ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2025; 16(1); 48-58

Investigation of Activities For Reading Comprehension Skills: A G-Theory Analysis

Gülden KAYA UYANIK*   Serap ATAOĞLU**

Abstract

This study aimed to investigate the effectiveness of activities prepared to improve reading comprehension skills based on the variables of number of raters, evaluation criteria, and number of activities. Twelve raters evaluated five reading comprehension activities created by the researcher. A descriptive survey method grounded in a quantitative approach was employed. The study utilized five reading comprehension activities, commonly used in high school textbooks, and a rubric developed by the researcher, consisting of sixteen criteria based on relevant literature. After performing reliability and validity analyses on the rubric, the experts assessed the activities using this tool. The data collected from their evaluations were analyzed through generalizability theory. The EduG program was used to estimate variance values for both main and interaction effects according to generalizability theory, calculate the scores' reliability using G and Φ (Phi) coefficients, and conduct decision (D) studies. The findings revealed that each of the reading comprehension activities used to improve students' comprehension skills is different from each other. Additionally, it was concluded that increasing the number of criteria included in the rubric and increasing the number of expert raters would lead to a more accurate and effective evaluation of the activities.

Keywords: reading comprehension, activity, rater, rubric, generalizability theory

Introduction

Reading, one of the four basic language skills, plays an important role in language teaching and is defined as a receptive skill. The necessity of reading skills is not only crucial for the content of language learning, but also for other courses. Despite the use of various technological tools in today's education systems and the continuous development of these tools, reading maintains its place and importance in education and training practices, and education and training activities are widely based on reading skills (Smith, Snow, Serry, & Hammond, 2021). Meanwhile, research findings outlined in the literature (Floris & Divina, 2015; Hunt & Beglar, 2005) emphasize that reading alone is not enough. This skill is fully utilized and serves its purpose when the content of the text is understood. Reading comprehension skills are very important in educating individuals capable of thinking, questioning, producing, and inferring. In instances where students are unable to read fluently and comprehend the text adequately, it cannot be asserted that the act of reading has fully achieved its intended purpose. Reading comprehension includes readers' ability to recognize and perceive symbols in the text, thinking skills, and lifelong knowledge and experiences. The interest and desire for reading, the intended goals of reading, one's opinions about reading, and the location where the act of reading takes place all influence the process of reading comprehension (Akyol, 2005). Individuals who can comprehend what they read can be successful in various fields, including social, scientific, political, economic, and so on. The healthy execution of comprehension and expression skills in mother tongue lessons also affects students' success in other courses. Understanding the problem is very important in order to solve the problems encountered in the lessons (Güvendir, 2014).

* Assoc. Prof. Dr., Sakarya University, Faculty of Education, Sakarya-Türkiye, guldenk@sakarya.edu.tr, ORCID ID: 0000-0002-8100-6994
** Milli Eğitim Bakanlığı Sakarya, Sakarya-Türkiye, imra5425@gmail.com, ORCID ID: 0000-0002-3849-0493

To cite this article:
Kaya Uyanık, G., & Ataoğlu, S. (2025). Investigation of activities for reading comprehension skills: A G-Theory analysis. *Journal of Measurement and Evaluation in Education and Psychology, 16*(1), 48-58. https://doi.org/10.21031/epod.1548738

Received: 12.09.2024
Accepted: 18.02.2025

Each country conducts its own national exam (for Turkey - ABIDE) to identify and improve students' reading comprehension skills, and there are international practices such as PIRLS, TIMSS, and PISA. One of the purposes of these large-scale exams is to evaluate the effectiveness of teaching methods and materials that help students acquire reading skills and comprehension skills and the impact of these skills on other academic achievements (Mullis et al., 2009).

