to develop an analytic rubric in the current study.

In this study, a track was prepared to evaluate the map reading skills of the students through orienteering activities. In order to complete the track, the students were expected to reach a total of 5 targets. During this process, the students were expected to demonstrate basic skills in map literacy such as holding the map, finding location/direction, recognizing signs and symbols, using landforms, and managing time. In the literature, it is stated that individuals employ some competencies, when they come up with the map. These skills are map reading and interpretation (Ooms *et al*., 2016; Ünlü, 2021, p. 388), making sense of signs and symbols (Sönmez, 2010; Ünlü, 2021, p. 388), finding the direction (Golledge *et al*., 2000), coordinate and location determination through landforms (Çepni, 2019; Wiegand, 2006, p. 150), using the scale and measuring the distance (Ünlü, 2021, p. 388). Therefore, these skills were considered as the criteria expected from the students and a rubric was prepared accordingly. In addition to ensure the validity of the criteria, a pilot study was conducted with 2 different students who took the course before. During the pilot study, the competencies were classified by taking expert opinions according to the skills used by the students.

In the selection of the track, attention was paid to selecting the points where the students could apply their geographical skills, recognize the landforms in the area and calculate the distance between the targets, and the professional orienteering map drawn by the Orienteering Burdur Provincial Representative was used as the map. Legal permissions were obtained from the relevant unit for the use of the map. The students' ability to reach the targets by using the map along the track was scored with the rubric. The most important advantage of this tool is that it provides detailed information for each performance component.

The criteria to be used in the study are listed as items:

1) Basic skills (holding map, finding location/direction)
2) Recognizing signs and symbols (legends)
3) Using landforms and
4) Managing time (This skill was added as a criterion due to the nature of both map reading and orienteering).

In this study, it was decided to use 4 degrees in order to prevent overlap between degrees in the rubric and to reveal the difference between students. The lowest attribute regarding performance was defined as 1 point (beginner level), and the highest attribute was defined as 4 points (fully successful).

2.2.1. *Taking expert opinion and pilot application*

The prepared draft rubric was sent to 4 experts (1 specialized in the field of Social Science Education, 2 specialized in the field of Geography, and 1 specialized in the field of Turkish Education). In line with the suggestions from the experts, some corrections were made on some attributes and the rubric was given its final form. Afterwards, 2 students who had previously taken a map skills course were observed in the orienteering track and were scored with the draft rubric by two raters. In line with the results obtained from the pilot application, some corrections were made to the rubric. Opinions about the criteria and attribute in the rubric were received from 7 experts, including 1 measurement and evaluation expert, 3 geography faculty members, 1 orienteering specialist, 1 social studies faculty member and 1 social studies graduate student. The experts were asked to evaluate the criteria and attributes in the following four categories;

1) The item represents the attribute measured

2) The item needs minor revision

3) The item needs major revision and

4) The item does not represent the attribute.

Davis' (1992) technique was considered for the content validity index (CVI). The CVI values were obtained by dividing the number of respondents to the 1st and 2nd categories among the experts by the total number of experts. As the CVI value was found to be higher than 0.80, it was concluded that the content validity is acceptable (Davis, 1992; as cited in Yurdugül, 2005). A minimum of 3 and a maximum of 20 experts are recommended for this technique.

Table 1. *Content validity index (CVI) of the rubric.*

Criteria	CVI
Basic skills	1.00
Recognizing signs and symbols	0.85
Using landforms	1.00
Managing time	1.00

According to Table 1 as a result of the content validity study, the experts mostly evaluated the items in the first or the second category. Only in the criterion of using landforms, corrections in the third category were suggested by an expert for the 2nd attribute, so the CVI value was found to be lower than the other criteria. However, since the CVI values were above 0.80, it was thought that sufficient evidence was obtained for the content validity of the rubric, and it was decided to include the criteria and attributes in the rubric. At this stage, minor revisions were made in line with the suggestions.

2.3. Data Collection Process

In order to examine the orienteering map reading skills of the pre-service social studies teachers, a track was prepared in the region located in the Burdur Central City Forest. Orienteering is a sport that aims to reach the targets positioned on the map of a place which is previously unknown with the help of a map and compass as soon as possible (Tanrıkulu, 2011). It is among the skills that individuals should have in line with some daily needs such as reading maps or finding a location on the map or reaching the target by finding direction in the land (Tuna & Balcı, 2013; Ünlü & Yıldırım, 2017). For this purpose, each student was taken to the track one by one. Any student did not see another student going to the track. Five raters observed the student who went on the track by moving simultaneously with the student on the track. The raters were not affected by each other at this stage; they only observed the student's state of reaching the targets and scored at the same time. In this study, the ideal time to complete the organized track was taken to be 18 minutes. This time was determined in the pilot study by the experts by taking the average of the students' time to complete the track.

The reason for the selection of this area was that it contained examples of many of the landforms. There were examples of landforms such as valleys, hills, streams and ridges in the area. The presence of these landforms in the area is important in terms of map reading skills, as they facilitate the determination of the reference point. In addition to using the landforms, the students were expected to complete the track according to the targets on the map by reasoning on the basis of the signs and symbols on the map, making quick decisions and determining the distance. Speed tests in map reading (Dong *et al*., 2018; Lobben, 2007) are important for decision-making, spatial perception and orientation through space. When there is no time limit, it is possible for students to reach the end of the track knowingly or unknowingly. Considering this factor, a time limit was set for the completion of the track, and they were expected to reach the final target within this time limit.

Before implementation, the raters were trained by researchers on the use of analytical rubrics to ensure the validity and reliability of the scoring. At this stage, the use of criteria and attributes in the analytical rubric and the errors that may interfere with scoring were emphasized. Another point considered in the study is that raters do not communicate with students. Since 5 raters followed the students at the same time and the raters did not know the students, it was thought that the error regarding bias was reduced as much as possible.

2.4. Data Analysis

In this study, MFRM was used to analyze student ability, the difficulty of tasks, severity/leniency and bias of the raters in the analysis of data scored with a rubric. Three facets; student (n=10), task (n=4) and rater (n=5), were determined for the analysis of the data. For MFRM, Minifac (FACETS) program was used.

2.4.1. *Facet Rasch Model (MFRM)*

MFRM (Facet model, Linacre, 1994, p. 129) is a measurement model which is an extension of the one-parameter Rasch model and which enables a detailed analysis of the variables that may have a potential impact on testing or assessment. MFRM models the score given to the student as a function of more than one variable. In this respect, it is similar to regression models. In this model, each of the sources of variability (facets) that affects the performance of individuals, such as the student's probability of success in an item, the individual's ability, the difficulty of the item, or the severity/leniency of the rater, is included in the model as an independent variable (Randall *et al*., 2009). The model is expressed by the following formula for three facets (individual, task and rater) (Eckes, 2009; Linacre, 2021; Randall *et al*., 2009).

$$\ln\left(P_{nijk}/P_{nij(k-1)}\right) = B_n - D_i - C_j - F_k \qquad (2)$$

P_{nijk}: the probability that the individual n will get the score k from the rater j in task i,

$P_{nij(k-1)}$: the probability that the individual n will get the score k-1 from the rater j in task I,

B_n: ability of the individual n,

D_i: item difficulty level of task i,

C_j: severity level of rater j,

F_k: difficulty of scale category k relative to scale category k-1.

MFRM summarizes the scoring patterns as the main effects of rater, task, individual, and other facets, if any. In this model, the contribution of each facet and whether it works as expected or not can be examined independently of the other facets. MFRM can show the effects of different elements on the facets at the individual level (Myford & Wolfe, 2003).

In more detail, it can provide information about which raters are more severe or lenient, which raters do not use the scoring criteria consistently, and which tasks are more difficult to score. At the same time, with the bias analysis, MFRM answers the questions of whether the rater's severity is constant in the subgroups, whether it changes with time, and whether the severity changes according to the rater type and the task/item type.

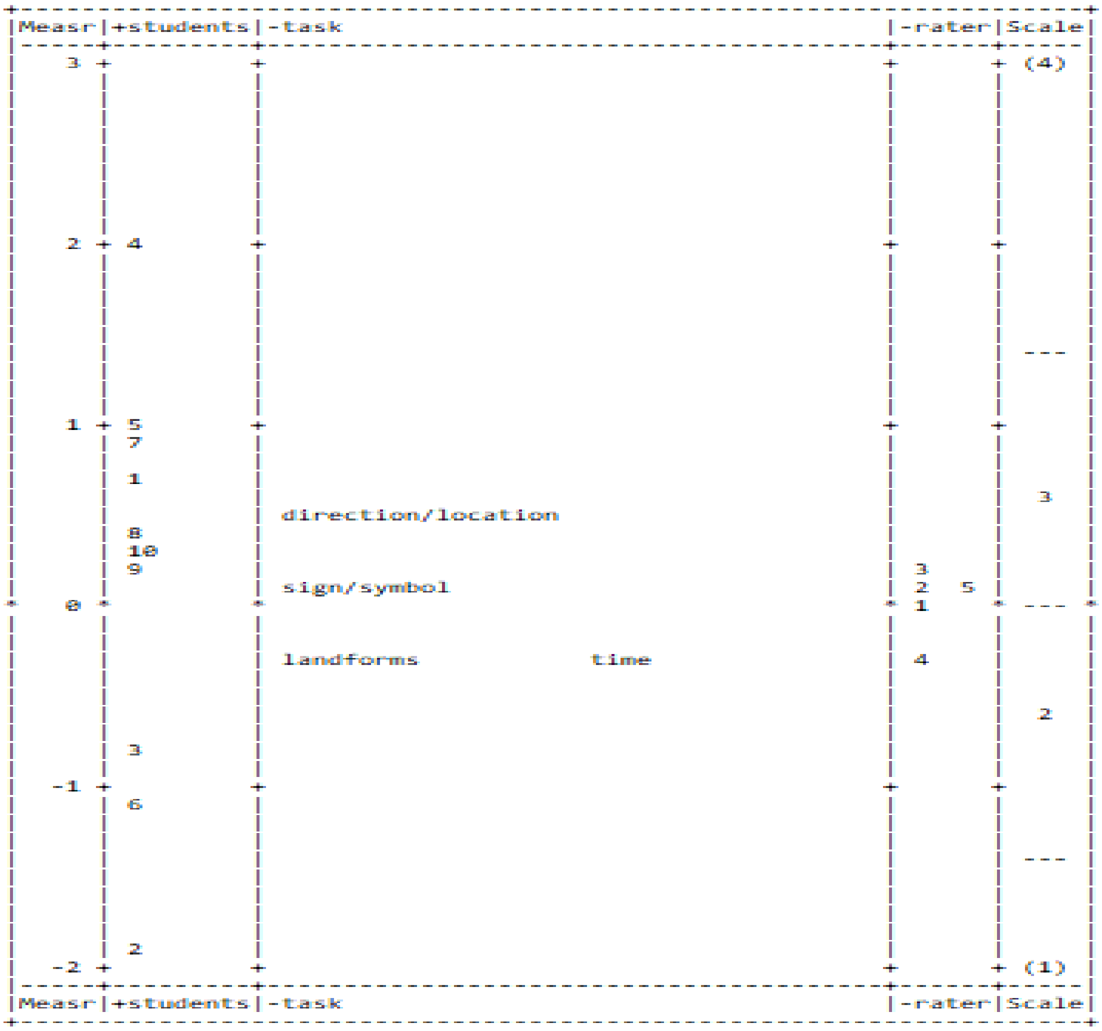
MFRM can cover many models such as rating scale, partial credit, the linear logistic, and mixed Rasch (Eckes, 2009). The Rating Scale Model (RSM; Andrich, 1978) was used in the current study.

3. FINDINGS

For model-data fit, less than about 5% of the standardized values (z-score) of the data used in the analysis must be greater than or equal to 2 in absolute value, or less than 1% of them must be greater than or equal to 3 in absolute value (Linacre, 2021). In this study, 0.5% of the total standardized values (standardized residual is 3.2) [1 out of 200 (10 students x 5raters x 4tasks)] are outside +/- 3 intervals. In addition, the ratio of the standardized value outside +/- 2 intervals is 3% (standardized residual are 2.36, -2.79, -2.21, -2.04, -2.63, -2.04) [6 out of 200 (10 students x 5raters x 4tasks)]. Therefore, it can be said that model data fit is achieved. The Facets program theoretically offers a logit scale (variable map) ranging from -∞ to ∞. On this scale, when the rater facet is negatively oriented, positive logit values indicate severe scoring (low score) and negative logit values indicate lenient scoring (high score) for the rater. When an individual facet is positively oriented, positive logit values for individuals indicate high ability. On the other hand, for the negatively oriented item or task facet, higher values indicate more difficult items (Güler, 2014; Randall *et al*., 2009).

In Figure 1, the logit scale obtained from the MFRM analysis is given as “Students”, “Tasks (dimensions of expected behaviours from students)” and “Raters”.

Figure 1. *Logit scale for three facets.*



```
+--------------------------------------------------------------------+
|Measr|+students|-task                                  |-rater|Scale|
|-----+---------+---------------------------------------+------+-----|
|   3 +         +                                       +      + (4) |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|   2 + 4       +                                       +      +     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      | --- |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|   1 + 5       +                                       +      +     |
|     | 7       |                                       |      |     |
|     |         |                                       |      |     |
|     | 1       |                                       |      |     |
|     |         |                                       |      |  3  |
|     |         | direction/location                    |      |     |
|     | 8       |                                       |      |     |
|     | 10      |                                       |      |     |
|     | 9       |                                       | 3    |     |
|     |         | sign/symbol                           | 2  5 |     |
*   0 *         *                                       * 1    * --- *
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         | landforms           time              | 4    |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |  2  |
|     |         |                                       |      |     |
|     | 3       |                                       |      |     |
|     |         |                                       |      |     |
|  -1 +         +                                       +      +     |
|     | 6       |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      | --- |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     |         |                                       |      |     |
|     | 2       |                                       |      |     |
|  -2 +         +                                       +      + (1) |
|-----+---------+---------------------------------------+------+-----|
|Measr|+students|-task                                  |-rater|Scale|
+--------------------------------------------------------------------+
```

In Figure 1, in the column where the students are located, it is seen that student number 4 is the most successful and student number 2 is the most unsuccessful. In the task (criteria/skills) column, which includes expected behaviours in students’ map reading performances, it is seen that the most difficult skill to be performed by students is the "basic skills (holding map/finding direction and location)", followed by the skills of "recognizing signs/symbols and calculating distance". It can be said that the skills most easily performed by students are "using landforms" and "managing time". According to the measurement results regarding the raters, the rater who scored the lowest and exhibited severe scoring behaviour was the 3rd rater, and the rater who scored the highest and showed lenient scoring behaviour was the 4th rater. Measurement reports related to all the facets are given in Table 2, Table 3 and Table 5 for individuals, tasks and raters, respectively.

Table 2. *Measurement report for the students.*

Students	Measure	St. error	Infit	Outfit
4	2.03	.41	.92	.77
5	.97	.28	.78	.77
7	.94	.27	.60	.58
1	.66	.26	.39	**.39**
8	.40	.25	1.35	1.35
10	.28	.25	2.20	**2.22**
9	.16	.25	.75	.76
3	-.83	.28	.38	**.38**
6	-1.08	.30	1.18	1.24
2	-1.89	.40	1.14	1.25
Mean	.16	.29	.97	.97
St. Deviation	1.15	.06	.54	.56
Reliability = .93, Separation index = 3.68, RMSE= .30, Chi-square = 97.7, SD = 9, *p* = .00				

In Table 2, the results of the analysis regarding the map reading skills of the students are listed from the most successful student to the most unsuccessful student. As Table 2 shows, the logit values of students measures vary between -1.89 and 2.03. The student with the highest map reading skill is number 4 and the student with the lowest map reading skill is number 2. The infit and outfit statistics in Table 2 show the degree of fit between the data and the model and the sensitivity to unexpected responses (Kaya Uyanık *et al*., 2019). Infit and outfit values are expected to be 1.00. However, it is stated in the literature that the range between 0.5 and 1.5 is acceptable (Linacre, 2021; Turner, 2003). Accordingly, it can be said that the infit and outfit indices of the students numbered 1, 10 and 3 are outside the specified range and thus, these students did not exhibit an acceptable performance. It is seen that the separation index obtained as a result of the analysis is 3.68 and the reliability index is .93. The reliability of the separation index for facets takes a value between 0 and 1, while the separation index ranges from 1 to infinity. The reliability of the separation index is similar to Cronbach's Alpha coefficient but the interpretation of these values varies according to the facets (Myford & Wolfe, 2003). For the student facet, the reliability of the separation index is expected to be close to 1.0 (Sudweeks *et al*., 2004). Reliability refers to how well the elements are discriminated against for reliable identification of a facet. The separation index refers to the values that show how much the elements on each facet are discriminated. Large differences between structures or elements within a facet provide high reliability of separation coefficients (Randall *et al*., 2009). The separation index and the reliability of this index are interpreted similarly. However, there is no upper limit for the separation index (Myford & Wolfe, 2003). The separation index for the student facet is expected to be large in order to reflect the difference between students (Sudweeks *et al*,2004). The results of the analysis show that the skill levels of the students can be reliably separated from each other. On the other hand, it can be said that there is a significant difference in the map reading skills of the students (χ^2= 97.7, p<.01).

Table 3. *Measurement report for the task (criteria/skills) included in the rubric.*

Criteria	Measurement	St. Error	Infit	Outfit
Basic skills (holding map/direction/location)	.45	.18	1.09	1.26
Recognizing signs and symbols	.15	.18	.96	.96
Managing time	-.28	.18	1.29	1.09
Using landforms	-.32	.18	.60	.58
Mean	.00	.18	.99	.97
St. Deviation	.37	.00	.29	.29
Reliability = .77, Separation index = 1.81, RMSE= .18, Chi-square = 12.9, SD = 3, *p* = .00				

As can be seen in Table 3, the most difficult criterion is "holding map/finding direction-location" (.45 logit). This value can also mean that this is the criterion most severe scored by the raters. According to Table 3, the easiest criterion is "using landforms". Infit and outfit indices for criteria are between 0.5 and 1.5. The separation index was found to be 1.81, and the reliability of the criteria in the rubric in terms of discriminating between the students was found to be 0.77. This measure indicates that raters are reliable to distinguish among criteria and the criteria were not equally challenging to the students (Sudweeks *et al*., 2004). Accordingly, it can be said that the difficulties related to the criteria differ significantly from each other (χ^2=12.9, *p*<.01). The measurement report regarding the scoring categories (1-4) of the analytical rubric is given in Table 4.

Table 4. *Category statistics.*

"	Frequency (*f*)	Percentage (%)	Mean Measurement	Expected Measurement	Outfit
1	40	20	-1.02	-1.07	1.0
2	53	27	-.28	-.21	.9
3	51	26	.50	.51	1.0
4	56	28	1.15	1.11	.9

Table 4 shows that the scoring categories of the task in the rubric were preferred at almost the same rate. The first category was preferred by 20%, the second category by 27%, the third category by 26% and the fourth category by 28%. This shows that there is no behaviour tending towards the centre (not overusing a certain category of the rubric) (Engelhard, 1994; Myford & Wolfe, 2003).

Table 5. *Measurement report for the raters.*

Rater	Measurement	St. Error	Infit	Outfit
3	.18	.20	.88	.78
2	.10	.20	1.04	1.07
5	.10	.20	.98	.89
1	-.04	.20	.95	.95
4	-.34	.20	1.09	1.17
Mean	.00	.20	.99	.97
St. Deviation	.20	.00	.08	.15
Reliability = .05, RMSE = .20, Discrimination index = .23, Chi-square = 4.1, *df* = 4, *p*=.39 Inter-rater agreement opportunities: 400 Exact agreements: 269 = 67.2% Expected: 156.1 39.0%, Rasch-Cohen's Kappa = .46				

In Table 5, it is seen that the most severe rater is number 3 (.18 logit) and the most lenient rater is number 4 (-.34 logit). When the infit and outfit indices are examined, it is seen that they are between 0.5 and 1.5 and they are acceptable. Unlike the individual and item facets, the separation index on the rater facet is expected to be close to zero (Linacre, 2021). The reliability of the separation index, on the other hand, reflects undesirable variability between raters in terms of severity/leniency. It is preferred that the separation index reliability is low for the rater facet (Myford & Wolfe, 2003). When raters don't differ in terms of severity, the rater separation reliability will be close to 0. By contrast, when raters are of a highly dissimilar degree of severity, the rater separation reliability will be close to 1 (Eckes, 2015). In Table 5, the separation index for the rater facet is .23 and reliability is .05. This value indicates that the raters did not score differently from each other. In addition, the Chi-square value regarding whether there is a difference between the raters is not statistically significant (χ^2 = 4.1, p>.05). Accordingly, it can be said that there is no significant difference between the raters in terms of severity/leniency. In Table 4, the observed (67.2%) and expected agreement values (39%) between raters are given. The Rasch-Cohen's Kappa statistic calculated based on the difference between these percentages (Observed%-Expected%)/(100-Expected%) was found to be 0.46. In the Rasch model, these values are required to be close to 0.00. If the Rasch-Cohen Kappa statistic is between 0.2 and 0.4, it can be said that there is a little more agreement between the raters than modelled (Linacre, 2021).

3.1. Bias Interaction

One of the most important advantages of MFRM is that rater biases can be determined by analyzing the interaction effects between all the surfaces included in the study. In this respect, the fact that the Chi-Square value is meaningful and the t-value is outside the range of ±2 are indicators of the differentiated rater severity/leniency behaviour, that is, the rater bias (Eckes, 2009). In this study, the findings of 3 interaction types, rater x student, rater x task and student x task, were given. There were 50 interactions for the rater x student interaction (*N*=5 and *N*=10). The bias results showed that the t-values remained between the ±2 limits, with the smallest and largest -0.72 and 0.83 values, respectively, and that the rater x student interaction was not statistically significant (χ^2 = 9.1, d.f. =50 p>.05). For the rater x task interaction (*N*=5 and *N*=4), it was seen that the t values of a total of 20 interactions were in the range of ±2, the smallest and largest -0.81 to 1.39, and were not statistically significant (χ^2 = 6, d.f. =20 *p*>.05). Finally, it was observed that 5 (39.25%) of the 40 interactions (*N*=10 and *N*=4) for the student x task interaction were outside the ±2 range of the t-value. The interaction results obtained biased were given in Table 6.

Table 6. *The t-values that are meaningful in student x task interaction.*

Student	Task	Expected value	Observed value	Bias	St. error	*t* value
8	Basic skills (holding map/direction/location)	12.10	7	-1.66	.75	-2.20
5	Recognizing signs and symbols	15.67	11	-1.15	.50	-2.28
6	Basic skills (holding map/direction/location)	7.35	12	1.45	.49	2.98
2	Basic skills (holding map/direction/location)	6.13	9	1.47	.57	2.58
10	Managing Time	14.66	5	-3.76	1.44	-2.62

According to Table 6, it is seen that the 8th, 5th and 10th students performed lower than expected (*t* = -2.20, *t* = -2.28 and *t* = -2.62) in “basic skills”, “recognizing signs and symbols” and “managing time” tasks, respectively. The 6th and 2nd students, whose t-values were obtained as 2.98 and 2.58, respectively, performed higher than expected in the task called basic skills.

4. DISCUSSION and CONCLUSION

In this study, it was aimed to evaluate the map skills of Social Studies pre-service teachers scored with a rubric during orienteering activity with the Many Facet Rasch Model. In this context, the data were evaluated according to the severity/leniency of the raters and the difficulty of the students in exhibiting the behaviour.

The result of the MFRM analysis in the study was related to the difficulty level of the criteria in the rubric. It was seen that the most difficult criterion in map reading skills (the most severe scored skill in map reading skills) is "basic skills (holding map, finding direction/location)". The first thing to do in reading maps is to hold the map correctly and place it in the space according to the direction of the map. A map that is not placed according to its direction in the space is difficult to help the individual. It is essential for individuals to be able to determine their exact location in order to make geographical applications in the space. The results obtained from this study showed that students had difficulties in determining direction/location. Tuna *et al*. (2012) similarly stated that individuals in different education, age and gender groups in Türkiye are poor at reading maps, determining the exact location and placing the map in its original position. In addition, Carswell (1971) in his study on map-based information interpretation in the TTMS [Test of Topographic Map Skills] test related the deficiencies and problems in the students' ability to interpret maps with the inadequacy of teaching processes in educational environments. Streeter & Vitello (1986) stated that in map reading skills, students correctly form the direction and route they aim according to situations such as their daily preferences, habits, experiences and needs. Thus, it was seen that the individual needs of the students directly affect the creation of directions and routes in the use of maps. In line with the findings obtained in the current study, it was concluded that the second most difficult skill to be acquired by students is "recognizing signs and symbols". Balcı (2015) argues that the individual should be able to read the scales of the maps he/she uses during his/her practices in the space. Individuals should be able to adapt the scale values on the map to the actual values in the space. The mistake made at this stage can create inconsistency in estimating distances. However, in his study, he stated that most of the pre-service geography teachers did not have difficulty in establishing a relationship between the scale of the map and the actual values during the field applications. The contradiction between the finding of the current study and the finding of Balcı (2015) is thought to be due to the fact that the number of geography courses taken by the pre-service social studies teachers is less than the number of students receiving education in the field of geography education. Ooms *et al*., (2012) stated that when the studies on eye tracking between students who took map skills courses and novices who had not previously received training on maps were examined, recognizing signs and symbols and mastering them (having taken the map skills course) had a positive effect on rapid decision-making and interpretation processes in individuals. It was seen that the skill found to be the easiest by the students in map reading skills or scored most lenient by the raters is "using landforms". In the studies (Çalışkan, 2015; Özgen, 2010), it is stated that the landforms in geomorphology, which is one of the basic disciplines of physical geography, are difficult to recognize and comprehend in the classroom environment. The reflection of topography on the map to make sense of the space is essentially related to spatial thinking skills (Yayla, 2019). Studies have shown that orienteering practices improve the ability to accurately recognize landforms (Tuna & Balcı, 2013). According to Wiegand (2006), since orienteering and scouting activities are generally voluntary activities, map education in these areas is limited. However, geography education given within the scope of a curriculum and map education in schools plays an active role in map teaching because they are systematic and programmed. At the same time, according to Gilhooly (1988), it is claimed that maps that provide contour information, provide permanence in the minds of individuals for longer periods in the process of learning the map and in developing mapping skills. The effective use of maps and the determination of landforms with the isohypse method are also possible through orienteering applications (Görmez, 2021). Similarly, Balcı (2015) stated that most of the participants did not have difficulty in reading the landforms. The finding obtained in the current study is consistent with the literature.

Another result of the study is that the skill of "managing time" was found to be easy by the students (scored lenient). In the current study, the students were asked to reach 5 targets in a period of 18 minutes in an area of approximately 5.7 hectares. While some participants used this time very effectively and quickly to reach the target and to complete the track, some participants could not complete it within the time limit and continued to search for the targets. In this connection, it was observed that there was a speed difference between the participants in terms of perceiving the space and interpreting it on the map. In some studies conducted on the basis of these differences, it has been seen that differences in quick thinking are an important parameter that creates individual differences in relation to neuro-physiological processes (Akcan, 2016; Sperdin *et al.,* 2009). The time elapsed between the time the stimulus triggers and the time the response appears is called the "reaction time". Reaction time reveals the ability to make quick decisions during the performance exhibited under the effect of space, time and other parameters in the environment (Akcan, 2016; Tamer, 2000). The existing research on the quick decision-making is focused on quick reading of maps by individuals (Lobben, 2007), eye movement measurement (Dong *et al*., 2018) and the importance of speed for map reading skills. In the current study, the skill of managing time was found to be easy by the students, which is thought to be because of the fact that the students used their eye movements effectively to use the orienteering map and tried to code the space in their minds in order to reach the targets quickly.

As a result of the MFRM analysis, although raters were selected from different fields such as geography, tourism and orienteering, the reliability of the inter-rater separation index was found to be close to zero (.05). This result indicates that there is no difference between the raters. At the same time, the fact that the Rasch-Cohen's Kappa statistic, calculated with the help of the values obtained from the model, had a value greater than 0.00, showed that the raters made consistent assessments and that the agreement between the raters was moderate (Eckes, 2009; Linacre, 2021). Thus, it was concluded that the reliability between the raters was established (Şata, 2019; Tobaş, 2020). Based on the results of the current study, it can be said that the use of analytical rubrics increases the level of objectivity by increasing the consistency between the scores and that it is a valid and reliable tool in assessing map reading skills.

Another result reached in this study was that the raters behaved almost equally in the categories included in the rubric. Range narrowing is observed when raters overuse any category of a rating scale (Wind, 2018). On the other hand, in the central tendency behaviour, aggregation occurs at the midpoint of the scoring scale (Myford & Wolfe, 2003). The central tendency behaviour and narrowing of the range threaten the validity of assessments as they prevent students from separating their performance correctly (Saal *et al.,* 1980). For this reason, it can be said that the analytical rubric developed in the study gives valid results in distinguishing successful and unsuccessful students (Tobaş, 2020).

As Stemler (2004) points out, getting average scores among raters may cause a systematic difference. Therefore, before calculating a summary score, it should be demonstrated that there is no rater bias. According to the results of the MFRM analysis, there was no finding indicating bias in rater-task and rater-student interaction. That is, raters behaved at the expected level in the criteria. According to Hung *et al*. (2012) such a result indicates that raters' interpretation of the rating scale is not different. In this case, it can be said that the opinions, beliefs or personality traits of the raters do not interfere with the scoring (Myford & Wolfe, 2003). However, student-by-task interaction results showed that 5 (39.25%) of them were biased. This result indicates that some tasks are easier or more difficult for these students (Engelhard & Myford, 2003).

5. DISCUSSION and CONCLUSION

Within the framework of the difficulties experienced by the students in their map reading skills, it is seen that some deficiencies in education in terms of geographical skills continue even at the higher education level. When the map reading skill is not imparted in a quality manner in educational environments, its reflections in society are clearly visible. For this reason, it can be said that practical activities aimed at imparting map reading skills to students should be integrated into curricula. It is stated that map reading skills are attempted to be imparted to students by teachers through lecturing and question-and-answer methods. However, in the 21st century, it is stated that there is a need for new methods in which the student will be active through the constructivist approach (Akengin *et al*. 2016). Large map activities (Anthamatten *et al*., 2018), field studies (Artvinli, 2021), and orienteering activities (Ayuldeş, 2021; Tanrıkulu, 2011; Tuna & Balcı, 2013; Yiğit, 2021) are examples that can be used in classroom assessment. In the current study, it was observed that the pre-service teachers had difficulty with some of the criteria in the rubric. Determining the factors affecting the map skills of the students or on which roads these skills can be raised to the highest level should be investigated. In cases where students have difficulty in understanding signs and symbols in reading maps, it can be contributed to recognizing symbols with digital games (Da Silva, 2015) and increasing imagination with simulation. On the other hand, students can be introduced to the findings in various applications such as GPS and satellite images about holding the map and finding directions, and it can be provided to create environments where students will encounter them more. The current work, of course, is about observing the current situation of students. Various experimental studies and studies that will reveal how these skills are affected can be included.

In the current study, an analytical rubric was used to assess map reading skills by means of orienteering activities. Rubrics are powerful tools as they not only improve student performance, but also clarify teacher expectations. Scores are expected to be more reliable when rubrics are used (Goodrich, 1997; Li & Lindsey, 2015). The results provided evidence of scoring rubric reliability with MFRM. In light of the results obtained, it can be said that the MFRM can be used to measure map reading skills. In addition, there is a need to develop appropriate tools (rubrics, checklists, rating scales) for the assessment of these performance-based practices. It can be suggested to researchers that similar tools should be developed for younger age groups (preschool, primary school).

The biggest limitation of this study is that the number of students included in the study was low because orienteering is time-consuming. Similar studies can be carried out by increasing the number of students or raters. It is expected that increasing the rater or students' number may contribute to the reliability and consistency of scores (Erguvan & Dünya, 2020).

Although the criteria in the rubric in this study overlapped with the international literature (Wiegand, 2006, p. 1), map reading skills with orienteering activity were examined only in a particular institution and in one area. Similar studies can be designed for more easily accessible environments such as the schoolyard or school environment. However, it should be noted that it may not be easy to find all landforms in these areas. However, it is considered appropriate to use the analytical rubric developed in the study in the field conditions where suitable geographical elements are available.

Acknowledgments

The authors thank Associate Professor İlker Yiğit, Lecturer Uğur Tozkoparan, and Burdur Orienteering Provincial Representative Mehmet Sadık Duman for their contributions as raters in the data collection process of the study.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Mehmet Akif Ersoy University, GO 02/2022/472.

Authorship Contribution Statement

Seyma Uyar: Investigation, Resources, Methodology, Supervision, Validation, Formal Analysis, and Writing-original draft. **Onur Yayla**: Investigation, Resources, Participant recruitment, Structuring and application of the orienteering process and Writing-original draft. **Hidayet Zunber**: Investigation, Resources, Participant recruitment, application of the orienteering process.

Orcid

Seyma Uyar https://orcid.org/0000-0002-8315-2637
Onur Yayla https://orcid.org/0000-0002-8710-3701
Hidayet Zunber https://orcid.org/0000-0002-8797-2835

REFERENCES

Abbak, A. (2021). Harita okuma ve yorumlama becerisinin incelenmesi. *Journal of Social Sciences And Education, 4*(1), 158-180. https://doi.org/10.53047/josse.829665

Adams, W.P. (1972). Geography and orienteering. *Journal of Geography, 71*(8), 473-480. https://dx.doi.org/10.1080/00221347208985332

Akcan, İ.O. (2016). *Orienteering elite athletes of the relationship between the visual reaction time with decision-making styles* [Unpublished Master's Thesis, Gazi University]. Publication No. 426905. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Akengin, H., Tuncel, G., & Cendek, M.E. (2016). Öğrencilerde harita okuryazarlığının geliştirilmesine ilişkin sosyal bilgiler öğretmenlerinin görüşleri [The social sciences teachers' opinions about developing map literacy of students]. *Marmara Geographical Review, 34*(1), 61-69. https://dergipark.org.tr/en/pub/marucog/issue/24661/260863

Akkuş, Z., & Kuzey, M. (2018). Ortaokul öğrencilerinin harita ve yön becerilerine sahip olma ve bu becerileri yaşama aktarabilme durumları üzerine bir değerlendirme [An examination on having map students and implementing these skills to their life]. *Journal of National Education, 47*(218), 201-234. https://dergipark.org.tr/en/download/article-file/556496

Aksoy, H., & Ünlü, M. (2012). Coğrafya derslerinde harita becerisine yönelik uygulamaların öğrenci tutumlarına etkisi [Effect of student attitudes applications ability maps in geography]. *Marmara Geographical Review, 26*(1), 16-41. https://dergipark.org.tr/tr/pub/manucog/issue/473/3875

Aktürk, V., Yazıcı, H., & Bulut, R. (2013). Sosyal bilgiler dersinde animasyon ve dijital harita kullanımının öğrencilerin mekân algılama becerilerine yönelik etkileri [The effects of using maps in social studies class on students' space perception skills]. *Marmara Geographical Review, 28,* 1-17. https://dergipark.org.tr/tr/pub/marucog/issue/475/3922

Alım, M., & Girgin, M. (2012). Coğrafya dersleri için kabartma harita yapımı [Relief map making for geography lessons]. *Eastern Geographical Review, 16*(25), 183-192. https://dergipark.org.tr/tr/pub/ataunidcd/issue/2453/31238

Andrich, D. (1978). A rating formulation for or dered response categories. *Psychometrika, 43*(1), 561-573.

Anthamatten, P., Bryant, L.M., Ferrucci, B.J., Jennings, S., & Theobald, R. (2018). Giant maps as pedagogical tools for teaching geography and mathematics. *Journal of Geography, 117*(5), 183-192. https://doi.org/10.1080/00221341.2017.1413413

Arıkan, A., & Aladağ, E. (2019). The effect of orienteering course on map literacy skills of students at school of physical education and sports. *International Journal of Geography and Geography Education (IGGE), 40*(1), 124-138.

Artvinli, E. (2021). *Development of map skills of students with orienteering social studies course in secondary school*. In E. Artvinli (Ed.) IGU proceedings of geographical education sessions (pp. 105-109). https://igc2021.org

Atayeter, Y., Yayla, O., Tozkoparan, U. & Sakar, T. (2018). Sosyal bilgiler öğretmen adaylarının mekânsal düşünme becerilerinin incelenmesi (Burdur ili örneği) [Examination of spatial thinking skils of social studies teacher candidates (example of Burdur province)]. *Multidiscipliner Studies 4 (Educational Sciences), 1*(1), 29-45.

Atit, K., Weisberg, S.M., Newcombe, N.S., & Shipley, T.F. (2016). Learning to interpret topographic maps: Understanding layered spatial information. *Cognitive Research: Principles and Implications, 1*(1), 1-18. https://doi.org/10.1186/s41235-016-0002-y

Ayuldeş, M. (2020). *The effect of the application of orienteering practices in primary education 6th grade social studies education on students' academic achievement and map literacy levels* [Unpublished Master's Thesis, Trabzon University]. Publication No. 652851. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Bahar, H.H., Sayar, K. & Başıbüyük, A. (2010). İlköğretim öğrencilerinin kroki okuma becerilerinin incelenmesi (Erzincan örneği) [The study of sketch map reading skills of the students in primary school (Erzincan sample)]. *Fırat University Journal of Social Sciences, 20*(1), 229-246. https://dergipark.org.tr/tr/pub/firatsbed/issue/45190/565886

Băițan, G.F. (2022). Orienteering–a necessary sports discipline for training the military. *Bulletin of "Carol I" National Defence University, 11*(1), 110-116. https://doi.org/10.53477/2284-9378-22-67

Baird, J.A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability: A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling*. Oxford University Centre for Educational Assessment.

Balcı, A. (2015). Coğrafya öğretmen adaylarının coğrafi arazi uygulamalarındaki harita okuryazarlıklarını tespitine yönelik bir araştırma [A study on the determination of map literacy of geography teacher candidates in geographical land applications]. *The Journal of Academic Social Science, 10*(3), 16-35.

Bartz, B.S. (1970). Maps in the classroom. *Journal of Geography, 69*(1), 18-24. https://doi.org/10.1080/00221347008981738

Bednarz, S.W., Acheson, G., & Bednarz, R.S. (2006). Maps and map learning in social studies. *Social Education, 70*, 398-404. https://www.socialstudies.org/system/files/publications/articles/se_700706398.pdf

Bednarz, W.S. (2001). Thinking spatially: ıncorporating geographic information science in pre and post secondary education. In L. Teoksessa Houtsonen and M. Tammilehto (Eds.), *Innovative practices in geographical education* (Pp. 3-7). Helsinki: Proceedings of The Helsinki Symposium of The IGU Commission on Geographical Education.

Borich, G.D., & Bauman, P.M. (1972). Convergent and discriminant validation of the french and spatial visualization factors. *Educational and Psychological Measurement, 32*(4), 1029-1033. https://doi.org/10.1177/001316447203200418

Buğdaycı, İ., & Bildirici, İ.Ö. (2009). Harita kullanımının coğrafya eğitimindeki önemi [The ımportance of map usage in geography education]. *TMMOB Harita ve Kadastro Mühendisleri Odası, 12*(1), 11-15.

Büyüköztürk, S., Kılıç-Çakmak, E., Akgün, O.E., Karadeniz, Ş., & Demirel, F. (2019). *Bilimsel araştırma yöntemleri* [Scientific research methods] (26th Ed.). Ankara: Pegem Akademi Publication.

Candan, G. (2019). Coğrafya eğitiminde oryantiring etkinliklerinin kullanımı [Use of orienteering activities in geography education]. *Geography For All, 1*(2), 19-26.

Carbonel-Carrera, C. & Hess Medler, S. (2017). Spatial orientation skill ımprovement with geospatial applications: report of a multi-year study. *ISPRS International Journal of Geo-Information, 6*(9), 278. https://doi.org/10.3390/ijgi6090278

Carbonell-Carrera, C., & Bermejo Asensio, L.A. (2016). Augmented reality as a digital teaching environment to develop spatial thinking. *Cartography and geographic information science, 44*(3), 259-270. https://doi.org//10.1080/15230406.2016.1145556

Carlin, B.P., & Louis, T.A. (2008). *Bayesian methods for data analysis*. CRC Press.

Carswell, R.J. (1971). The role of the user in the map communication process: Children's abilities in topographic map reading. *Cartographica: The International Journal for Geographic Information and Geovisualization, 8*(2), 40-45. https://doi.org/10.3138/UV32-7523-G562JKJ4

Chen, J.L., & Stanney, K.M. (1999). A theoretical model of wayfinding in virtual environments: proposed strategies for navigational aiding. *Presence, 8*(6), 671-685. https://doi.org/10.1162/105474699566558

Çalışkan, O. (2015). *Coğrafya eğitimi ve arazi çalışmaları* [Geography education and field studies]. Pegem Academia Publication.

Çepni, O. (2019). Location analysis. In B. Aksoy, B. Akbaba & B. Kılcan (Eds.). *Social studies skills education* (1st ed.) (pp. 367-386). Pegem Academia Publication.