Several factors can help students develop reading comprehension skills. Some of these factors include the development of organs that assist reading, the use of strategies for reading comprehension, and enjoying the act of reading (Kim et al. 2021). Furthermore, factors such as the content and structural features of the text, its attractiveness to the student, its sufficiency regarding vocabulary and grammar rules, the student's level of knowledge, desire to read, and internalization of the text are also influential in reading comprehension (Baştuğ et al. 2019). Considering all these factors under one concept, it is indispensable to carry out activities within education and training activities for the improvement of reading comprehension skills, which is crucial for every age period. The activities designed for this purpose aimed at reading comprehension skills are among the instruments constantly used in education and training activities. Considering the reports of national and international large-scale exams, it is seen that the basis of all problems is the ability to understand and interpret what is read. To facilitate the development of these skills, activities targeting different age groups are designed within the educational processes, and these prepared activities are frequently employed in lessons.

The dictionary definition of "activity," which we frequently hear in the teaching field via the constructivist approach, is "the state of being active." Activities are significant in developing individuals' language skills, ensuring permanent learning, and helping them develop the habit of reading (Clarke et al., 2010). Activities are used to teach students both specific and general skills. The specific aim of the activities is to transform the learning outcome into behavior. The general aim is to equip students with skills such as creative thinking, critical thinking, etc. In other words, while the activities ensure the acquisition of the determined outcomes, they also play an important role in differentiating students' perspectives and making what they learn permanent. Students can improve their reading skills and gain creative thinking skills with the help of activities that they can relate to their lives and include problems they may encounter (Başpınar, 2013). Considering their effects on the acquisition and development of desired skills, it is possible to say that the activities have an important contribution to the teaching process. Activities allow students to develop their language and thinking skills by providing them with relevant acquisitions in a suitable period based on a predetermined plan, thus enabling them to learn easily, quickly, permanently, and systematically (Güneş, 2017).

When it comes to the activities prepared for reading comprehension in educational processes, the potential obstacles in the evaluation of these activities should also be considered. It is seen from the studies that one of the factors affecting the literature is the issue of who and how the activities prepared for reading comprehension skills will be evaluated. (Long & Pang, 2015; Myford & Wolfe, 2003; Snyder, Caccamise & Wise, 2005; Şata & Karakaya, 2021; Wiseman, 2012). In this regard, studies that reveal the effect of the rater in reading comprehension activities and the characteristics of the rubric used in scoring are required. One of the theories that helps to reveal the effect of these statistical properties is the Generalizability Theory (G-Theory).

Generalizability (G) Theory is based on the analysis of variance and is similar to Classical Test Theory (CTT). However, unlike the CTT, in G-theory, the sources of the error rate in the observed scores can be obtained in detail. In G-theory, error rates can be determined separately for each error source and for the interaction of these sources (Shavelson and Webb, 1991). G theory uses the concepts of the "facet, object of measurement, condition, and design." The concept of facet is the definition used for each of the sources of variability in the universe (Brennan, 2001). The source of variance in the universe whose effect is examined for the research purpose is the "object of measurement." In the theory, variance due to the object of measurement is desirable, while large variance due to facets is undesirable. The different levels that facets have are called conditions (Guler, Kaya Uyanik, & Tasdelen Teker, 2012). For instance, consider a scenario in which five raters evaluate a 10-question examination administered to a class of 50 students. In this context, the students' exams represent the objects of measurement, while raters and questions serve as sources of error, which are examined and treated as facets. The condition for the rater facet in the study was five, whereas the condition for the question facet was ten.

Another concept that needs to be addressed in G-theory is the designs of facets. Crossed or nested are the types of designs that are considered in this theory (Shavelson & Webb, 1991). A crossed design is when all the conditions of one facet are associated with all the conditions of another facet. A nested design is a type of design in which a condition of one facet is associated with some conditions of another facet. These designs also differ in terms of notation, with the crossed design represented by "x" and the nested design represented by ":" (Shavelson & Webb, 1991).