Da Silva, C.N. (2015). Interactive digital games for geography teaching and understanding geographical space. *Creative Education, 6*(1), 692-700. http://dx.doi.org/10.4236/ce.2015.67070

Darakçı, S. (2014). Sosyal bilgiler öğretim programı ve ders kitaplarında harita kullanımı [Use of map in the social studies teaching program and textbooks]. *Mehmet Akif Ersoy University Journal of The Institute of Educational Sciences, 4*(3), 15-31. https://dergipark.org.tr/tr/pub/ebed/issue/22328/239289

Darling-Hammond, L., Adamson, F., & Abedi, J. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning* (p. 52). Stanford Center for Opportunity Pollcy in Education.

Demiralp, N. (2006). Coğrafya eğitiminde harita ve küre kullanım becerileri [Map and globe usage skills ingeography education]. *Turkish Journal of Educational Sciences, 4*(3), 323-343. https://dergipark.org.tr/en/pub/tebd/issue/26119/275163

Demirci, A., Karaburun, A., & Ünlü, M. (2013). Ortaöğretim kurumlarında cbs tabanlı projelerin uygulanması ve etkinliği [Implemantation and effectiveness of GIS-based projects in secondary schools]. *Journal of Geography, 112*(5), 214-228. https://doi.org/10.1080/00221341.2013.770545

Dong, W., Jiang, Y., Zheng, L., Liu, B., & Meng, L. (2018). Assessing map-reading skills using eye tracking and bayesian structural equation modelling. *Sustainability, 10*(9), 1-13. https://doi.org/10.3390/su10093050

Eckes, T. (2009). Many-facet rasch measurement. In S. Takala (ed.), *Reference supplement to the manual for relating language examinations to the common european framework of reference for languages: Learning, teaching, assessment* (section H). Council Of Europe/Language Policy Division

Engelhard, G.Jr. & Myford, C.M. (2003). Monitoring faculty consultant performance in the advancedplacement English literature and composition program with a many-faceted

Rasch model. *College Board Research Report, 2003*(1), ETS RR-03-01. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01893.x

Engelhard, G.Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement, 31*(2), 93- 112 https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Erguvan, I.D., & Aksu Dunya, B. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia 10*(1), 1-20. https://doi.org/10.1186/s40468-020-0098-3

Ertuğrul, Z. (2008). *The determination of 6th grade primary school students? map and globe usage skills* [Unpublished Master's Thesis, Gazi University]. Publication No. 219055. http://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Gersmehl, P.J., & Gersmehl, C.A. (2007). Spatial thinking by young children: neurologic evidence for early development and "educability". *Journal Of Geography, 106*(5), 181-191. https://doi.org/10.1080/00221340701809108

Gilhooly, K.J., Wood, M., Kinnear, P.R., & Green, C. (1988). Skill in map reading and memory for maps. *The Quarterly Journal of Experimental Psychology Section A, 40*(1), 87-107. https://doi.org/10.1080/14640748808402284

Golledge, R.G., Jacobson, R.D., Kitchin, R., & Blades, M. (2000). Cognitive maps, spatial abilities, and human wayfinding. *Geographical review of Japan, Series B., 73*(2), 93-104. https://doi.org/10.4157/grj1984b.73.93

Goodrich, H.G. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-17.

Görmez, E. (2021). Ortaokul öğrencilerinin harita okuryazarlık becerisi yeterlilikleri üzerine bir çalışma [A study on map literacy skill competences of secondary school students]. *Journal of Van Yüzüncü Yıl University Faculty of Education, 18*(2), 712-733. https://doi.org/10.33711/yyuefd.1029178

Güler, N. (2014). Analysis of open-ended statistics questions with many facet rasch model. Eurasian Journal of Educational Research, 55(1), 73-90. https://dx.doi.org/10.14689/ejer.2014.55.5

Güneş, G. & Öztürk-Demirbaş, Ç. (2020). Sosyal bilgiler öğretmen adaylarının harita kullanabilme beceri düzeylerinin farklı değişkenler açısından incelenmesi [Investigation of social studies teacher candidates' level of map using skills in terms of different variables]. *MANAS Journal of Social Studies, 9* (4), 2145-2158. https://dergipark.org.tr/tr/pub/mjss/issue/57186/675955

Haiyang, S. (2010). An application of classical test theory and many-facet rasch measurement in analyzing the reliability of an english test for non-english major graduates. *Chinese Journal of Applied Linguistics (Foreign Language Teaching & Research Press), 33*(2), 87-102.

Hung, S., Chen, P. & Chen, H. (2012). Improving creativity performance assessment: A rater effect examination with Many Facet Rasch Model. *Creativity Research Journal, 24*(4), 345-357. https://doi.org/10.1080/10400419.2012.730331

Huynh, N.T., & Sharpe, B. (2013). An assessment instrument to measure geospatial thinking expertise. *Journal of Geography, 112*(1), 3-17. https://doi.org/10.1080/00221341.2012.682227

International Cartographic Association (ICA) (2022,11). International Cartographic Association (ICA). https://icaci.org/

İncekara, S., Karatepe, A., & Karaburun, A. (2008). Ortaöğretim coğrafya derslerinde cbs yoluyla harita okuma becerisinin kazandırılmasına yönelik bir uygulama [An application about how to teach topographic map reading with gis in secondary school geography courses]. *Marmara Geographical Review, 17*(1)*,* 97-110. https://dergipark.org.tr/tr/pub/marucog/issue/464/3735

Jo, I. & Bednarz, S.W. (2014-a). Dispositions toward teaching spatial thinking through geography: conceptualization and an exemplar assessment. *Journal of Geography, 113*(5), 198-207.

Jo, I. & Bednarz, S.W. (2014-b). Developing pre-service teachers' pedagogical content knowledge for teaching spatial thinking through geography. *Journal of Geography in Higher Education, 38*(2), 301-313.

Jo, I. (2007). *Aspect of spatial thinking in geography textbook* [Unpublished Doctoral Dissertation]. Seoul National University.

Jo, I. (2011). *Fostering a spatially literate generation: explicit instruction in spatial thinking for preservice teachers* (Unpublished doctoral dissertation). Office of Graduate Studies of Texas A&M University, Texas, USA.

Karasar, N. (2005). *Bilimsel araştırma yöntemi* [*Scientific research method*]. Nobel Publication Distribution.

Kaya Uyanık, G., Güler, N., Taşdelen Teker, G., & Demir, S. (2019). Fen bilimleri dersi etkinliklerinin çok yüzeyli rasch modeliyle analizi [The analysis of elementary science education course activities through many-facet rasch model]. *Kastamonu Education Journal, 27*(1), 139-150. https://doi.org/10.24106/kefdergi.2417

Kaymakçı, S. (2015). Tarih öğretiminde harita becerilerinin gerekliliği üzerine bir çalışma [A study on the necessity of map skills in history teaching]. *Ataturk University Journal of Social Sciences Institute,19*(3), 127-154. https://dergipark.org.tr/tr/pub/ataunisosbil/issue/45086/563148

Kıroğlu K. (2006). *Yeni ilköğretim programları (1-5. sınıflar)* [New primary education programs (1-5th grade)]. Pegem Academia Publication.

Kızılçaoğlu, A., & Ünlü, M. (2008). 9. sınıfta dilsiz haritaların kullanımına yönelik aktivite önerileri [Suggestions of outline map activities for the 9the grade geography course]. *Marmara Geographical Review, 17*(1), 45-67. https://dergipark.org.tr/tr/pub/marucog/issue/464/3732

Kızılçaoğlu, A. (2007). Harita becerilerine pedagojik bir bakış [A pedagogical look at map skills]. *Selcuk University Journal of Social Studies, 18*(1), 341-358. https://dergipark.org.tr/tr/pub/susbed/issue/61794/924246

Koç, H. & Karatekin, K. (2016). Sosyal bilgiler öğretmen adaylarının harita okuryazarlık düzeylerinin çeşitli değişkenler açısından incelenmesi [Investigation of social studies teacher candidates' map literacy levels in terms of various variables]. *Bolu Abant İzzet Baysal University Journal of Education, 16* (USBES Special Issue II), 441-461. https://app.trdizin.gov.tr/publication/paper/detail/TWpNNE5qSTRPQT09

Koç, H., & Bulut, İ. (2014). Gestalt kuramının öğrencilerin harita okuma ve yorumlama beceri düzeyleri üzerine etkisini belirlemeye yönelik bir inceleme [An investigation into the ımpacts of gestalt theory on learners map literacy skills]. *Marmara Geographical Review*, (30), 1-19. https://doi.org/10.14781/mcd.26337

Koç, H., Aksoy, B. & Çifçi, T. (2017). Farklı lisans programlardaki öğrencilerin harita okuryazarlık düzeylerinin çeşitli değişkenler açısından incelenmesi: Cumhuriyet üniversitesi örneği [An examination of map literacy levels of students from various undergraduate programmes according to several variables: Cumhuriyet university sample]. *Erzincan University Journal of Education Faculty, 19*(3), 301-321. https://doi.org/10.17556/erziefd.331083

Koláčný, A. (1969). Cartographic information—a fundamental concept and term in modern cartography. *The Cartographic Journal, 6*(1), 47-49. https://doi.org/10.1179/caj.1969.6.1.47

Lee, J.W. (2005). *Effect of gis learning on spatial ability* (Unpublished Doctoral Dissertation). A Dissertation Graduate Studies of Texas A&M University, Texas, USA.

Li, J. & Lindsey, P. (2015). Understanding variations between student and teacher application of rubrics. *Assessing Writing, 26*(1), 67–79. https://doi.org/10.1016/j.asw.2015.07.003

Linacre, J.M. (1994). Constructing measurement with a many-facet rasch model. In Wilson, M. (Ed.), *Objective measurement: Theory into practice*, (2nd Ed.) (pp. 129-144). Ablex.

Linacre, J.M. (2021). *A user's guide to facets rasch-model computer programs*. Available online www.winsteps.com

Lobben, A.K. (2007). Navigational map reading: Predicting performance and identifying relative influence of map-related abilities. *Annals of The Association of American Geographers, 97*(1), 64-85. https://doi.org/10.1111/j.1467-8306.200700524.x

Maier, P.H. (1996). *Spatial geometry and spatial ability–how to make solid geometry solid*. In selected papers from the annual conference of didactics of mathematics (pp. 63-75).

Meyer, J.M. (1973). Map skills instruction and the child's developing cognitive abilities. *Journal of Geography, 72*(6), 27-35.

Murakoshi, S. (1997). Navigational planning in orienteering. *The Journal of Navigation, 50*(2), 321-327. https://doi.org/10.1017/S0373463300023948

Myford, C.M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many facet rasch measurement: Part 1. *Journal of Applied Measurement, 4*(4), 386-422. https://researchgate.net

Newcombe, N.S., Uttal, D.H., & Sauter, M. (2013). Spatial development. In P. D. Zelazo (Eds.), *The Oxford handbook of developmental psychology (Vol. 1): Body and mind* (pp. 564–590). Oxford University Press.

Ooms, K., De Maeyer, P., Dupont, L., Van der Veken, N., Van de Weghe, N., & Verplaetse, S. (2016). Education in cartography: What is the status of young people's map-reading skills?. *Cartography and Geographic Information Science, 43*(2), 134-153. https://doi.org/10.1080/15230406.2015.1021713

Ooms, K., De Maeyer, P., Fack, V., Van Assche, E., & Witlox, F. (2012). Interpreting maps through the eyes of expert and novice users. *International Journal of Geographical Information Science, 26*(10), 1773-1788. https://doi.org/10.1080/13658816.2011.642801

Öcal, A. (2007). *Investigation of special cognition skills of 6th grade students in primary school social studies course* [Unpublished Doctoral Dissertation, Gazi University]. Publication No. 207102. http://tez.tok.gov.tr/UlusalTezMerkezi/giris.jsp

Özcan, F. & Uzun, F.V. (2016). Sosyal bilgiler öğretmen adaylarının çeşitli değişkenlere göre harita okuma özyeterlilik ve başarı düzeyleri [Review of map reading self-efficacy and success level of social studies teacher candidates according to several variables]. *Journal of Dicle University Ziya Gökalp Faculty of Education, 29*(1), 387-400. http://dx.doi.org/10.14582/DUZGEF.764

Özgen, N. (2010). Bilim olarak coğrafya ve evrimsel paradigmaları [Geography as a science and its evaluative paradigms]. *Ege Coğrafya Dergisi, 19*(2), 1-26. https://dergipark.org.tr/en/pub/ecd/issue/4872/66894

Pala, Ş.M. & Başıbüyük, A. (2019). Matematik becerisinin sosyal bilgiler derslerindeki harita grafik ve tablo okuma becerilerine etkisi. *Uluslararası Sosyal Bilgilerde Yeni Yaklaşımlar Dergisi, 3*(1), 41-56.

Randall, J. & Engelhard, G. Jr. (2009). Examining teacher grades using rasch measurement theory. *Journal of Educational Measurement, 46*(1), 1-18. https://doi.org/10.1111/j.1745-3984.2009.01066.x

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428. https://doi.org/10.1037/0033-2909.88.2.413

Safi, H. (2010). *Teachers' views on the development of space perception skills in the social studies curriculum* [Unpublished Master's Thesis, Marmara University]. Publication No. 979907. http://tez.tok.gov.tr/UlusalTezMerkezi/giris.jsp

Sarıgül, O. (2021) Yaşantımızın vazgeçilmez araçları: Haritalar [Indispensable Tools of Our Lives: Maps]. Koç H., Ergün A. (Eds.), *Bilginin görsel ifadesi: Haritalar* [Visual Expression of Knowledge: Maps] (1st ed.) (pp. 2-31). Pegem Academy Publication.

Schnotz, W., & Kulhavy, R.W. (Eds.). (1994). *Comprehension of graphics.* Elsevier.

Sönmez, Ö.F. (2010). *Map skills in primary social studies education* [Unpublished Doctoral Dissertation, Gazi University]. Publication No. 298446. http://tez.tok.gov.tr/UlusalTezMerkezi/giris.jsp

Sönmez, Ö.F. (2019). *Map literacy*. In B. Aksoy, B. Akbaba & B. Kılcan (Eds.), *Social studies skills education* (1st ed.) (pp. 219-232). Pegem Academia Publication.

Sönmez, Ö.F., & Aksoy, B. (2012). İlköğretim ikinci kademe öğrencilerinin harita beceri düzeylerinin belirlenmesi [Determination of map skill levels of primary school second level student]. *Electronic Turkish Studies, 7*(1), 1905-1924. https://www.acarindex.com/dosyalar/makale/acarindex-1423933856.pdf

Sönmez, Ö.F. & Aksoy, B. (2013). Cumhuriyetten günümüze ilköğretim programlarında harita becerileri [Map skills in primary education curriculum from the republic to the present]. *Turkish Journal of Social Research, 171*(171), 269-288. https://dergipark.org.tr/en/pub/tsadergisi/issue/21497/230500

Sperdin, H.F., Cappe, C., Foxe, J.J., & Murray, M.M. (2009). Early, low-level auditory-somato sensory multisensory interactions impact reaction time speed. *Frontiers in Integrative Neureo Science, 3*(1), 1-10. https://doi.org/10.3389/neuro.07.002.2009

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation, 9*(1), 4. https://doi.org/10.7275/96jp-xz07

Stumpf, H., & Eliot, J. (1999). A structural analysis of visual spatial ability in academically talented students. *Learning and Individual Differences, 11*(2), 137-151. https://10.1016/S1041-6080(00)80002-3

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261. https://doi.org/10.1016/j.asw.2004.11.001

Şanlı C. (2021). Coğrafya öğretmen adaylarının mekânsal kavramlara ilişkin bilişsel yapılarının incelenmesi [Examining the cognitive structures of geography teacher candidates regarding spatial concepts]. *Journal of History School, 14*(51), 1060-1084. https://doi.org/10.29228/Joh.49537

Şanlı C. (2021). Mekânsal düşünme becerisi testinin geliştirilmesi [Developing a spatial thinking ability test]. *International Journal of Eurasia Social Sciences, 12*(43), 1-18. https://dx.doi.org/10.35826/ijoess.2858

Şanlı C., & Sezer, A. (2019). Coğrafya öğretiminde mekânsal düşünme ölçeği: Türkçe'ye uyarlama geçerlik ve güvenirlik çalışması [Teaching spatial thinking through geography disposition inventory: a validity and reliability study on its Turkish adaptation]. *Aegean Geographical Journal, 28*(2), 213-225.

Şanlı, C. (2019). Coğrafya öğretmen adaylarının mekânsal düşünme becerisine ilişkin görüşleri [Geography Teacher Candidates' Views on Spatial Thinking Skills]. J*ournal of Anatolian Cultural Research, 3*(3), 215-233. http://www.ankad.org/index.php/Ankad/article/view/59

Şanlı, C. (2020). Mekansal düşünme becerisinin sosyal bilgiler ders kitapları sorularında analizi [The analysis of spatial thinking skills in the questions included within social sciences coursebooks]. *International Journal of Geography and Geography Education (IGGE), (42),* 118-132. https://doi.org/10.32003/igge.724028

Şata, M. (2019). *The investigation of the effect of rater training on the rater behaviors ın the performance assessment process* [Unpublished Doctoral Dissertation, Gazi University]. Publication No. 626117. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Tamer, K. (2000). *Sporda fiziksel-fizyolojik performansın ölçülmesi ve değerlendirilmesi* [*Measurement and evaluation of physical-physiological performance in sports*]. Bağırgan Publication.

Tanrıkulu, M. (2011). Harita ve pusulanın farklı bir kullanım alanı: oryantiring [A different use of maps and compass: orienteering]. *Journal of National Education, 41*(191), 120-126. https://dergipark.org.tr/en/pub/milliegitim/issue/36191/406895

Taş, H.İ. (2006). Coğrafya eğitiminde görselleştirmenin önemi: Mekânsal algılamaya pedagojik bir bakış [The ımportance of vitualizations in geographic education: an educational aproaches to spatial context]. *Eastern Geographical Review, 11*(16), 211-237.

Taşlı, İ., Çelik, H., & Taşlı, M. (2007). Yapılandırmacı öğretim ve sosyal bilgilerde harita kullanım durumlarının bazı değişkenler açısından incelenmesi (demirci-gördes örneği) [Investigation of map usage cases in constructivist teaching and social studies in terms of some variables (Demirci-Gördes sample)]. *Celal Bayar University Journal of Social Sciences, 5*(2), 57-68.

Tobaş, C. (2020). *Examination of the differential rater behaviours in performance evaluation with many facet rasch measurement* [Unpublished Master's Thesis, Gazi University]. Publication No. 629923. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Tuna F., Demirci, A., & Gültekin, N. (2012). Temel coğrafi bilgi ve beceriler toplumda ne ölçüde kullanılıyor? yön, konum ve harita becerilerinde mevcut durum analizi [To what extent are basic geographic ınformation and skills used in society? current situation analysis in direction, location and map skills]. *Journal of National Education, 42*(195), 211-227. https://dergipark.org.tr/en/pub/milliegitim/issue/36172/406706

Tuna, F., & Balcı, A. (2013). Oryantiring uygulamalarının coğrafya öğretmen adaylarının özyeterlik algılarına etkisi [The effects of orienteering on prospective geography teachers' selfefficacy perceptions]. *Marmara Geographical Review, 27,* 1-14. https://dergipark.org.tr/tr/pub/marucog/issue/474/3890

Turner, J. (2003). *Examining an art portfolio assessment using a many-facet rasch measurement model* [Unpublished Doctoral Dissertation, Boston College]. https://www.proquest.com/pagepdf/305343181?accountid=37161

Tümertekin, E., & Özgüç, N. (2019). *Beşeri coğrafya: İnsan, kültür, mekân* [*Human geography: People, culture, spatial*]. Çantay Kitabevi.

Ünlü, M. (2021). *Harita becerileri.* In H. Koç & A. Ergün (Eds.). *Bilginin görsel ifadesi: Haritalar* [*Map Skills*. In H. Koç & A. Ergün (Eds.). *Visual Expression of Knowledge: Maps*] (1st ed.) (pp. 388-418). Pegem Academia Publication.

Ünlü, M., & Güncegörü Aksoy, H. (2013). Coğrafya derslerinde harita becerilerine yönelik uygulamaların öğretmen tutumlarına etkisi [The influence of the applications for maps teachers' attitudes in geography lessons]. *Marmara Geographical Review, 27,* 58-71. https://dergipark.org.tr/tr/pub/marucog/issue/474/3893

Ünlü, M., & Yıldırım S. (2017). Coğrafya dersi öğretim programına bir coğrafi beceri önerisi: mekânsal düşünme becerisi [A geographical skill suggestion to geograp teaching curriculum: spatial thinking skill]. *Marmara Geographical Review, 35*(1), 13-20. https://doi.org/10.1478/mcd.291018

Wiegand, P. (2006). *Learning and teaching with maps.* Routledge.

Wilson, J.A. (2017). *Orienteering, the map and child development* (Organised outdoor play with maps). Department of Natural and Built Environment.

Wind, S.A. (2018). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement, 43*(2), 159-171. https://doi.org/10.1177/0146621618789391

Yayla, O. (2019). *The effect of spatial thinking skills and academic achievements of teaching applications based on spatial technologies in social studies education* [Unpublished Master's Thesis, Trabzon University]. Publication No. 568309. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Yiğit, T. (2020). *Effect of spatial thinking skills of students orienteering application in social studies* [Unpublished Master's Thesis, Kastamonu University]. Publication No. 647306. https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Yurdugül, H. (2005). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması* [*Using content validity ındices for content validity in scale development studies*]. XIV. National Educational Sciences Congres, 1, 771-774. https://yunus.hacettepe.edu.tr/~yurdugül/3/indir/PamukkaleBildiri.pdf

APPENDIX

Appendix 1: A Rubric for Orientiring Map Reading Skills

A Rubric for Orientiring Map Reading Skills					
Criteria	**Categories**				
	1	**2**	**3**	**4**	**Score**
Holding map and finding direction and location	He/She hold the map upside down, unable to find direction and location.	He/She hold the map right but could not find direction or location.	He/She hold the map right, finded the direction, found the approximate location using triangulation.	He/She holds the map right, finded the direction, pinpointed multiple triangulation points and found the exact location.	
Recognizing signs and symbols	………………	He/She recognized the signs and symbols given on the map, but could not calculate the distance.	………………	He/She recognized the signs and symbols given on the map and calculated the distance with a compass.	
Using landforms	………………	………………	………………	………………	
Managing time	………………	………………	………………	………………	

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 283–299

https://doi.org/10.21449/ijate.1121814

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Learning through teaching: Teaching the nature of scientific inquiry in online outdoor learning environments

Eda Erdas Kartal[1,*], **Gunkut Mesci**[2]

[1]Kastamonu University, Faculty of Education, Department of Educational Sciences, Kastamonu
[2]Giresun University, Faculty of Education, Department of Science Education, Giresun

ARTICLE HISTORY

Received: May. 26, 2022
Revised: Nov. 1, 2022
Accepted: Nov. 10, 2021

Keywords:

Nature of scientific inquiry,
Outdoor learning,
Teacher training,
Learning through teaching.

Abstract: This study aims to examine the developments of 50 pre-service teachers' NOSI views during a 14-week implementation in the online outdoor learning environment. This is an experimental study that examines each participant's views and changes about NOSI using an open-ended questionnaire (VASI), and follow-up interviews. The data were analyzed by using content analysis. Almost all participants positively improved their views through the explicit/reflective approach and teachers' own experiences by practicing. In this study, the views of pre-service teachers developed more clearly after preparing lesson plans and their teaching practices. This is an indication that NOSI teaching, which does not provide the experience of conveying their learning outcomes to their practices to the participants is limited on its own and that the importance of "learning through teaching" in teachers' in-service and pre-service training on this subject should not be overlooked. Online teacher education in outdoor learning environments might be used in the development of NOSI views of pre-service teachers. We think that it is important to investigate the effect of this training on teacher education. These types of training might create a more economical and sustainable alternative for the development of NOSI views of wider groups of pre-service and in-service teachers.

1. INTRODUCTION

Science and technology are constantly changing and societies are expected to keep up with this rapid change and development. In this regard, raising science-literate individuals who can keep up with the changes has become the primary target of science curricula (American Association for the Advancement of Science (AAAS), 1993; Ministry of National Education (MoNE), 2018; National Research Council (NRC), 2012; Next Generation Science Standards (NGSS), 2013). Different definitions of scientific literacy have been examined and there are three different interpretations of the word "literate". These are literate as learned, literate as competent, and literate as able to function minimally in society (Laugksch, 2000). While interpretations of the concept of literacy move from "informed" to "function in society" from past to present, today, increasing emphasis is placed on scientific literacy qualities to cope with situations encountered in daily life (Laugksch, 2000). Scientific literacy means having scientific knowledge, the nature of scientific knowledge, and how it is produced and using this knowledge to solve problems

*Corresponding Author: Eda Erdas Kartal ✉ erdaseda@gmail.com Kastamonu University, Faculty of Education, Department of Educational Sciences, Kastamonu, Türkiye

encountered in daily life. Scientific literacy also requires being aware of how science, technology, and society affect each other and having positive attitudes and value judgments about science and technology (NRC, 2012; Organization for Economic Co-operation and Development (OECD), 2003). Individuals with scientific literacy skills can distinguish science from non-science, use scientific knowledge in problem solving, and think scientifically. They are aware of the role of experiments in science. They know the theories that form the basis of science, how they are achieved and why they are widely accepted. They know the elements of scientific research, the importance of proper inquiry, relying on objective evidence, and deductive reasoning and logical thought processes (Norris & Philips, 2003).

Although scientific literacy includes understanding the content of science, it is much more than that. Students must have an understanding of science subjects as well as the nature of science (NOS) and the nature of scientific inquiry (NOSI) to be scientifically literate (Bartels & Lederman, 2022). As the main component of science literacy, scientific inquiry involves traditional science processes, which refers to combining these processes with scientific knowledge, scientific reasoning, and critical thinking to develop scientific knowledge (Lederman et al., 2014). Scientific inquiry is the whole of systematic research activities carried out by scientists to understand and explain the natural world (Lederman & Lederman, 2012; NRC, 2000). It is important to have scientific inquiry skills, but the fact that students have scientific inquiry skills does not mean that they have knowledge of the NOSI. Teachers usually focus on doing inquiry in schools and assume that students will know how scientific inquiry is done by doing scientific inquiry (Bell et al., 2003). However, students can make scientific inquiries without knowing how and why scientists continue their work (Lederman et al., 2019).

Scientific inquiry should be emphasized as a skill and understanding (NGSS, 2013). Participating in simple inquiry experiences and knowing inquiry procedures without knowing the NOSI is not enough for students to understand the epistemology of science and achieve the objectives that are targeted by scientific inquiry (Lederman, 2006; Wong and Hodson, 2010). The NOSI expresses the characteristics of the scientific inquiry process (Lederman et al., 2014). It is necessary to explain the source of the information we have and why we believe it to teach not only the process of creating scientific knowledge but also the characteristics of this process, that is, to gain an adequate understanding of the features (components of scientific inquiry) (Osborne, 2014; Schwartz, 2004). The aspects of NOSI are defined as follows: (1) All scientific research begins with a question, but it does not always have to test a hypothesis, (2) There is no single, step-by-step scientific method used in all scientific research, (3) Research questions guide the scientific inquiry process, (4) Not all scientists who do the same can achieve the same results, (5) Scientific inquiry procedures can have an impact on the results, (6) There should be consistency between research findings and data collected, (7) Scientific data and scientific evidence are not the same, (8) Combining previously known and collected data develops scientific explanations (Lederman et al., 2014). Researchers and reform documents emphasize the importance of developing students' scientific inquiry skills, as well as their views of the abovementioned features of the scientific inquiry process (Lederman et al., 2019; NGSS, 2013; NRC, 2020).

1.1. Problem Statement

Although it is emphasized in international documents that the foundation of scientific literacy should be established from kindergarten, more importance is given to reading and mathematics in early grades (Aydemir et al., 2017; Bartels & Lederman, 2022). Allocating more time to reading and mathematics in early classes causes science education to remain in the background in these classes. At an early age, children are interrogative and inquisitive by nature. During this period, children's imaginations are also quite strong. The first experiences that children have in this period are extremely important and these experiences form the basis for their future lives (Alisinanoğlu & Özbey, 2011; Çamlıbel Çakmak, 2014). Studies show that children develop an understanding of basic scientific concepts and can use basic scientific process skills at early ages (Opfer & Siegler, 2004). To raise the science-literate individuals of the future, children need to spend this period, in which they learn quickly and the lasting impact of the concepts they learn, productively in terms of science education. Unfortunately, students continue to graduate from high school without science literacy skills due to the lack of time for science teaching in early grades (Roberts & Bybee, 2014). Bartels and Lederman (2022) showed in their research that students' understanding of science, scientists, and how scientists work did not change from the first grade to the fifth grade. The findings of Bartels and Lederman (2022) are a tragic indicator that students fail to make progress in terms of scientific literacy at early grades.

Science teaching, which is recommended from kindergarten onward, should focus not only on science content but also on applications and understanding what science is as a body of knowledge (NRC, 2013). The teaching of NOSI usually begins in middle school, but recent studies have revealed that early graders (kindergarten to K5) also have the capacity to understand some features of scientific inquiry, so it should be started at the earliest age possible (Bartels & Lederman, 2022; Lederman et al., 2019; Tytler & Peterson, 2003). The findings of the limited number of studies conducted with younger children show that these children's views on NOSI are limited (Bartels & Lederman, 2022; Lederman, 2012; Lederman & Bartels, 2018; Lederman et al., 2013; Lederman & Lederman, 2004; Penn et al., 2021).

The attitudes of children toward science and the process of learning science are highly affected by the knowledge, attitudes, and behaviors of the teacher (Yurt, 2015). Thus, teachers are important actors in the process of adopting scientific inquiry in science teaching and developing students' views (Schwartz & Lederman, 2002). Teachers' lack of understanding of scientific inquiry is one of the obstacles to applying it to their lessons (Roehring & Luft, 2004). It is important for teachers to understand NOSI, which guides scientific research and forms the basis of scientific knowledge (Zion & Mendelovici, 2012). Most studies (Baykara & Yakar, 2020; Crawford et al., 2005; Crawford et al., 2010; Dudu, 2014; Karışan et al., 2017; Lederman et al., 2019; Mesci et al., 2020; Mesci & Kartal, 2021; Wang & Zhao, 2016;) have aimed at identifying and developing the views of secondary and high school teachers. The findings of the limited number of studies conducted with early graders' teachers show that these teachers/pre-service teachers have naive views and misconceptions about NOSI (Aydemir et al., 2017; Deniz & Akerson, 2013; Koyunlu-Ünlü, 2020; Perez & Diaz-Moreno, 2022). Considering the limited number of studies aimed at improving NOSI views of pre-service and in-service teachers, there is still a need for dissemination of these studies.

1.2. Theoretical Framework

Science is closely intertwined with real life. Classroom and laboratory environments create some limitations for science teaching about relating science subjects to real life. This may cause difficulties in understanding science subjects. Outdoor learning is of great importance in terms of connecting the theoretical knowledge learned at school with real life and learning the events comparatively. This study was framed by "teaching in an outdoor learning environment", which is a type of teaching carried out by examining an event or phenomenon in its real natural environment, according to a previous plan made for achievements in science teaching (Rickinson et al., 2004). Recent studies have found that outdoor learning environments increase children's motivation to learn (Andiema, 2016) and increase their interest and achievement in science courses (Dori & Tall, 2000), but teachers mostly do not prefer to perform these activities (Tatar & Bağrıyanık, 2012). It is very important for teachers to include outdoor learning environments that affect students' interests, attitudes, and learning levels in the learning-teaching process in their professions (Kubat, 2018). Thus, it is necessary to provide pre-service teachers with experience on how science issues can be handled in outdoor learning environments. Recent studies have suggested that aspects of NOSI should be deemed as science content (i.e., Lederman, 2019; Mesci & Schwartz, 2017; Schwartz et al., 2008). In this context, teaching NOSI experienced by pre-service teachers in outdoor learning environments may be useful in improving views about the components of NOSI.

It is argued that one of the most effective teaching approaches in teaching the nature of scientific inquiry is the explicit/reflective approach (Aydeniz et al., 2011; Bell et al., 2003; Erdas-Kartal et al., 2018; Lederman, 2019; Mesci et al., 2020; Metin-Peten, 2022). For example, in one of these studies, Perez and Diaz-Moreno (2022), in their study where they examined the evolution of NOSI concepts of pre-service primary teachers after they were immersed in a specific teaching module focusing on NOSI, revealed that explicit/reflective approach-based NOSI teaching improved participants' views. Teachers may plan and teach NOSI courses effectively by improving their knowledge and awareness about NOSI (Mesci et al., 2020). Therefore, it should not be forgotten that the training to be offered to the teachers about NOSI should include explicit/reflective instruction on NOSI as well as providing the opportunity to practice. Bringing teachers and pre-service teachers together for such professional development support can be difficult and costly in many cases. To expand such professional development support, the possibilities offered by technology should be evaluated. Being unable to keep up with rapid technological developments is one of the important problems in catching up with the current age, so the use of new technologies in education is encouraged in many countries in the world. It is necessary to benefit from the opportunities offered by rapidly changing technologies in teacher education and in developing teachers' professional standards (Gelişli, 2015). The distance teaching approach, which makes it possible to provide educational services to the masses by using the developed and enriched resources of communication and education technology, is an important option that is suitable for effective and continuous use in pre-service and in-service teachers. One of the distance teacher education training models is web-based education, in other words, online training (Burns, 2011). This model is used in the vast majority of countries where access to Internet is high and technological skills are becoming widespread in school or home settings (Gelişli, 2015). We think it is important to investigate the effect of this training on teacher education. These types of training might create a more economical and sustainable alternative for the development of NOSI views of wider groups of pre-service and in-service teachers.

Based on the above-mentioned literature, this study aimed to develop the NOSI views of pre-service teachers with online training to be given in outdoor learning environments. Online NOSI training in outdoor learning environments is theoretically framed by the learning theory of Reid et al.'s (1989) 5-stage model under the constructivist approach. The first stage of this theory is engagement, which is described as ‘the time during which students acquire information and engage in an experience that provides the basis for, or content of, their ensuing learning’ (Reid et al.,1989). The second stage is exploration, which can be an open-ended process where learners follow their instincts. Transformation is the stage where the knowledge that the learner participates in and discovers can be restructured into a form that allows presentation (the next stage) but, more importantly, into a format from the instructor's point of view. This is usually a lesson plan preparation phase in teacher development programs, which ensures learning objectives. Presenting the transformed knowledge gives the learner time to reflect on the process and content, internalize it, and develop a deeper level of understanding. This section may coincide with the microteaching section in teacher development programs. The transformation and the resulting presentation are not the end of the process. The final stage is a reflection that can take many forms, usually in the form of oral presentations, reflection essays, posters, or creating a newspaper/magazine (Pritchard, 2017). Kolb's (1984) experiential learning model has also emphasized that the most important element in the learning process is the learner's own experiences. According to the experiential learning model, the individual should first engage with a certain concrete experience activity in the teaching process (Brock & Cameron 1999). Then, the individual should observe objectively and carefully in the reflective observation stage and analyze concrete experiences to reach certain judgments (Brock & Cameron, 1999).

1.3. Purpose of the Study

This study aimed to develop the NOSI views of pre-service teachers through online training to be given in outdoor learning environments. In this context, the following questions guided this study:

1.How is the change in the NOSI views of pre-service teachers after the outdoor learning course in online settings?
2.What are the pre-service teachers' views on the impact of the outdoor learning course on their NOSI views?

2. METHOD

A single-group experimental design was used in this study to explore the impact of online NOSI instruction to be given in outdoor learning environments on pre-service teachers' NOSI views (Creswell, 2012).

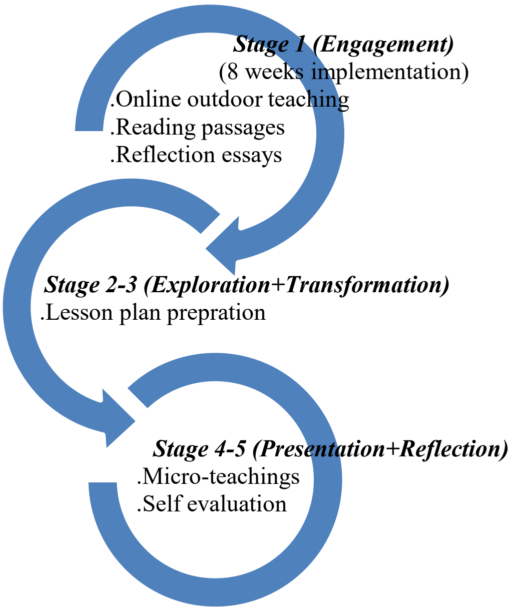
2.1. Participants

The sample of this study consisted of 50 pre-service elementary (25) and preschool (25) teachers who were teaching at a public university in northeastern Turkey. The participants were selected among those who took the undergraduate course, namely, "outdoor learning environments", a common elective undergraduate course for pre-service teachers, and volunteered to participate in the research. None of the participants had taken any course related to NOSI or the nature of science until then.

2.2. Context of the Study and Data Collection

At the beginning of the semester, pre-service teachers were asked to fill out the Views About Nature of Scientific Inquiry (VASI) Questionnaire (Lederman et al., 2014), and follow-up semi-structured interviews were implemented. The context of the study and data collection procedures are summarized in Figure 1 below.

Figure 1. *Context of the study and data collection procedure.*



2.2.1. *Stage 1. (Engagement)*

During the eight weeks of implementation, the researchers made a live video from the places mentioned in Table 1 below and made an interactive presentation for each week (for those who could not attend the live broadcast, it was recorded and uploaded to the system for further watch). Every week, special emphasis was placed on NOSI, and the importance of NOSI aspects within the related socio-scientific issues was discussed explicitly, especially in the selected outdoor places (see Table 1). For example, while discussing fossils belonging to creatures that lived in ancient years, the concepts of scientific data and evidence and their differences were discussed on a science museum tour. After the elucidation of dinosaur bone activity was carried out, the combination of previously known and collected data develops scientific explanations. Another example is that during online visit to the laboratories in the university, it was explicitly discussed what the scientific inquiry is and what features it is built on, the importance of the research question in science, and how it affects the research process. It was clearly expressed that there are different methods in science by interviewing the professors from different fields in both social science and science and emphasized the differences in the methods in their studies. In the online meetings that followed, reading passages were given to the pre-service teachers every week and discussions were made on both these reading passages and the things learned in the outdoor places visited online. In addition, every week, students were asked to write a daily reflection essay about what they had learned on that week. After the 8-week online outdoor teaching focusing on the NOSI, the mid-VASI questionnaire was completed by the pre-service teachers.

Table 1. *First eight weeks of implementation.*

Week	Outdoor Environments	Explanations	NOSI Aspects Intended to Teach	NOSI Generic Activities
1	Seven Mills Nature Park	Investigation of plants and animals in danger of extinction	-Begins with question -Data/Evidence	Tricky track
2	Meteorology Center	Investigation of the cause and effects of global climate change	-The same procedures do not get the same results -Inquiry procedures influence results	Global warming
3	University Laboratories	Interviewing professors from different fields and discussing different methods in science.	-Multiple scientific methods -Conclusions consistent with data	Future scientists
4	Blood Donation Center	Investigation for blood groups and the importance of donating blood	-Conclusions consistent with the data	Where does my genetics come from?
5	Gas and Electricity Generation and Storage Facility	Knowledge about recycling, environmental problems and solutions that may arise as a result of human activities	-Begins with question -Procedures by the question asked -Data/Evidence	My project for environmental problems
6	Hydroelectric Power Plant	Transformation of energy. The benefits and harms of hydroelectric power plants	-Data/Evidence -Conclusions consistent with the data	Argumentation (Is hydroelectric power plant harmful or useful?)
7	Archeology Museum	Having knowledge about fossils and creatures that lived years ago.	-Explanations are developed from the data -Data/Evidence	Dinosaur bones
8	International Airport	Observing the effect of friction force on kinetic energy	-Procedures by the question asked -Conclusions consistent with the data	Airplane runway and aircraft tires

2.2.2. *Stage 2-3. (Exploration + Transformation)*

In the remaining weeks, each pre-service teacher was asked to prepare a lesson plan by associating at least one NOSI aspect with a socio-scientific topic. First, a sample plan was introduced by the researchers, and then the pre-service teachers were asked to prepare their plans.

2.2.3. *Stage 4-5. (Presentation + Reflection)*

After the researchers gave feedback on the plans made, each pre-service teacher had the opportunity to present their plans in the outdoor learning environment of their choice. They video-recorded their presentations and sent them to the researchers. Each student was asked to write a self-evaluation essay in which they evaluated themselves and the whole process after microteaching. In these essays, they were asked to express their strengths and weaknesses and the parts of the process that they had the most difficulty with.