In G-Theory, there are two different coefficients, G and Phi, as reliability coefficients. The main difference between these coefficients is that the sources of variance of the object of measurement considered in the assessments are examined in relative and absolute terms. The G coefficient is used for relative assessment, and the Phi coefficient is a usable absolute assessment (Brennan, 2001).

In G-theory, reliability can be obtained for two different cases called Generalizability (G) and Decision (D) studies. The G study is concerned with generalizing to the universe based on the universe in which the measurements are made, thus aiming to provide information about the sources of variability in the sample. In the D study, scenarios are created for a specific purpose by using the information obtained in the G study, and decision-making is aimed at these scenarios (Brennan, 2001; Guler, Kaya Uyanik, & Tasdelen Teker, 2012; Nalbantoglu & Gelbal, 2011).

This study, which emphasizes the importance of reading and reading comprehension skills in educational activities, aimed to question the effectiveness of the activities prepared for reading comprehension by evaluating them by different raters and increasing their efficiency by identifying their deficiencies. For this purpose, a completely crossed randomized design of a (activity) x r (rater) x c (criterion) was created. With the design created, answers to the questions of variance values for the main and interaction effects and reliability of the test were sought. In addition, Decision (D) studies were conducted, and scenarios suitable for the features of the facets were created. In this regard, the main problem of the research is as follows:

What are the variance values for the main and interaction effects of the a (activity) x r (rater) x c (criterion) completely crossed randomized design and the reliability of the test as a result of the examination of the activities prepared for reading comprehension skills by different raters with the specified criteria? In the study, answers were sought to three sub-problems.

1. What are the variance values for the main and interaction effects in a (activity) x r (rater) x c (criterion) completely crossed randomized design?

2. What are the G and Φ (Phi) coefficients calculated for the reliability of scores in a (activity) x r (rater) x c (criterion) completely crossed randomized design?

3. What are the reliability values obtained from scenarios created with different numbers of raters and criteria in a (activity) x r (rater) x c (criterion) completely crossed randomized design?

Method

Research Design

In this study, which was conducted to examine the activities prepared to measure reading comprehension skills by different raters and to make suggestions for increasing the efficiency and effectiveness of these activities, the "descriptive survey" design, one of the quantitative research methods, was used. The descriptive survey model aims to describe a past or ongoing situation as it exists. The individual, object, or event to be researched is defined within its own conditions, as it is. The researcher does not attempt to intervene, influence, or change shape (Karasar, 2010). The main purpose of this model is to describe and explain the situation in detail (Çepni, 2010).

Study Group

The study group consisted of 12 raters who are teachers working in measurement and evaluation centers located in different provinces of Turkey and are experts in their fields. Demographic information of the raters is given in Table 1.

Table 1

Demographic Information of the Study Group

Variables	Category	Frequency (f)	Percent (%)
Gender	Female	8	66,7
	Male	4	33,3
Profession	Turkish language and literature	4	33,3
	Curriculum development	4	33,3
	Measurement and Evaluation	4	33,3
Educational Status	Bachelor	2	16,7
	Master	6	50
	Doctorate	4	33,3
Professional seniority	6-10 years	1	8,3
	11-15 years	6	50
	16-20 years	3	25
	21-25 years	2	16,7
	Total	12	100

Table 1 shows that, of the participants, 8 were female, 4 were male, 4 were experts in Turkish language and literature, 4 were in curriculum development, and 4 were in measurement and evaluation. Two participants had bachelor's degrees, six had master's degrees, and four had doctorate degrees. It was observed that the least experienced participant had 6 years of seniority and 92% had more than 10 years of experience. As can be understood from these data, the study was conducted with experienced and expert evaluators.

Data Collection Tools

In this study, five activities to measure reading comprehension skills in the 10th and 12th grade Skill-Based Turkish Language and Literature Books written by the researcher within the General Directorate of Secondary Education (GDSE) of the Ministry of National Education (MoNE) and a rubric consisting of 16 items created by the researcher to determine the suitability of these activities were used as data collection tools.