At the end of 14 weeks, the post-VASI questionnaire was completed again, and follow-up interviews were carried out to determine the progress of the pre-service teachers' views of the NOSI. In the final interview, the views of pre-service teachers on the effect of the outdoor learning course on NOSI views were revealed.

2.3. Data Collection Tools

The VASI was developed by Lederman et al. (2014) and adapted into Turkish using the retranslation method by Çavuş-Güngören and Öztürk (2016). The VASI questionnaire is a context-based 7 open-ended questionnaire that explores the views of students in the 6th-grade or above, teachers, and pre-service teachers about the aspects of scientific inquiry targeted by the National Science Education Standards (Lederman et al., 2014). Due to the nature of the questionnaire, participants are challenged to think critically about scientific inquiry and the underlying reasons for their thoughts. It is emphasized that this reasoning should be examined further with follow-up interviews (Lederman et al., 2014). It is preferred that the VASI be given under controlled conditions with no set time limit for completion. VASI responders typically take 30-45 minutes to complete the questionnaire. Participants are encouraged to write as much information as possible on relevant items and to provide illustrative examples to help support their explanations.

2.4. Data Analysis

In its analysis, the VASI Questionnaire developers presented a table, the questions of which corresponded to NOSI aspects (Lederman et al., 2014 p.75). Analyses were made based on this table. In addition, all items on the VASI questionnaire were analyzed holistically to generate a profile of each pre-service teacher's views across the targeted aspects of NOSI. For example, if a participant states that researchers who use different methods in one item can achieve the same or different results people who achieve the same results in another item should have followed the same method, the participant is considered to be in mixed view. It should be emphasized that the answers given to the items in the VASI are not independently scored as correct or incorrect and the participant's view on the relevant aspect of NOSI is classified according to the NOSI continuum scale, taking into account the responses to all items holistically. Using the NOSI views continuum scale, a profile for each participant was developed, describing their views on a continuum from naive "-" to mixed "(+)" to increasing levels of informed "+, ++, +++" (Schwartz et al., 2008). If pre-service teachers had an insufficient view or an incompatible view about the targeted NOSI aspects, their responses were coded as naive (-). The pre-service teachers' responses were coded as informed if they had a sufficient view about the targeted aspect that was compatible with the literature. The informed level "+", "++", and "+++" varies depending on the explanations given appropriate examples with their sentences. The pre-service teachers' responses were coded as mixed "(+)" if their responses showed inconsistency within the questionnaire or during the interviews.

The final interviews were analyzed by using content analysis. The content analysis consisted of coding data, creating categories and themes from codes, and visualizing data (McMillan & Schumacher, 2010). A reasonable amount of data (20%) (Lincoln & Guba, 1985) was reviewed and analyzed by the authors and two leading independent experts. After the analyses were completed, the researchers discussed the analysis findings until 90% agreement was reached. The authors analyzed the remainder of the data based on the commonalities obtained in the inquiry audit.

To increase the consistency of the research, two field experts were consulted about the results of the analysis. To ensure the verifiability of the findings, information about the sample from which the data were collected was presented (Merriam, 2018). The researchers conducting this study have experience and research in teaching NOSI. They also have experience in conducting qualitative research. These increase the verifiability of the findings. To increase the credibility of the findings obtained, the two researchers worked together in the data collection and analysis process. A semi-structured interview form was used to collect in-depth focused data and the data were ensured to reach a saturation point. To ensure the transferability of the research, direct quotations from the participants were made while presenting the findings.

3. FINDINGS

According to the analysis, the pre-service teachers had mostly mixed or naive views regarding the targeted NOSI aspects at the beginning of the study. Pre-service teachers generally had naive views in some NOSI aspects, such as “scientific data are not the same as scientific evidence”, “all scientists performing the same procedures may not get the same results”, “inquiry procedures are guided by the question asked”, and “scientific investigations all begin with a question and do not necessarily test a hypothesis” (see Figure 2). Some of the representative quotes expressed by the pre-service teachers are provided below.

“Scientific research mostly does not start with a question.” (PST12_pre-VASI)

“Data are correct or incorrect results that come from the experiment. However, the evidence is exact information.” (PST41_ pre-VASI Interview)

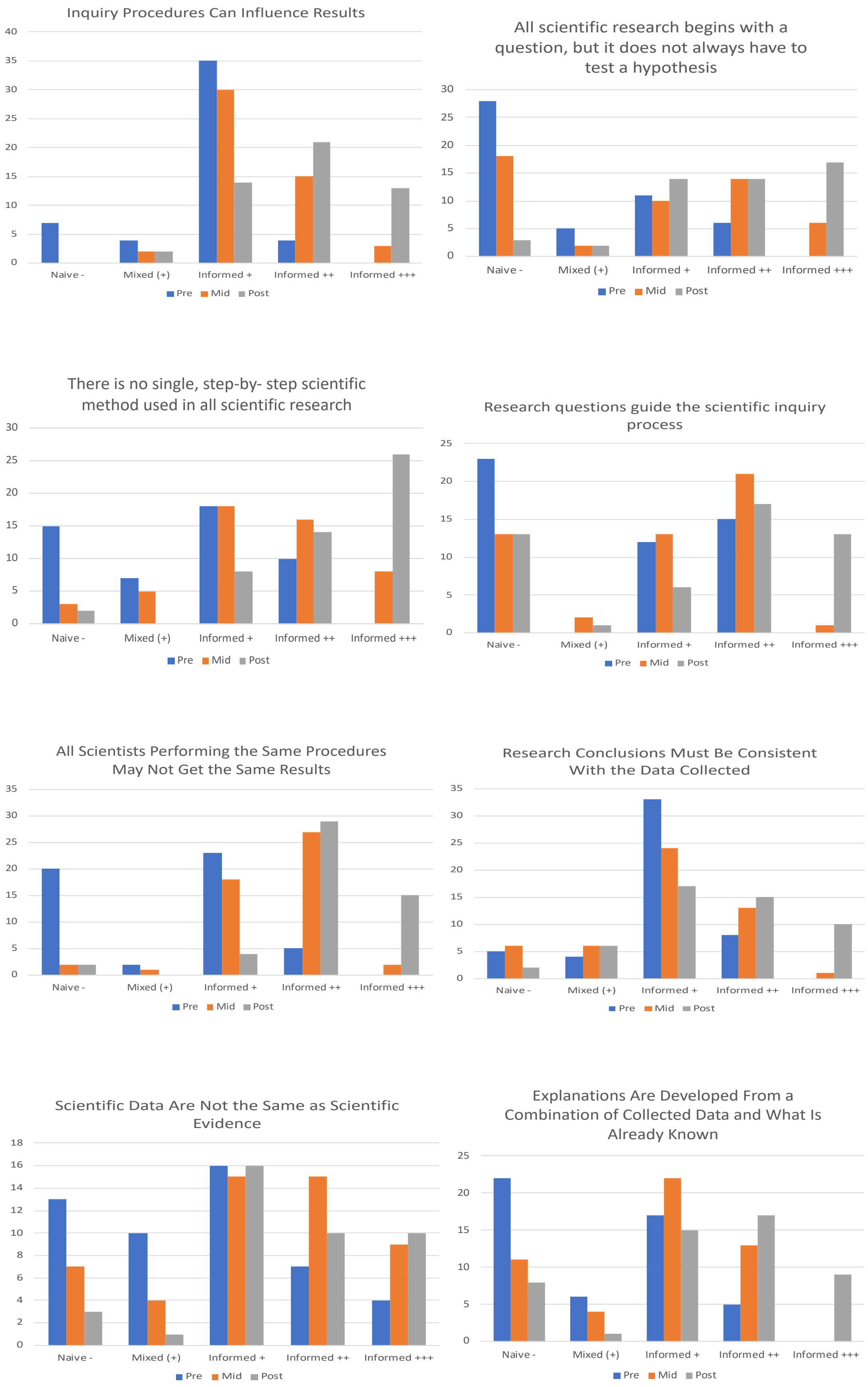
After the first eight weeks of online NOSI teaching in outdoor learning environments, a positive development was observed in the NOSI views of pre-service teachers, but this development was not at the desired level for all of them (see Figure 2). According to the analysis of mid-VASI responses and interviews, some pre-service teachers were still in the naive and mixed views of some NOSI aspects. The aspects with the highest improvement in the views of pre-service teachers are "there is no single scientific method", "scientists can reach different results even if they follow the same procedures", and "the inquiry procedures affects the research results". Some representative quotations of the pre-service teachers' NOSI views in the middle of the study are provided below.

“There is no one single scientific method. Scientists can follow more than one method. Qualitative and quantitative research methods can be given as examples for different methods.” (PST14_mid-VASI)

“Scientists are people with different experiences, theoretical assumptions, cultures and imaginations, so even if the same methods are followed, different results may emerge.” (PST7_mid-VASI interview)

According to the analysis of pre-service teachers’ post-VASI responses and interviews, after the pre-service teachers prepared lesson plans and follow-up micro teachings, almost all participants dramatically improved their views of the NOSI. The shifting was mostly seen from the mixed view to an increased level of the informed range (see Figure 2).

Figure 2. *Participants' views on NOSI aspects.*



Some of the pre-service teachers still have naive views in only a few NOSI aspects (i.e., “questions guide the research process”, and “explanations consist of collected data and prior knowledge”). Some representative quotations of the pre-service teachers’ informed NOSI views at the end of the study are provided below.

“Data are collected through the observations or experiences. The evidence is an argument that we use to support our claims by interpreting the data.” (PST23_post-VASI interview)

“Scientific research starts with questions. For example, Leeuwenhoek asked himself the question of "What can I see if I examine the pond water in the garden?" and discovered unicellular microorganisms starting from the question". (PST48_post-VASI)

As a result of the analysis of final interviews, teachers thought that some factors might have affected their NOSI views throughout the course. These factors were "online NOSI instruction in outdoor learning environments", "lesson plan preparing + microteachings", "feedback", and "classroom discussions" (see Figure 3).

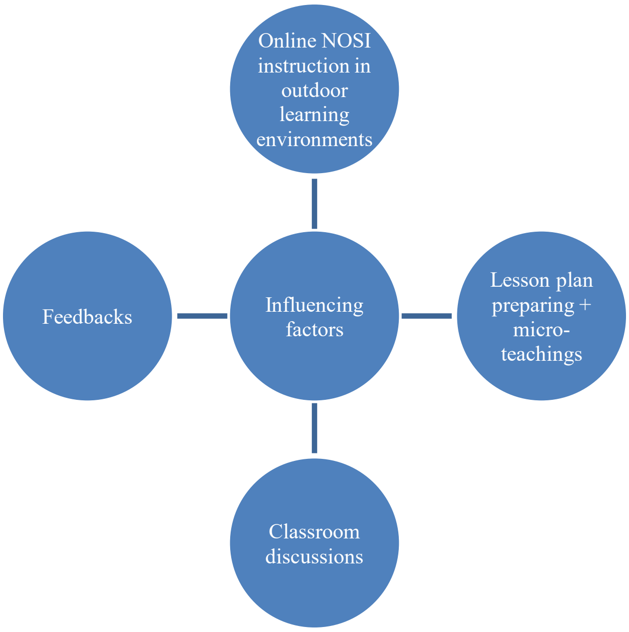
One of the factors stated by the pre-service teachers, which they think is effective, is the effect of online NOSI teaching in open-air learning environments that represents the first eight weeks of the course. In addition to the changes seen in Figure 2, the pre-service teachers also expressed how the first 8-week course affected their NOSI views.

"I think your online instruction in outdoor learning environments was very useful for me to understand the scientific inquiry. Therefore, I did not have much difficulty in preparing my plans." (PST17_Final interview)

"Your instruction on the nature of scientific inquiry by relating it to the contexts we encounter in daily life made it easier for me to understand the aspects." (PST33_Final interview)

"I think, I understood the nature of scientific inquiry better by watching your instructions in the records. I was able to reinforce what I had not fully understood by rewatching." (PST27_Final interview)

Figure 3. *Influencing factors on pre-service teachers’ NOSI views.*


Online NOSI instruction in outdoor learning environments
Feedbacks
Influencing factors
Lesson plan preparing + micro-teachings
Classroom discussions

The other factor is the effect of pre-service teachers' lesson plan preparation and teaching practices. The participants explained how the lesson plan preparation and teaching practices affected their NOSI views.

"In fact, these lesson plans and practices have been very useful for me to improve myself and understand the scientific inquiry. I noticed my misconceptions and had a chance to fix them all." (PST47_Final interview)

"I think that preparing lesson plans and follow-up practice improves our views about scientific inquiry. Now, I feel more confidence myself to teach the aspects of nature of the scientific inquiry." (PST7_Final interview)

"While I was preparing my lesson plan, I had the opportunity to review my view about scientific inquiry. I tried to make up for my deficiencies, I think that teaching something is the best way to learn it." (PST13_Final interview)

Another factor stated by the pre-service teachers, which they thought to be effective, was feedback from the instructors. They expressed the importance of feedback on their NOSI views.

"Seeing each other's plans and your feedback helped us to improve ourselves." (PST12_Final interview)

"Thanks to the feedback we received from our teacher in the lesson, we saw our shortcomings, which gave me an idea about how I could do it more appropriately in my last plan." (PST40_Final interview)

"After each lesson plan and teaching practice, I had the opportunity to make up for the deficiencies thanks to the feedback I received from you." (PST15_Final interview)

Last but not least, pre-service teachers thought that their NOSI views might also be influenced by weekly classroom discussions just made after their teaching practices. They expressed as:

"I realized that I did not understand the nature of scientific inquiry at first, or rather, I had difficulty in understanding it. Once I understood the topic, I had a hard time applying it to my plan and activity, but after two exercises and classroom discussions, I thought I understood it better." (PST23_Final interview)

"I thought that I understood it, but when I tried to put it into practice, I realized that I did not quite understand it. After our weekly discussions and the examples I saw, I think I understood the scientific inquiry better." (PST34_Final interview)

4. DISCUSSION and CONCLUSION

The findings indicate that almost all participants improved their views of NOSI in a positive manner through explicit/reflective online outdoor NOSI teaching and teachers' own experience through lesson planning and practice. Explicit reflective teaching is an effective method for developing learners' NOSI views (Lederman, 2019; Mesci et al., 2020; Schwartz and Crawford, 2004). However, the findings of the current study show that when pre-service teachers are provided with the opportunity to prepare a lesson plan and practice after explicit/reflective NOSI instruction, their NOSI views dramatically improve. The findings of this study are valuable, as they show how 'learning through teaching' makes a dramatic change in participants' NOSI views. Explicit/reflective NOSI teaching, which does not provide participants with the experience of transferring what they have learned to their practices, is limited on its own (Mesci et al., 2020). The importance of "teaching experience" in improving pre-service teachers' NOSI views in teacher training programs should not be overlooked. Other studies in the literature also confirm that lesson plan preparation and teaching experience are effective in developing participants' NOSI views (Gess-Newsome, 2002; Lederman and Lederman, 2004; Lederman and Lederman, 2012; Lotter et al., 2009, Mesci et al., 2020).

As Lederman and Lederman (2012) argued, learners more easily adopt what they see from the acts of their peers, rather than what is modeled by professional educators. In the present study, the pre-service teachers not only provided explicit/reflective NOSI instruction and teaching experience but also had the opportunity to see and criticize each other's plans and practices (reflective observations) and received feedback. In parallel with the current study, research shows that reflective discussions and mentors' feedback support the pedagogical development of teachers and pre-service teachers and facilitate the implementation of effective teaching strategies (Melville et al., 2008; Singer, 2005; Yung et al., 2007). Based on the literature (Lederman and Lederman, 2012; Lotter et al., 2009; Mesci et al., 2020), these interactive dialogs and discussions within the group and the feedback received during the process are effective in clearing their current misconceptions about NOSI and improving their naive views.

In the present study, the experiences in outdoor learning environments may also affect the development of pre-service teachers' views. Outdoor learning experiences provide individuals with awareness of the science-society relationship that classroom-based learning environments cannot gain; that is, they provide real contexts from life and offer a more realistic learning experience by practicing (Akgül & Arabacı, 2020; Gürsoy, 2018). Based on the emphasis that different alternative teaching methods should be examined in NOSI teaching (Lederman et al., 2019), the NOSI views of pre-service teachers in outdoor learning environments were developed. The current study shows examples of how to use an explicit/reflective approach to socio-scientific issues in outdoor learning environments. The design of the study may add to the literature and may also be of interest to science educators. In another study, Deniz and Akerson (2013) developed primary school teachers' NOS and NOSI views by integrating explicit reflective teaching with language arts. Mesci et al. (2020) developed pre-service teachers' NOSI views through argumentation-based NOSI teaching in laboratories. These studies may encourage researchers who want to find alternative ways of teaching NOSI in different contexts.

This study is uniquely focused on the changing or unchanging views in outdoor learning environments by using fully online teaching. Developing the views of pre-service teachers related to NOSI with this alternative method (online-outdoor learning) may set an example for NOSI teaching in fully online education. In addition, it is obvious that effective in-service learning activities to be planned in online learning environments can contribute to the professional development of more teachers economically. Although the findings of our study are consistent with other studies in the literature, it should be considered that the findings of this study are limited to the context. Developing teachers' views and teaching skills on NOSI is not easy and takes a long time (Lederman & Lederman, 2012). It is also important to investigate the long-term effects of the results of this study on the participants.

Finally, it is known that early graders' teachers are mostly experts in language teaching (Akerson, 2007) and generally do not have a strong science background (Anderson, 1999). It is essential to provide content and pedagogical knowledge and teaching experience for teaching inquiry-based science to teachers who do not see themselves as science teachers (Lederman & Lederman, 2004). Therefore, these teachers may not be able to teach science effectively without the support of a well-designed professional development even though they are encouraged to teach science (Deniz & Akerson, 2013). Thus, these pre-service and in-service teachers need more opportunities to learn to teach the NOSI than middle and high school science teachers due to their low science background. Non-science major pre-service teachers and researchers who will work with teachers may consider this. Considering the emphasis in international documents on laying the foundations of science literacy from the kindergarten (Lederman & Bartels, 2018), further studies should be conducted to investigate and develop the views of children and teachers who have an indisputable influence on children's learning.

Acknowledgments

A part of this paper was presented at the National Association for Research in Science Teaching (NARST), Vancouver, BC, Canada, March 27-30, 2022.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Kastamonu University, 12.10.2020/3-35.

Authorship Contribution Statement

Eda Erdas Kartal: Determining the purpose of the research and the research design; implementing the activities, data collection, and analysis; and writing the introduction, method, and discussion parts of the article. **Gunkut Mesci:** Determining the purpose of the research and the research design; implementing the activities; data collection and analysis; and writing the methods, findings, and discussion parts of the article. Both authors checked the mutually written sections and made necessary revisions.

Orcid

Eda Erdas Kartal https://orcid.org/0000-0002-1568-827X
Gunkut Mesci https://orcid.org/0000-0003-0319-5993

REFERENCES

Akerson, V.L. (Ed.) (2007). *Interdisciplinary language arts and science instruction in elementary classrooms: Applying research to practice*. Erlbaum: NJ.

Akgül, G.D. & Arabacı, S. (2020). Okul dışı öğrenme ortamlarına yönelik fen bilgisi öğretmenlerinin görüşleri [The views of science teachers on the use and application of out of school learning environments]. *Uluslararası Eğitim Araştırmacıları Dergisi, 3*(2), 276-291.

Alisinanoğlu, F. & Özbey, S. (2011). *Okul öncesinde fen eğitimi* [*Preschool science education*]. Maya Akademi.

American Association for the Advancement of Science [AAAS] (1993). *Benchmarks for science literacy: A project 2061 report*. Oxford University Press.

Andersson, B. (1999). Pupils' conceptions of matter and its transformations. *Studies in Science Education, 2*(1), 53-85.

Andiema, N.C. (2016). Effect of child-centered methods on teaching and learning of science activities in preschools in Kenya. *Journal of Education and Practice, 7*(27), 1-9.

Aydemir, S., Ugras, M., Cambay, O., & Kilic, A. (2017). Prospective preschool teachers' views on the nature of science and scientific inquiry. *Üniversitepark Bülten, 6*(2), 74-87. https://doi.org/10.22521/unibulletin.2017.62.6

Aydeniz, M., Baksa, K. & Skinner, J. (2011). Understanding the impact of an apprenticeship-based scientific research program on high school student's understanding of scientific inquiry. *Journal of Science Education and Technology, 20*(4), 403-421. https://doi.org/10.1007/s10956-010-9261-4

Bartels, S., & Lederman, J. (2022). What do elementary students know about science, scientists, and how they do their work?. *International Journal of Science Education, 44*(4), 627-646. https://doi.org/10.1080/09500693.2022.2050487

Baykara, H. & Yakar, Z. (2020). Pre-service science teachers' views about scientific inquiry: The case of Turkey and Taiwan. *Turkish Online Journal of Qualitative Inquiry, 11*(2), 161-192. https://doi.org/10.17569/tojqi.618950

Bell, R.L., Blair, L.M., Crawford, B.A. & Lederman, N.G (2003). Just do it? Impact of a science apprenticeship program on high school students' understandings of the nature of science and scientific inquiry. *Journal of Research in Science Teaching, 40*(5), 487-509.

Brock, K.L. & Cameron, B.J. (1999). Enlivening political science courses with Kolb's learning preference model. *Political Science and Politics, 32*(2), 251-256.

Burns, M. (2011). *Distance education for teacher training: modes, models, and methods*. Education Development Center, Inc.

Camlıbel Çakmak, Ö. (2014). *Okul öncesi dönemde fen eğitimi ve öğretmenin rolü* [*Science education and the role of the teacher at the preschool period*]. *In M. Çetin & Ç. Şahin (Eds.), Okul öncesi dönemde fen eğitimi* [*Science education at the preschool period*] (s.29-47). Pegem Akademi.

Crawford, B.A., Zembal Saul, C., Munford, D. & Friedrichsen, P. (2005). Confronting prospective teachers' ideas of evolution and scientific inquiry using technology and inquiry-based tasks. *Journal of Research in Science Teaching, 42*(6), 613-637.

Crawford, B.A., Capps, D., Meyer, X., Patel, M. & Ross, R.M. (2010, April). Supporting teachers in complex situations: Learning to teach evolution, nature of science, and scientific inquiry. A paper presentation at the American Educational Research Association Annual Meeting-Denver, Colorado.

Creswell, J.W. (2012). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.). Pearson.

Çavuş Güngören S., & Öztürk E. (2016). *Turkish adaptation of the views about scientific inquiry VASI and examine pre-service mathematics teachers' views about scientific inquiry*. VI. International Congress on Research in Education (ICRE).

Deniz, H. & Akerson, V. (2013). Examining the impact of a professional development program on elementary teachers' views of the nature of science and nature of the scientific inquiry, and science teaching efficacy beliefs. *The Electronic Journal for Research in Science & Mathematics Education, 17*(3), 1-19.

Dori, Y.J., & Tall, R.T. (2000). Formal and informal collaborative projects: engaging in the industry with environmental awareness. *Science Education, 84*(1), 95-113.

Dudu, W.T. (2014). Exploring South African high school teachers’ conceptions of the nature of scientific inquiry: A case study. *South African Journal of Education, 34*(1), 1-18.

Erdas Kartal, E., Cobern, W.W., Dogan, N., Irez, S., Cakmakci, G., & Yalaki, Y. (2018). Improving science teachers’ nature of science views through an innovative continuing professional development program. *International Journal of STEM education, 5*(1), 1-10. https://doi.org/10.1186/s40594-018-0125-4

Gelişli, Y. (2015). Practices of distance education for teacher training: History and development. *Journal of Research in Education and Teaching, 3*(1), 313-321.

Gess-Newsome, J. (2002). The use and impact of explicit instruction about the nature of science and science inquiry in an elementary science methods course. *Science & Education, 11*(1), 55-67.

Gürsoy, G. (2018). Fen öğretiminde okul dışı öğrenme ortamları (Outdoor learning environments in science education). *Electronic Turkish Studies, 13*(11), 623 649. http://dx.doi.org/10.7827/TurkishStudies.13225

Karışan, D., Bilican, K., & Şenler, B. (2017). Bilimsel sorgulama hakkında görüş anketi: Türkçeye uyarlama, geçerlik ve güvenirlik çalışması (The adaptation of the views about scientific inquiry questionnaire: A validity and reliability study). *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 18*(1), 326-343. https://doi.org/10.17679/inuefd.307053

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall.

Koyunlu-Ünlü, Z. (2020). Improving pre-service teachers' science process skills and views about scientific inquiry. *Journal of Theoretical Educational Science, 13*(3), 474-489. http://dx.doi.org/10.30831/akukeg.626165

Kubat, U. (2018). Okul dışı öğrenme ortamları hakkında fen bilgisi öğretmen adaylarının görüşleri (Opinions of pre-service science teachers about outdoor education). Mehmet

Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 48(1), 111-135. https://doi.org/10.21764/maeuefd.429575

Laugksch, R.C. (2000). Scientific literacy: A conceptual overview. *Science Education, 84*(1), 71-94.

Lederman, J.S. & Lederman, N.G. (2004, April). *Early elementary students' and teacher's understandings of nature of science and scientific inquiry: Lessons learned from project ICAN*. Paper Presented at the Annual Meeting of the National Association for Research in Science Teaching, Vancouver, British Columbia.

Lederman, N.G. (2006). Research on nature of science: reflections on the past, anticipations of the future. *Asia-Pasific Forum Science Learning and Teaching, 7*(1), 1-11

Lederman, J.S. (2012). *Development of a valid and reliable protocol for the assessment of early childhood students' conceptions of the nature of science and scientific inquiry*. A Paper Presented at the Annual Meeting of the National Association of Research in Science Teaching, Indianapolis, IN.

Lederman, N., & Lederman, J. (2012). *Nature of scientific knowledge and scientific inquiry: Building instructional capacity through professional development*. In B.J. Fraser, K. Tobin & C.J. McRobbie (Eds.), Second international handbook of science education (24th ed., pp. 335–359). Springer

Lederman, J.S., Bartels, S.L., Liu, C. & Jimenez, J. (2013). *Teaching the nature of science and scientific inquiry to diverse classes of early primary-level students*. A Paper Presented at the Annual Meeting of the National Association for Research in Science Teaching (NARST), San Juan, PR, USA.

Lederman, N.G., Antink, A., & Bartos, S. (2014). Nature of science, scientific inquiry, and socio scientific issues arising from genetics: A pathway to developing a scientifically literate citizenry. *Science & Education, 23*(2), 285-302.

Lederman, J.S., Lederman, N.G., Bartos, S.A., Bartels, S.L., Meyer, A.A., & Schwartz, R.S. (2014). Meaningful assessment of learners' understandings about scientific inquiry-The views about scientific inquiry (VASI) questionnaire. *Journal of Research in Science Teaching, 51*(1), 65-83. https://doi.org/10.1002/tea.21125

Lederman, N.G. (2019). Contextualizing the relationship between the nature of scientific knowledge and scientific inquiry. *Science & Education, 28*(1), 249 267. https://doi.org/10.1007/s11191-019-00030-8

Lederman, J.S., Lederman, N.G., Bartels, S., Jimenez, J., Akubo, M., et al. (2019). An international collaborative investigation of beginning seventh-grade students' understandings of scientific inquiry: Establishing a baseline. *Journal of Research in Science Teaching, 56*(4), 486-515. https://doi.org/10.1002/tea.21512

Lederman, J.S., & Bartels, S.L. (2018). *Assessing the ultimate goal of science education: Scientific literacy for all*! In S. Kahn (Ed.), Toward inclusion of all learners through science teacher education (pp. 277–285). Brill.

Lincoln, Y.S., & Guba, E.G. (1985). *Naturalistic inquiry*. Sage.

Lotter, C., Singer, J., & Godley, J. (2009). The influence of repeated teaching and reflection on pre-service teachers' views of inquiry and nature of science. *Journal of science teacher education, 20*(6), 553-582. https://doi.org/10.1007/s10972-009-9144-9

McMillan, J.H., & Schumacher, S. (2010). *Education research: Evidence-based inquiry* (7th ed.). Pearson Education, Inc.

Merriam, S.B. (2013). *Nitel araştırma: Desen ve uygulama için bir rehber* (S.Turan, Trans.) [A guide to qualitative research patterns and practice] (Trans. S. Turan). Nobel Press.

National Research Council [NRC] (2000). *Inquiry and the national science education standards*. National Academy Press.

National Research Council [NRC] (2012). *Inquiry and the national science education standards*. National Academic Press.

National Research Council (NRC). (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing*? National Academies Press.

NGSS Lead States (2013). *Next generation science standards: For states, by states*. The National Academy Press.

Norris S.P., & Phillips, L.M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(1), 224-240.

Melville, W., Fazio, X., Bartley, A., & Jones, D. (2008). Experience and reflection: Pre-service science teachers' capacity for teaching inquiry. *Journal of Science Teacher Education, 19*(1), 477-494.

Mesci, G., & Schwartz, R.S. (2017). Changing pre-service science teachers' views of nature of science: Why some conceptions may be more easily altered than others. *Research in Science Education, 47*(2), 329-351. https://doi.org/10.1007/s11165-015-9503-9

Mesci, G., Çavuş Güngören, S., & Yesildag Hasancebi, F. (2020). Investigating the development of pre service science teachers' NOSI views and related teaching practices. *International Journal of Science Education, 42*(1), 50-69. https://doi.org/10.1080/09500693.2019.1700316

Mesci, G., & Kartal, E.E. (2021). Science teachers' views on the nature of the scientific inquiry. *Bartın University Journal of Faculty of Education, 10*(1), 69-84. https://doi.org/10.14686/buefad.797246

Metin-Peten, D. (2021). Influence of the argument-driven inquiry with explicit-reflective nature of scientific inquiry intervention on pre-service science teachers' understandings of the nature of the scientific inquiry. *International Journal of Science and Mathematics Education, 20*(1), 921-941. https://doi.org/10.1007/s10763-021-10197-8

Ministry of National Education (MoNE) (2018). *Elementary and secondary science curriculum*. National Education Press.

Opfer, J.E., & Siegler, R.S. (2004). Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. *Cognitive psychology, 49*(4), 301-332.

Organization for Economic Co-operation and Development (OECD) (2003). *The PISA 2003 assessment framework-mathematics, reading, science, and problem-solving knowledge and skills*. OECD Publishing.

Osborne, J. (2014). *Scientific practices and inquiry in the science classroom*. In N. Lederman & S. Abell (Eds.), The handbook of research on science education, vol. II (pp. 579–599). Taylor and Frances Group.

Penn, M., Ramnarain, U., Kazeni, M., Dhurumraj, T., Mavuru, L., & Ramaila, S. (2021). South African primary school learners' understandings of the nature of scientific inquiry. *Education 3-13, 49*(3), 263-274. https://doi.org/10.1080/03004279.2020.1854956

Pérez, B.C., & Díaz-Moreno, N. (2022). Promoting pre-service primary teachers' development of NOSI through specific immersion and reflection. *EURASIA Journal of Mathematics, Science and Technology Education, 18*(3), 1-16. https://doi.org/10.29333/ejmste/11795

Pritchard, A. (2017). *Ways of learning: Learning theories for the classroom*. Routledge.

Reid, J., Forrestal, P. & Cook, J. (1989). *Small group learning in the classroom: Scarborough (Australia)*. English and Media Centre.

Rickinson, M., Dillon, J., Teamey, K., Morris, M., Choi, M.Y., Sanders, D., & Benefield, P. (2004). *A review of research on outdoor learning*. Field Studies Council.

Roberts, D.A., & Bybee, R.W. (2014). Scientific literacy, science literacy, and science education. *Handbook of Research on Science Education, 2*(1), 545-558.

Roehrig, G.H., & Luft, J.A. (2004). Constraints experienced by beginning secondary science teachers in implementing scientific inquiry lessons. *International Journal of Science Education, 26*(1), 3-24

Schwartz, R.S., & Lederman, N.G. (2002). "It's the nature of the beast": The influence of knowledge and intentions on learning and teaching nature of science. *Journal of Research in Science Teaching, 39*(3), 205-236.

Schwartz, R.S. (2004). *Epistemological views in authentic science practices: A cross discipline comparison of scientists' views of nature of science and scientific inquiry* [Unpublished doctoral dissertation]. Oregon State University.

Schwartz, R.S., & Crawford, B.A. (2004). *Authentic scientific inquiry as a context for teaching nature of science: Identifying critical elements for success*. In L. Flick & N.G. Lederman (Eds.), Scientific inquiry and nature of science: Implications for teaching, learning, and teacher education (pp. 331–356). Kluwer Publishing Co.

Schwartz, R.S., Lederman, N., & Lederman, N. (2008, March). *An instrument to assess views of scientific inquiry: The VOSI questionnaire*. Paper presented at the international conference of the National Association for Research in Science Teaching. Baltimore, MD.

Singer, J. (2005). *Integrating technology and pedagogy: The ideas, the shift and the targets*. In S. Rhine & M. Bailey (Eds.), Integrated technologies, innovative learning: Insights from the PT3 program (pp. 199–215). International Society for Technology in Education.

Tatar, N., & Bağrıyanik, K.E. (2012). Opinions of Science and Technology Teachers about Outdoor Education. *Ilkogretim Online, 11*(4), 883-896.

Tytler, R., & Peterson, S. (2003). Tracing young children's scientific reasoning. *Research in Science Education, 33*(4), 433-465.

Wang, J. & Zhao, Y. (2016). Comparative research on the understandings of nature of science and scientific inquiry between science teachers from Shanghai and Chicago. *Journal of Baltic Science Education, 15*(1), 97.

Wong, S., & Hodson, D. (2010). More from the horse's mouth: What scientists say about science as a social practice. *International Journal of Science Education, 32*(11), 1431-1463.

Yung, B., Wong, S., Cheng, M., Hui, C., & Hodson, D. (2007). Tracking pre-service teachers' changing conceptions of good science teaching: The role of progressive reflection with the same video. *Research in Science Education, 37*(1), 239–259.

Yurt, Ö. (2015). *Okul öncesi dönemde fen eğitimi [Science education at preschool period]*. In F. Şahin (Ed.), *Her yönüyle okul öncesi eğitim: Okul öncesi dönemde fen eğitimi [Preschool education in all aspects: Science education at preschool period],* (ss.11-20). Hedef CS Yayıncılık ve Mühendislik.

Zion, M., & Mendelovici, R. (2012). Moving from structured to open inquiry: Challenges and limits. *Science Education International, 23*(4), 383–399.

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 300–316

https://doi.org/10.21449/ijate.1115317

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Click here for the Turkish version of this article.

Adaptation of teachers' perceptions of grading practices scale to Turkish and examination of measurement invariance

Yesim Ozer Ozkan[1,*], Meltem Acar Guvendir[2], Emre Guvendir[3]

[1]Gaziantep University, Faculty of Education, Department of Educational Sciences, Türkiye
[2]Trakya University, Faculty of Education, Department of Educational Sciences, Türkiye
[3]Trakya University, Faculty of Education, Department of Foreign Languages Education, Türkiye

ARTICLE HISTORY

Received: May 11, 2022

Revised: Oct. 19, 2022

Accepted1: Nov. 1, 2022

Keywords:
Grading practices,
Scale adaptation,
Teachers' perceptions,
Student achievement,
Measurement invariance.

Abstract: The purpose of this research is to adapt the Teacher Perceptions of Grading Practices Scale into Turkish and to examine the measurement invariance. This scale, which examines teachers' perceptions of grading methods, has six components: importance, usefulness, student effort, student ability, teachers' grading patterns, and perceived self-efficacy of the grading process. Before adapting the scale, permission was first acquired from the researcher who developed it. To ensure linguistic comparability, bilingual translators were recruited in the second phase. The semantic, experiential, conceptual, and idiomatic equivalence between the two variants of the scale were evaluated. The original and adapted scales were administered to a group of English teachers twice at a predetermined interval, and the consistency between the two applications was analyzed due to the fact that the language employed in the original test was a widely spoken group. Confirmatory Factor Analysis (CFA) was used to examine the factor structure of the original scale. Cronbach's α and McDonald's ω coefficients were calculated for the reliability of the data obtained from the scale. Finally, the measurement invariance of the scale according to gender was examined by using Multiple Group Confirmatory Factor Analysis (MGCFA), and it was determined that the measurement model fulfilled the criteria of complete gender-group invariance.

1. INTRODUCTION

Measurement and evaluation are intertwined processes that entail detection and decision-making. While measuring entails observing certain circumstances, events, or things and describing the findings with numbers or symbols, evaluation is making a decision based on an objective or criterion associated with the measurement obtained at the end of this process. In this respect, no evaluation can be made without measurement. Teachers must conduct measurements in order to make judgments about their students' achievement. With this in mind, they aim to elicit information regarding their students' achievement with the tests or assignments they have utilized.

*Corresponding Author: Meltem Acar Guvendir ✉ meltemacar@trakya.edu.tr Trakya University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

Measurement and evaluation are primarily concerned with student achievement. The purpose of post-instruction evaluation is to measure and interpret the change in student behavior induced by the teaching activities. The performance of students is compared to established guidelines or norms. As a result of the evaluation, feedback is provided for all instructional components, and quality feedback is typically the most important part of learning. (Biggs, 2001; Eraut, 2004). At this stage, what matters is that feedback is provided on time, sufficiently, and consistently. (Harlen, 2005; Serban, 2004).

Throughout this process, the teacher attempts to provide pupils with feedback regarding their progress based on the grades they have earned. Grading is the method of allocating a student to a continuum based on impressions, evidence, or a combination of the two (Anderson, 2018). But, what is the purpose of grading? Is it absolutely required to assign grades to students in order to evaluate them?

Campbell (1921) claimed that grading serves two critical functions. The first objective is to urge students to exert greater effort, and the second goal is to offer teachers information to help them improve their instruction. Bailey & McTighe (1996) stated that a third aim of grading is to provide information about student learning to a variety of populations that need and/or require information about how well students are learning or advancing in order to make appropriate judgments about them. The grades serve as a means of disseminating student success to students, parents, teachers, postsecondary institutions, and employers.

Salend and Duhaney (2002) further extended the purposes of grading to achievement, progression, effort, comparison, instructional planning, program effectiveness, motivation, communication, education and career planning, relevance, and accountability. The grading procedure serves as a demonstration of the teacher's knowledge of the program objectives. Simultaneously, the teacher can ascertain the students' learning issues and tailor their instruction to their specific needs. Thus, the program's effectiveness can be determined. Grading is used to track students' progress in learning over time, to compare students' competencies, and to monitor students' progress and efforts. This way, feedback may be provided to families with students and the level of support required can be determined. Thus in this manner, grading enables students to develop career strategies. Finally, grades are used to determine whether or not a student is eligible to graduate from a program. Consequently, indicators of academic achievement can be provided.

The teacher's role in this evaluation process is to select the behaviors that best reflect a student's progress, to develop and implement measurement methods, and to interpret the results appropriately (Küçükahmet, 2005). Gardner *et al.* (1997) identified the following critical points that a teacher should consider when assigning grades:

1. Explain the school's grading system to students in advance.
2. State explicitly the grading rules and requirements.
3. Assign grades based on objective evidence.
4. Ascertain that pupils comprehend the examination guidelines.
5. Connect the questions to what is being taught in class.
6. Never tolerate student cheating.
7. Ascertain that the exam grades are appropriate for the intended purpose.
8. Whenever possible, never alter the grade assigned.
9. Make every effort to share the exam results as soon as possible.

Furthermore, Masters (1987) and Messick (1984) emphasized the need to embrace students' evolving and partially correct ideas rather than label them as 'wrong.' According to them, it is critical to focus on each student's individual development rather than compare them to one another.

The question of how to evaluate students fairly has long been an intriguing one, both theoretically and practically, particularly for psychologists (Meyer, 1908). A student's grade is a summary of his/her accomplishments. Notifying students of this grade level can also be handled separately. Because while a grade may motivate students to learn or boost their self-confidence, it may also have the opposite impact, diminishing the student's desire to learn or disrupting their psychology. In addition to the variables that teachers must consider when assigning grades, Gardner *et al.* (1997) proposed that the following aspects should be emphasized when notifying students of their grades.

1. If students have concerns or reservations regarding their grades, explain the reasons,
2. Inform students about the grading criteria.
3. Notify the students' parents through letter, either individually or as a group.
4. Avoid being abrasive in your provisions.
5. Maintain a balance of oral, written, and multiple-choice examinations.
6. Keep in mind that each grade should provide an opportunity for students to remedy their weaknesses.

Another thing to keep in mind is that it is important to tell the student not just her grade but also how she can improve her performance (Masters, 1987; Messick 1984).

Numerous studies have been carried out in the literature on the extent to which teachers follow the important points stated above by Gardner *et al.* (1997). These studies show that most teachers do not know how to appropriately evaluate or grade students (Brewer & deMarrais, 2015). This is especially true for teachers working in regions where the need for teachers is high and socio-economic income is low (Redding & Smith, 2016). Due to teachers' lack of training on this issue, teachers determine students' grades based on variables other than evidence of student performance (Guskey, 2015). his combination of student accomplishment and process variables can lead to score pollution that does not correctly reflect students' grades, as well as impede academic mastery and access to accurate information about academic achievement by students, families, and other education system stakeholders (Green, Johnson, Kim, & Pope, 2006).

Although teachers agree that grades should not be assigned for non-academic subjects (Frisbie, Diamond, & Ory 1979), Guskey & Bailey (2001) and Andersson (1998) argue that teachers generally avoid assigning grades solely on the basis of achievement and that when they do, they consider other factors in addition to success. Brookhart *et al.* (2016) suggest similarly that grades are typically a composite of numerous factors that teachers value (e.g., effort, ability, study habits, engagement, and participation), and that these factors vary significantly depending on what teachers believe. McMillan, Myran, & Workman (2002) used the term "chaotic grading" to refer to this type of grading. Guskey and Link (2019) propose that integrating both achievement scores and process evaluation results in end-of-term grades may result into score pollution that fails to acknowledge the information on academic competence.

A grade may represent academic achievement alone (Bailey & McTighe, 1996) or some combination of academic achievement and one or more other factors (e.g., effort, attendance, classroom participation, and/or behavior). It is much easier to interpret a grade that represents only academic achievement. If grades are based on a combination of scores from key exams, essays, quizzes, projects, and reports, as well as evidence from homework, punctuality in delivering assignments, classroom participation, study habits, and effort, the result will be a mess (Guskey, 2011).

In school, teachers decide which students pass or fail based on their grades, which are mostly determined by the written exams that students take (Koç, 1981). However, most teachers do not possess the necessary skills to assure the validity of the measurement tools they employ (Öztürk, 1988). Teachers, in particular, struggle with developing questions that are appropriate for their students' levels (Acar Güvendir & Özer Özkan, 2016). Furthermore, teachers' grades are inconsistent, regardless of whether they utilize answer keys or not when grading written examinations (Kan, 2005). Additionally, teachers might incorporate success or external factors into their measurement and evaluation processes (Semerci, 1993; Topal, 2020). According to the Ministry of National Education's [MoNE] (2005) report, the "monitoring and evaluating learning and development" competence area has the lowest average on the self-assessment scale used to evaluate teachers' self-evaluation of the qualifications included in the draft "teaching profession general competences." In other words, teachers frequently feel insecure about their measurement and evaluation abilities. Similarly, studies show that teachers in several sectors of elementary, secondary, and high school education lack measurement and evaluation skills (Adıyaman, 2005; Çakan, 2004; Erdal, 2007; Erdemir, 2007).

As a result, fair, transparent, and effective grading procedures and methods are required to aid all students in reaching higher academic standards. However, it is apparent that teachers are incompetent at all stages of the grading process, from the development of the measurement tool through its implementation. When teachers grade students, they also take into account a variety of variables other than the grade. In this context, it is important to discover teachers' perspectives on grading processes. To investigate teachers' perceptions, Liu (2004) and Liu, O'Connell, and McCoach (2006) developed the "The Teachers' Perceptions of Grading Practices Scale" in English and Chinese. The purpose of this study is to construct a Turkish version of this scale whose validity and reliability have been established in different cultures.

It is also significant to look for evidence of measurement invariance, which is required for group comparisons based on the modified "Teachers' Perceptions of Grading Practices Scale". Since the validity and reliability are based on the measurements obtained from the measurement tool, the test and item statistics calculated to obtain information about the level of validity and reliability only reflect the characteristics of the individuals in the group (Crocker & Algina, 1986). As a result, the evidence regarding the validity and reliability of measures taken in different groups may vary. The psychometric properties of the measurements acquired may be a result of the individuals' unique features or they may be a result of the measurement tool. Thus, measurement invariance investigations disclose the circumstances under which observed variables are valid and reliable between groups (Vandenberg & Lance, 2000). The other goal of this study is to find out if the measuring tool can be used to compare different groups. To do this, a measurement invariance study will be done on the "Teacher Perceptions of Grading Practices" across gender groups.

As a result, this scale, which was adapted and whose measurement invariance was investigated between groups, might be utilized as a tool in future intercultural comparisons of teachers' grading practices. This scale may also be used to make different decisions concerning the grading processes of teachers working in Türkiye. As a consequence, it was deemed necessary to investigate the scale's validity and reliability, as well as its measurement invariance.

2. METHOD

In this section, the scale adaptation steps are explained in detail. The following steps were followed for scale adaptation (Deniz, 2007; Hambleton, 1996; Hambleton, Meranda, & Spielberger, 2005; Hambleton & Patsula, 1999).

1. Permission has been received for the adaption study.
2. Field specialists were consulted on the scale's adaptability.

3. Measurement specialists were consulted on the scale's adaptability.

4. To ensure language comparability, translators fluent in both cultures were chosen. Two translators performed the translation, and the translated version of the scale was reviewed and approved by three translators.

5. A back-translation was done.

6. It was determined if the two variants of the scale were semantically, experientially, conceptually, and idiomatically equivalent.

7. A pilot application was conducted.

8. Confirmatory factor analysis (CFA) was used to examine the factor structure of the original scale.

9. Various approaches for determining reliability were utilized.

After the adaption, measurement invariance was used to determine how teachers' responses to the scale varied by gender. In the measurement invariance process, configural, metric, scalar, and strict invariance stages were followed.

2.1. Scale Adaptation Process

The measurement tool adapted in this study is the Teachers' Perceptions of Grading Practices Scale, which was developed in English and Chinese by Liu (2004) and Liu *et al.* (2006) to determine teachers' perceptions of the practices they use in the grading process. this instrument measuring teachers' perceptions of grading practices has six factors (Importance, Usefulness, Student Effort, Student Ability, Teachers' Grading Habits, Perceived Self-efficacy of Grading Process). It is 5-point Likert rating scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree). The fit indices for the hypothesized six-factor model with 40 items were as follows: Chi-square (χ^2) = 1562.67, degree of freedom (*df*)= 687, *p*< 0.001. Confirmatory Fit Index (CFI) = .80, Root Mean Square Error of Approximation (RMSEA) = 0.067 (90% Confidence interval of 0.062 to 0.071), and χ^2 /*df* = 2.277. The reliability coefficient of the whole scale is 0.73.

Permission to adapt the scale was received by e-mail from the researchers who developed it. In the second stage, translation and back-translation processes were carried out by researchers as well as three lecturers working in English language educators who are proficient in both cultures. The researchers examined whether the two scale forms were semantically, experientially, conceptually, and idiomatically equivalent. Due to the presence of a group that spoke the original test language, the original and adapted scales were administered twice, one month apart, to a group of English teachers, and the consistency of the two applications was investigate product-moment correlation coefficient was calculated to determine the relationship between the two scales' scores (two-term, normally distributed scores). The correlation coefficient obtained was 0.86, indicating a positive, high, and significant relationship (p<0.05) between the two applications. CFA was performed to examine whether the factor structure of the original scale was the same in its Turkish version. Cronbach's α and McDonald's ω reliability coefficients were estimated during the scale's reliability research.

To begin with, the CFA analysis's assumptions were tested in order to verify the scale's factor structure. First of all, it was checked whether there was missing data in the data and it was observed that there was no missing data. One of its underlying assumptions is that there are no versatile extreme values. This assumption was made using Mahalanobis distances. A total of 549 teachers responded to the scale. However, 52 outliers determined by Mahalanobis distances were removed. The second assumption is that the sample size for factor analysis must be adequate. The Kayser-Mayer-Olkin (KMO) test was used to analyze this, and because the value obtained was 0.918, the sample size was large enough for factor analysis (Leech, Barrett, & Morgan, 2005). Another assumption is normality. Since CFA is a multivariate analysis, it requires a multivariate normality assumption. This was done by using Henze-Zirkler's test, which showed that the data did not meet the assumption of multivariate normality. (hz= 1.082; p<0.05). When the variables observed in the CFA did not show normal distribution, the WLS method was preferred since the Weighted Least Squares Method (AGL-WLS Weighted Least Square Estimation) was used as the parameter estimation method (Bollen, 1989; Schermelleh-Engel, Moosbrugger, & Müller, 2003).

CFA, which was conducted to reveal how the original factor structure of the scale was in its Turkish form, was carried out using LISREL software (v. 8.71; Jöreskog & Sörbom, 2004). Cronbach's α and McDonald's ω coefficients were calculated using Jamovi software (v. 1.8; The Jamovi Project).

Multiple group confirmatory factor analysis (MGCFA) was used for the measurement invariance of the scale according to gender groups (Jöreskog & Sörbom, 1993). For measurement invariance, configural, metric, scalar, and strict invariance models were established and the difference between the CFI and RMSEA values obtained in each model from the CFI and RMSEA values obtained with the configural invariance model was taken. ΔCFI and ΔRMSEA (Chen, 2007; Cheung & Rensvold, 2002) values were used as decision criteria in the analysis of stepwise models for measurement invariance in gender groups. According to Chen (2007), in samples larger than 300, -0.010≤ ΔCFI and ΔRMSEA≤ 0.015 values are the cut-off points for the invariance decision. These values were utilized as the cut-off point for this study to ensure that measurement invariance was attained or not. For measurement invariance, "Lavaan" (http://cran.r-project.org/web/packages/lavaan/index.html) and "semTools" (http://cran.r-project.org/web/packages/semTools/index.html) available in R software packages are used. The package (http://cran.r-project.org/web/packages/MVN/index.html) was used for multivariate normality checking.

2.2. Study Group

There are 497 teachers in the study group. In terms of gender distribution, females made up 59.8% of the study group, while males made up 40.2%. In terms of school type, 28.8% work at elementary schools, 47.9% at secondary schools, and 23.3% attend work at high schools. Associate degree instructors make up 1.6% of the research group, undergraduate teachers make up 78.1%, and graduate teachers make up 20.3%. When their distribution is examined in terms of professional seniority, 6.4% have less than one year, 9.7% have 1-3 years, 9.5% have 4-5 years, 26.8% have 6-10 years, 17.1% have 11-15 years, 14.1% have 16-20 years, and 16.5% have more than 20 years of service. The data were obtained in the spring semester of the 2020-2021 academic year.

3. FINDINGS

Descriptive statistics and reliability coefficients for the six sub-factors of the scale of teacher perceptions regarding grading practices are presented in Table 1.

Table 1. *Descriptive statistics and reliability coefficients of the scale.*

	Mean	SD	Cronbach's α	McDonald's ω
Importance	3.51	0.85	0.93	0.94
Usefulness	3.55	0.69	0.91	0.92
Student Effort	3.91	0.54	0.77	0.78
Student Ability	4.04	0.58	0.92	0.93
Teachers' Grading Habits	3.72	0.53	0.67	0.70
Perceived Self-efficacy of Grading Process	2.81	0.67	0.68	0.70

The reliability values obtained for the scale's six sub-factors were found to be high. The fact that the obtained values exceed 0.70 indicates a high degree of reliability. CFA was performed to obtain evidence of the scale's factor structure. The initial CFA involved 427 participants. The scale's 40th item was insignificant. While filling out the scale, the researchers inserted the item into the first item of the relevant factor and retested 70 respondents, taking into account the high likelihood of quitting, becoming fatigued, or responding without reading the final item. As a result of CFA performed on 497 participants, item 40 (z=-1.5, p=0.13) was excluded because it was not significant (p>0.05). The 39th item on the scale has a standardized estimation value of less than 0.30, indicating that it contributes very little to the factor. As a result, this item was removed from the scale due to its low factor load. Permission was obtained from the researcher who developed the scale at the stage of removing these items. Table 2 shows the standardized regression values and the unstandardized regression coefficients for the other 38 items.

Table 2. *Factor loadings of the scale of teachers' perceptions of grading practices.*

Factor	Indicator	Estimate	SE	95% Confidence Interval		Z	p	Stand. Estimate
				Lower	Upper			
Importance	I1	0.86	0.04	0.78	0.93	22.01	< 0.00	0.82
	I2	0.79	0.04	0.72	0.86	22.30	< 0.00	0.83
	I3	0.72	0.04	0.65	0.79	20.22	< 0.00	0.77
	I4	0.88	0.04	0.81	0.95	25.58	< 0.00	0.90
	I5	0.89	0.04	0.83	0.95	26.02	< 0.00	0.91
	I6	0.82	0.04	0.75	0.89	21.77	< 0.00	0.81
Usefulness	I7	0.63	0.04	0.56	0.70	17.05	< 0.00	0.69
	I8	0.72	0.04	0.65	0.80	19.78	< 0.00	0.76
	I9	0.60	0.03	0.53	0.66	18.43	< 0.00	0.73
	I10	0.70	0.04	0.62	0.79	16.24	< 0.00	0.66
	I11	0.75	0.04	0.68	0.81	21.47	< 0.00	0.81
	I12	0.74	0.04	0.67	0.81	20.52	< 0.00	0.78
	I13	0.75	0.04	0.68	0.82	21.22	< 0.00	0.80
	I14	0.70	0.04	0.62	0.78	17.14	< 0.00	0.69
	I15	0.61	0.03	0.55	0.67	20.93	< 0.00	0.79
	I16	0.42	0.04	0.33	0.50	9.93	< 0.00	0.44
Student Effort	I17	0.54	0.03	0.48	0.59	18.56	< 0.00	0.76
	I18	0.64	0.03	0.57	0.70	18.59	< 0.00	0.77
	I19	0.46	0.03	0.40	0.53	14.09	< 0.00	0.62
	I20	0.32	0.04	0.23	0.40	7.12	< 0.00	0.34
	I21	0.43	0.03	0.36	0.49	12.40	< 0.00	0.56
	I22	0.43	0.03	0.36	0.50	12.57	< 0.00	0.57
Student Ability	I23	0.50	0.03	0.45	0.55	19.64	< 0.00	0.76
	I24	0.56	0.02	0.51	0.60	24.56	< 0.00	0.88
	I25	0.62	0.02	0.58	0.67	27.80	< 0.00	0.94
	I26	0.62	0.02	0.58	0.66	28.51	< 0.00	0.95
	I27	0.60	0.02	0.55	0.64	25.84	< 0.00	0.90
	I28	0.45	0.04	0.38	0.53	12.21	< 0.00	0.52

Table 2. *Continues.*

Teachers' Grading Habits	I29	0.30	0.05	0.20	0.40	5.77	< 0.00	0.29
	I30	0.37	0.05	0.28	0.47	7.54	< 0.00	0.38
	I31	0.45	0.03	0.39	0.52	13.16	< 0.00	0.62
	I32	0.45	0.03	0.39	0.52	13.93	< 0.00	0.65
	I33	0.51	0.04	0.43	0.59	12.46	< 0.00	0.58
	I34	0.43	0.04	0.35	0.50	11.06	< 0.00	0.53
Perceived Self-efficacy of Grading Process	I35	0.55	0.05	0.45	0.65	10.74	< 0.00	0.54
	I36	0.33	0.04	0.16	0.31	5.87	< 0.00	0.31
	I37	0.69	0.05	0.60	0.79	14.33	< 0.00	0.74
	I38	0.78	0.05	0.67	0.88	14.68	< 0.00	0.77

When Table 2 is seen, the standardized estimation values for factor loadings for all items vary between 0.30 and 0.89. According to Tabachnick and Fidell (2019), factor loads should be at a minimum of 0.32. Büyüköztürk (2002) categorized a load value of 0.60 or greater as high, and 0.30-0.59 as medium. As a result, all items pertaining to the factors are significant (*p*<0.01), and factor loads are greater than 0.30. The model's fit index values (χ^2= 1868.10, *df*=650, χ^2/df=2.87, RMSEA =0.06, CFI=0.97, NNFI=0.96) were significant at the 0.05 level of significance (*p*<0.05). When model fit indices are evaluated, χ^2/df value (2.69) is deemed acceptable by Schermelleh-Engel *et al.* (2003) and corresponds to a moderate fit, as it is less than 5, as defined by Sümer (2000). The RMSEA value shows that the fit is acceptable. NNFI and CFI values indicate a good fit of the model. Appendix 2 shows the path diagram for the six-factor model derived using DFA.

When Appendix 2 is examined, it is noticeable that the scale of 38 items with six variables was confirmed. The gender invariance of the six-factor construct was tested using multi-group CFA analyses. Multi-group confirmatory factor analysis was performed to demonstrate that the psychometric features of the scale did not remain constant across the groups to which it would be applied (Thompson, 2004). Table 3 summarizes the results of the tested invariance stages.

Table 3. *Results of measurement invariance obtained by gender.*

Stages	$\chi 2$	*d*	CFI	GFI	RMSEA	ΔRMSEA	ΔCFI
Configural Invariance	2544.58	1300	0.89	0.97	0.06	-	-
Metric Invariance	2568.44	1332	0.89	0.97	0.06	-0.00	0.00
Scalar Invariance	2630.89	1364	0.89	0.97	0.06	-0.00	-0.00
Strict Invariance	2707.36	1400	0.88	0.96	0.06	-0.00	-0.01

In order to determine the measurement invariance between the groups at the stages in Table 3, the difference values of the fit coefficients (ΔCFI and ΔRMSEA) were given by comparing the more limited models with the configural model. In accordance with Table 3, the fit indices as a result of multi-group CFA for configural invariance show that this stage is achieved. In other words, female and male teachers use the same conceptual perspectives in responding to scale items. The fit indices as a result of multi-group CFA for metric variance and the ΔCFI and ΔRMSEA values obtained as a result of the CFI and RMSEA difference tests were interpreted. The fit indices obtained show that the model fits well with the data. To test the metric invariance, the difference between the CFI and RMSEA values obtained in the configural invariance and metric invariance stages was examined, and it was seen that ΔCFI and ΔRMSEA for metric invariance were within acceptable limits (ΔCFI $\leq$0.01; ΔRMSEA $\leq$ 0.015). This finding shows that the factor loadings of the variables included in the model do not vary depending on a person's gender.

In the scalar invariance stage, fit indices are within acceptable limits. Scalar invariance was tested by comparing the CFI and RMSEA values obtained from configural invariance to the CFI and RMSEA values obtained from scalar invariance. When the findings were analyzed, it was discovered that the measurement model for the scale of teacher perceptions on grading processes fulfilled the scalar invariance requirement (ΔCFI $\leq$ 0.01; ΔRMSEA $\leq$ 0.015). After the scalar invariance stage, the strict invariance stage was tested.

Strict invariance fit indices are within accepted limits. The difference between the CFI and RMSEA values obtained during the configural and strict invariance phases indicated that the grading practices measurement model in gender subgroups fulfilled the strict invariance stage (ΔCFI $\leq$ 0.01; ΔRMSEA $\leq$ 0.015).

4. DISCUSSION and CONCLUSION

The purpose of this research was to analyze the validity and reliability of the Turkish version of the Teachers' Perceptions of Grading Practices Scale. CFA was performed to confirm the factor structure of the original scale in its Turkish form. Cronbach's α and McDonald's ω coefficients, which measure internal consistency, were used to check for reliability. A significant t value could not be found for the scale's 40th item (Students' engagement in the course outside of the test, social events, and other activities complicates my grading procedure.). While the scale provided satisfactory fit values, it was established that the t value for the 40th item was not significant and that the error variance for this item was also rather high. As a result, item 40 was eliminated from the scale. This item is meant to assess if instructors' non-grading status hinders their work when it comes to grading students. The reason why the item does not work in the Turkish form may be due to the attitude difference between the two cultures. Some of the teachers who answered this scale think that it is normal for them to consider their students' extracurricular situations while grading. Interviews were held with the teachers regarding this item. Teachers stated that while grading, variables other than grades (such as listening to the lecture, being respectful, doing their homework regularly) also affect their grading status. They stated that they reflect these non-academic variables on their exam scores in order to motivate students, and this is the right thing to do. This article may not function in Turkish owing to the cultural differences between the two cultures. According to several teachers who responded to this scale, it is natural for them to include their students' extracurricular activities while grading. Teachers were interviewed on this subject. Teachers indicated that during grading, they take into account aspects other than grades (such as listening to the lecture, being courteous, and doing their assignments on a consistent basis). They argued that they include these non-academic characteristics in their exam results in order to stimulate pupils, which is the correct thing to do. For instance, an English teacher at a fine arts high school remarked that she considers her students' talent while grading, and the administration even encourages them to do so. This should not be suggested in foreign literature, as it would influence the validity of the scores (Guskey, 2011; Guskey & Link, 2019). While Koç (1981) asserted that teachers largely determine their students' pass-fail status based on the results of written exams, Semerci (1993), Topal (2020), Guskey & Bailey (2001), and Andersson (1998) argue that teachers can incorporate factors outside the classroom into the measurement and evaluation process. Frisbie, Diamond, and Ory (1979) argue that grades should not be assigned for non-academic areas. Otherwise, grading will be chaotic (McMillan *et al.*, 2002) and will result in score pollution (Green *et al.*, 2006).

In addition, since the factor load of item 39 was 0.15 (<0.30), this item was also removed from the scale. When the English (it is difficult to measure student effort) and translated Turkish equivalents of this item are examined, it is clear that the statement is written as a factual statement rather than a perceived self-efficacy statement. Therefore, although the item is significant, it is thought that the factor load is therefore low. Since these last two items on the original scale did not work in the Turkish form, Liu, who developed the scale, was contacted and permission was requested to remove it. After the positive response from the scale developer, these two items were removed from the scale, and confirmatory factor analysis was found appropriate to be done again. The results obtained in the repeated analysis show that the 38-item scale is consistent with the six-factor original structure and is compatible with the data. Taking into account all of the coherence values, it is possible to conclude that the theoretical framework explains the relationships between the data acquired from the Turkish form of the scale. The internal consistency coefficients of the entire scale and its sub-factors were examined to determine reliability of the data obtained from the scale. Cronbach's α and McDonald's ω internal consistency coefficients are high on the basis of the entire scale and factors. As a result, the data acquired from the scale can be said to be consistent. As a result, the means obtained from these two groups formed by gender using this scale can be compared.

The measurement invariance of the adapted scale in different groups in terms of gender was determined by examining the ΔCFI and ΔRMSEA values obtained for the models. It was concluded that the grading practices measurement model met the condition of complete invariance because it included all of the configural, metric, scalar, and strict invariance stages in gender groups. Measurement invariance of the scale across cultures was examined by Liu (2008). In Liu's study, the factor loadings of the 39th and 40th items out of 40 items in the scale were not found to be similar in the two compared samples (China and the United States). This finding shows that the answers given to items 39 and 40 differ according to cultures. In this study, these items were removed from the scale as a result of CFA, and the measurement invariance according to gender was made over 38 items and the 38-item scale provided measurement invariance.

The study was carried out with 497 teachers. The research enlisted the help of 497 instructors. The original scale's factor structure was meant to be validated in the study, and measurement invariance in different groups was evaluated. Along with these procedures, convergent and divergent validity investigations can be carried out. Furthermore, the outcomes of studies using the adapted scale are expected to increase the evidence that the scale is both valid and reliable.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Trakya University, 25.05.2022 - 05/13.

Authorship Contribution Statement

Yesim Ozer Ozkan: Methodology, Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Meltem Acar Guvendir:** Introduction and discussion, Writing -original draft, Methodology, Investigation, Resources, and Formal Analysis. **Emre Guvendir**: Introduction and discussion, Writing, and Proofreading.

Orcid

Yesim Ozer Ozkan https://orcid.org/0000-0002-7712-658X
Meltem Acar Guvendir https://orcid.org/0000-0002-3847-0724
Emre Guvendir https://orcid.org/0000-0003-1226-9878

REFERENCES

Acar Güvendir, M., & Özer Özkan, Y. (2016). *Practicality of measurement and evaluation course in education.* Paper presented at V. Congress of Measurement and Evaluation in Education and Psychology, 1-3 September 2016, Antalya.

Adıyaman, Y. (2005). *İlköğretim 4., 6. ve 8. sınıflarında Türkçe dersine giren öğretmenlerin ölçme değerlendirme düzeyleri [The measurement and evaluation levels of teachers teach Turkish course in 4th, 6th and 8th classes in primary school]* [Unpublished Master Thesis, Kocatepe University].

Andersson, A. (1998). The dimensionality of the leaving certificate. *Scandinavian Journal of Educational Research, 42*(1), 25-40. https://doi.org/10.1080/0031383980420102

Anderson, L.W. (2018). A Critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives, 26*(49), 1-31. http://dx.doi.org/10.14507/epaa.26.3814

Bailey, J.M., & McTighe, J. (1996). Reporting achievement at the secondary level: What and how. In T.R. Guskey (Ed.), *Communicating student learning: 1996 Yearbook of the ASCD* (pp. 119–140). ASCD.

Bollen, K.A. (1989). *Structural equations with latent variables,* Wiley.

Brewer, T.J., & deMarrais, K. (2015). *Teach for America counter-narratives: Alumni speak up and speak out.* Peter Lang Incorporated, International Academic Publishers. https://doi.org/10.3726/978-1-4539-1556-1

Biggs, J. (2001) Assessment of student learning: Where did we go wrong? *Assessment Update, 13*(6), 6-11.

Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T., & Welsh, M.J. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848. https://doi.org/10.3102/0034654316672069

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı. [Factor analysis: basic concepts and using to development scale]. *Educational Administration in Theory & Practice, 32*(32), 470-483. https://dergipark.org.tr/en/pub/kuey/issue/10365/126871

Campbell, A.L. (1921). Keeping the score. *School Review, 29*(7), 510-519.

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Çakan, M. (2004). Öğretmenlerin ölçme-değerlendirme uygulamaları ve yeterlik düzeyleri: ilk ve ortaöğretim. [Comparison of elementary and secondary school teachers in terms of their assessment practices and perceptions toward their qualification levels]. *Ankara University, Journal of Faculty of Educational Sciences, 37*(2), 99-114. https://doi.org/10.1501/Egifak_0000000101

Deniz, Z. (2007). Psikolojik ölçme aracı uyarlama. [The Adaptation of psychological scales]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 40*(1), 1-16.

Eraut, M. (2004) A wider perspective on assessment, *Medical Education, 38(1),* 803-804. https://doi.org/10.1111/j.1365-2929.2004.01930.x

Erdal, H. (2007). *2005 ilköğretim matematik programı ölçme değerlendirme kısmının incelenmesi (Afyonkarahisar ili örneği). [The investigation of measurement & evaluation parts in the new elementary school mathematics curriculum (case of Afyonkarahisar)]* [Unpublished Master Thesis, Kocatepe University].

Erdemir, Z.A. (2007). *İlköğretim ikinci kademe öğretmenlerinin ölçme değerlendirme tekniklerini etkin kullanabilme yeterliklerinin araştırılması (Kahramanmaraş örneği).*

[Searching for the secondary education teachers' competence of being able to use the techniques of measurement and evaluation (example of Kahramanmaraş)] [Unpublished master thesis, Kahramanmaraş Sütçü İmam University].

Frisbie, D., Diamond, N.A., & Ory, J.C. (1979). *Assigning course grades*, Urbana, IL: University of Illinois Office of Instructional Resources. (ERIC Document Reproduction Service No. ED285496)

Gardner, W., Demirtaş, A., & Doğanay, A. (1997). Sosyal bilimler öğretimi. [Social sciences teaching] YÖK-Dünya Bankası. MEGEP.

Green, S.K., Johnson, R.L., Kim, D., & Pope, N.K. (2006). Ethics in classroom assessment practices: Issues and attitudes. *Teacher and Teacher Education, 23*(7), 999-1011. https://doi.org/10.1016/j.tate.2006.04.042

Guskey, T.R. (2011). Five obstacles to grading reform. *Educational Leadership, 69*(3), 16-21. https://uknowledge.uky.edu/edp_facpub/6

Guskey, T.R. (2015). *On your mark: Challenging the conventions of grading and reporting.* Bloomington, IN: Solution Tree Press.

Guskey, T.R., & Bailey, J.M. (2001). *Developing grading and reporting systems for student learning*, Corwin Press.

Guskey, T.R., & Link, L.J. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), 303-320. https://doi.org/10.1080/0969594X.2018.1555515

Hambleton, R.K. (1996). *Guidelines for adapting educational and psychological test*. National Center for Education Statistics (ED).

Hambleton, R.K. & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*(1), 1-30.

Hambleton, R.K., Merenda, P., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Lawrence S. Erlbaum Publishers.

Harlen, W. (2005) Teachers' summative practices and assessment for learning- Tensions and synergies, *The Curriculum Journal, 16*(2), 207-223. https://doi.org/10.1080/09585170500136093

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Jöreskog, K., & Sörbom, D. (2004). *LISREL [Computer Software]*. Lincolnwood, IL: Scientific Software, Inc. https://doi.org/10.1002/0471667196.ess1481

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (aynı) puanlayıcı güvenirliğine etkisi [The effect of using grading scale and answer key to grader's reliability]. *Eurasian Journal of Educational Research, 20(1),* 166-177.

Koç, N. (1981). *Liselerde öğrencilerin akademik başarılarının değerlendirilmesi uygulamalarının etkinliğine ilişkin bir araştırma [A research on the effectiveness of the applications of evaluating the academic achievement of students in high schools*] [Unpublished Master Thesis, Ankara University].

Küçükahmet, L. (2005). *Öğretimde planlama ve değerlendirme [Planning and evaluation in instruction].* Nobel Yayınları.

Leech, N.L., Barrett, K.C., & Morgan, G.A. (2005) *SPSS for intermediate statistics,use and interpretation* (2nd Edition). Lawrence Erlbaum.

Liu, X. (2004). *The initial validation of teacher's perception of grading practices*. Paper presented at the 2004 Northeastern Educational Research Association annual meeting, Measuring Teachers' Perceptions 14.

Liu, X. (2008). *Assessing measurement ınvariance of the teachers' perceptions of grading practices scale across cultures*. NERA Conference Proceedings 2008. 3. https://opencommons.uconn.edu/nera_2008/3.

Liu, X., O'Connell, A.A., & McCoach, D.B. (2006). *The initial validation of teachers' perceptions of grading practices*. Paper presented at the 2006 Annual Meeting of American Educational Research Association (AERA).

Masters, G. (1987). *New views of student learning: Implications for educational measurement.* Research working paper 87.11. University of Melbourne: Centre for the Study of Higher Education.

McMillan, J.H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203–213. https://doi.org/10.1080/00220670209596593

Ministry of National Education's [MoNE] (2005). *EARGED ilköğretim 1.-5. sınıf pilot uygulama sonuçlarının değerlendirilmesi [EARGED primary education 1.-5. Evaluation of class pilot application results. MoNE Publications.*

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215-237. https://doi.org/10.1111/j.1745-3984.1984.tb01030.x

Meyer, M. (1908). *The grading of students. Science*, 28(712), 243-250. https://doi.org/10.1126/science.28.712.243

Öztürk, B. (1988). *Lise sosyal bilimler dersleri öğretmenlerinin başarı testi hazırlamadaki yeterliliklerine ilişkin bir araştırma [A research on the competencies of high school social science teachers in preparing achievement tests]* [Unpublished Master Thesis, Gazi University].

Redding, C., & Smith, T.M. (2016). Easy in, easy out: Are alternatively certified teachers turning over at increased rates? *American Educational Research Journal, 53*(4), 1086-1125. https://doi.org/10.3102/0002831216653206

Salend, S.J., & Duhaney, L.M.G. (2002). Grading students in inclusive settings. *Teaching Exceptional Children, 34*(3), 8-15. https://doi.org/10.1177/004005990203400301

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74. http://www.mpr-online.de

Semerci, Ç. (1993). *Fırat Üniversitesinde öğrenci başarısının ölçülmesinde kullanılan yöntemler ile ölçme ve değerlendirmeye ilişkin görüşler [Opinions on the methods used in measuring student achievement at Fırat University and on measurement and evaluation]* [Unpublished Master Thesis, Fırat University].

Serban, A.M. (2004) Assesment of student learning outcomes at the institutional level. *New Directions For Comminity Colleges, 2004*(126), 17-27. https://doi.org/10.1002/cc.151

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar [Structural equation modeling: Basic concepts and applications]. *Türk Psikoloji Yazıları, 3*(6), 49–74.

Tabachnick, B.G. & Fidell, L.S. (2019). *Using multivariate statistics* (7th edition). Pearson.

The jamovi project (2021). *Jamovi*. (Version 1.8) [Computer Software]. Retrieved from https://www.jamovi.org

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association. https://doi.org/10.1037/10694-000

Topal, T. (2020). Öğretmen adaylarının bakış açısından sınıf öğretmenlerinin öğretim sürecinde gösterdikleri dönüt ve düzeltme davranışları [Feedback and correction behavior of the classroom teachers during the teaching process from the perspective of the teacher candidates]. *OPUS International Journal of Society Researches, Eğitim ve Toplum Özel Sayısı, 16*, 6150-6166. https://doi.org/10.26466/opus.825157

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

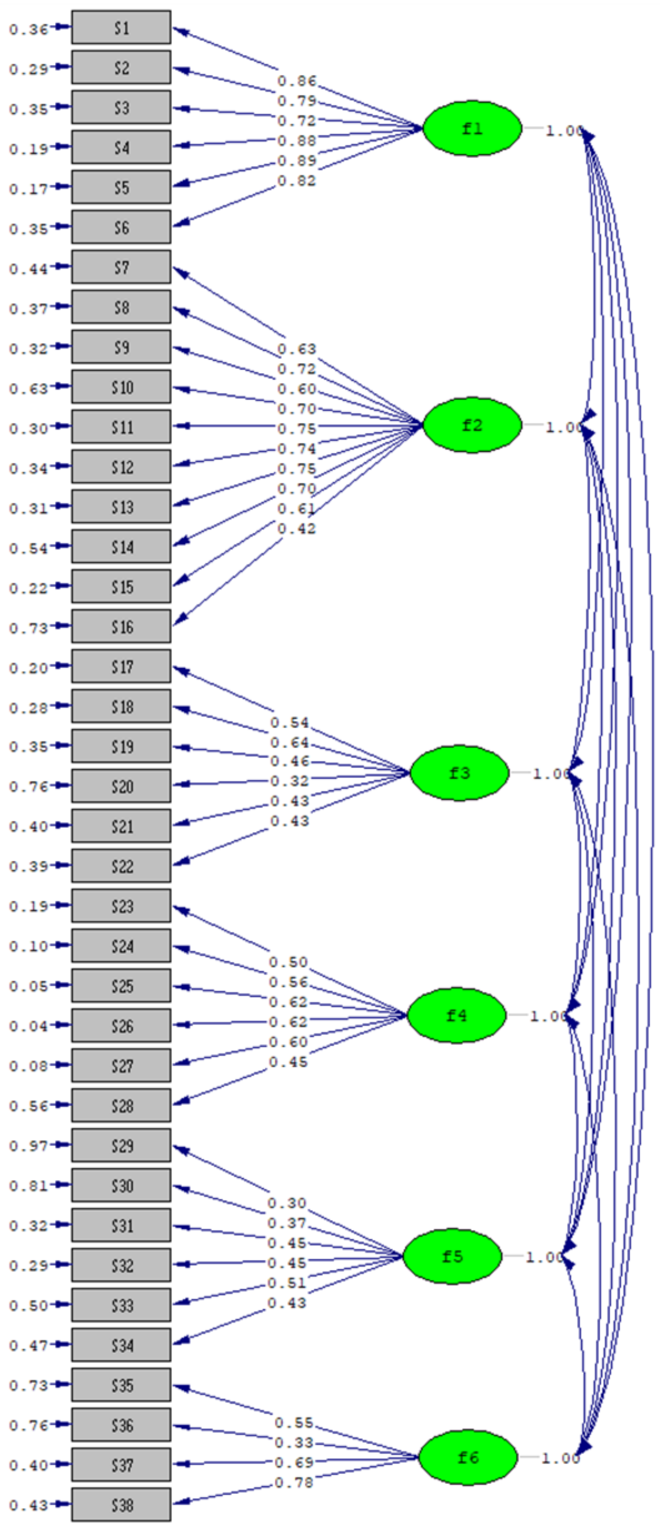
APPENDIX

Appendix 1. *Teachers' perceptions of grading practices scale* (Turkish version).

Faktör 1. Önem	**Kesinlikle Katılmıyorum**	**Katılmıyorum**	**Nötr**	**Katılıyorum**	**Kesinlikle Katılıyorum**
1. Not verme, öğrencilerin gelişimlerini değerlendirmek için önemli bir ölçüttür.					
2.Not verme, sınıf içi ölçme ve değerlendirmelerde önemli bir role sahiptir.					
3. Not verme, öğrencilerin akademik başarıları üzerinde olumlu bir etkiye sahiptir.					
4.Not verme uygulamaları, sınıf içi öğrenmelerin önemli ölçülerini oluşturur.					
5.Not verme uygulamaları, öğrenci başarısının önemli ölçümleridir.					
6. Not verme, öğrencilerin öğrenmeleri üzerinde güçlü bir etkiye sahiptir.					
Faktör 2. Yarar					
7. Not verme, öğrencileri ortalamanın üstünde, ortalama düzeyde ve ortalamanın altında olarak sınıflandırmama yardımcı olur.					
8. Not verme, öğretim yöntemimi geliştirmeme yardımcı olur.					
9.Verilen notlar öğrencileri iyi çalışmalar yapmaya teşvik edebilir.					
10. Not verme, hangi konuları öğreteceğime karar vermeme yardımcı olur.					
11. Not verme, öğrencilerin bir dersin içeriğindeki zayıflıklarını belirlemeye yardımcı olan iyi bir yöntemdir.					
12.Not verme, öğrencileri gelişimleri hakkında bilgilendirebilir.					
13.Not verme, öğrenci başarısı hakkında bilgi verir.					
14.Not verme, benim etkili bir öğretim uyguladığımın bir göstergesidir.					
15.Not verme, öğrencilerime geri bildirim sağlar.					
16.Yüksek notlar, öğrencileri öğrenmeye motive edebilir.					
Faktör 3. Öğrenci Çabası					
17.Not verirken öğrencinin çabasını göz önünde bulundururum.					
18.Daha fazla çaba gösteren öğrencilere daha yüksek karne notları veriyorum.					
19.Başarısız bir öğrenciyi çaba göstermesi halinde geçiririm.					
20.Verdiğim notlar, öğrencilerin verilen ödevleri tamamlayıp tamamlamadıklarına dayanır.					
21.Verdiğim notlar, öğrencilerin sınıfta derse katılma düzeylerine dayanır.					
22. Verdiğim notlar, öğrencinin gelişim düzeyine dayanır.					
Faktör 4: Öğrenci Yeteneği					
23.Not verirken öğrencilerin yetenek düzeylerini göz önünde bulundururum.					
24.Not verirken, öğrencilerin problem çözme yeteneğini göz önünde bulundururum.					
25.Not verirken, öğrencilerin eleştirel düşünme yeteneğini göz önünde bulundururum.					

26.Not verirken, öğrencilerin bağımsız düşünme becerilerini göz önünde bulundururum.					
27.Not verirken, öğrencilerin işbirliğine dayalı öğrenme yeteneğini göz önünde bulundururum.					
28.Not verirken, öğrencilerin yazma becerilerini göz önünde bulundururum.					
Faktör 5: Öğretmenlerin not verme alışkanlıkları					
29. Not verirken, imkanım olsaydı, rakamlardan ziyade harfleri (örn., A, B, C) kullanma eğiliminde olurdum.					
30. Bir öğrenci sınavda başarısız olursa, ona sınava girmek için ikinci bir şans daha sunarım.					
31.Öğrencilere sıklıkla ek puan kazanma fırsatı veririm.					
32.Not vermeyi bitirdikten sonra sıklıkla tüm sınıfın not dağılımına bakarım.					
33.Kendime özgü not verme yöntemim var.					
34.Değerlendirme ölçütleri konusunda sık sık meslektaşlarımla görüş alışverişinde bulunurum.					
Faktör 6: Not verme sürecinin algılanan öz-yeterliği					
35.Not verme, öğretmen olarak işimin en kolay parçasıdır.					
36.Bir öğrencinin çok çaba gösterdiğini fark etmek benim için kolaydır.					
37.Öğrenci başarısını tek bir notla veya puanla değerlendirmek benim için kolaydır.					
38.Not verirken, öğrencileri başarı açısından sıralamak benim için kolaydır.					

Appendix 2. *The path diagram of factor loadings of the scale of teachers' perceptions of grading practices.*


0.36 S1
0.29 S2
0.35 S3
0.19 S4
0.17 S5
0.35 S6
0.86
0.79
0.72
0.88
0.89
0.82
f1
1.00
0.44 S7
0.37 S8
0.32 S9
0.63 S10
0.30 S11
0.34 S12
0.31 S13
0.54 S14
0.22 S15
0.73 S16
0.63
0.72
0.60
0.70
0.75
0.74
0.75
0.70
0.61
0.42
f2
1.00
0.20 S17
0.28 S18
0.35 S19
0.76 S20
0.40 S21
0.39 S22
0.54
0.64
0.46
0.32
0.43
0.43
f3
1.00
0.19 S23
0.10 S24
0.05 S25
0.04 S26
0.08 S27
0.56 S28
0.50
0.56
0.62
0.62
0.60
0.45
f4
1.00
0.97 S29
0.81 S30
0.32 S31
0.29 S32
0.50 S33
0.47 S34
0.30
0.37
0.45
0.45
0.51
0.43
f5
1.00
0.73 S35
0.76 S36
0.40 S37
0.43 S38
0.55
0.33
0.69
0.78
f6
1.00