Activities

The activities used as one of the data collection tools of this study were selected from the Skill-Based Activity Books written by Turkish Language and Literature and Turkish teachers employed by the GDSE in the 2020-2021 academic year. Three of the selected activities are at 10th-grade level, and the other two are at 12th-grade level. One of the researchers who conducted this study wrote the activities for the GDSE. The written activities aim to identify and improve students' reading and reading comprehension skills. The activities used in the research aimed to develop reading skills among domain skills and critical thinking skills among general skills. The learning outcome-based activities were sent to field experts, curriculum development experts, measurement and evaluation experts, language experts, and guidance experts employed under the GDSE. The opinions of these experts were taken, the activities revised in line with the feedback received were finalized, and the activities collected in an interactive book were uploaded to the GDSE and made available to students and teachers.

Rubric

The rubric was developed by the researchers. First, a question pool of 26 items was prepared based on the literature to determine the effect of the activities used in the study on improving reading comprehension skills. Upon examination, repetitive items, items with little relationship with the content, and items with a broad scope were removed, and a trial form consisting of 16 items was prepared. The prepared form was presented to expert opinion. Opinions were received from two faculty members who were experts in the field of Turkish education and two faculty members who were experts in the field of measurement and evaluation. The items were arranged according to the feedback from the experts and the experts reached a consensus on the final version of the form. At the end of these processes, the final form consisting of 16 items was created.

Data Analysis

The data obtained from the activities examined through the rubric and the raters were analyzed according to the Generalizability Theory. In the study, the variance values for the main and interaction effects of the a (activity) x r (rater) x c (criterion) completely crossed randomized design, which was formed by analyzing the activities prepared for reading comprehension skills (object of measurement) by different raters with the specified criteria, were examined. G and Φ (Phi) coefficients were calculated for the reliability of the test scores of the design used in the study. In addition, decision (D) studies were conducted, and future scenarios were created. The EduG program was utilized to estimate the main and interaction effect’s variance values according to the generalizability theory, to calculate the reliability of the scores, and to carry out D studies.

Results

The variance values for the main and interaction effects of the a (activity) x r (rater) x c (criterion) completely crossed randomized design, which was formed by evaluating the activities prepared for reading comprehension by different raters using specified criteria, were investigated, and the results are given in Table 2.

Table 2

Variance Values and Total Variance Explanation Rates Estimated by the G Study Regarding the axrxc Design

Source of Variance	df	Sum of Squares	Mean Squares	Variance	%
a	4	99.09792	24.77448	7.56847	63.2
r	11	10.56146	0,96013	0.29312	2.8
c	15	5.05521	0.33701	0.00209	0.2
ar	44	43.02708	0.97788	1.63254	14.3
ac	60	21.96875	0.36615	1.08145	9.8
rc	165	12.05729	0.07307	0.01005	1.1
Arc,e	660	13.10625	0.01985	0.96780	8.6
Total	959	204.87396			100%

An analysis of the variance estimated and total variance explained ratios of the axrxc fully crossed randomized design in Table 2 shows that the variance component estimated for the main effect of activity (a) explains 63.2% of the total variance. In generalizability studies, the main effect taken as the object of measurement is evaluated as the variance of the universe score and refers to the differentiation between activities in this study in terms of the measured feature (Shavelson & Webb, 1991; Brennan, 2001; Guler, Kaya Uyanik, & Tasdelen-Teker, 2012; Kaya Uyanik & Guler, 2016). The ratio of the variance estimated for activities to the total variance should be large. This indicates that differences between activities can be revealed in the dimension obtained by measurement (Brennan, 2001; Kaya Uyanik & Guler, 2016). According to the results obtained in this study, it can be said that the evaluation of activities based on criteria can reveal the differences between activities. The variance component estimated for the rater main effect (0.29) explains 2.8% of the total variance. This value is the third smallest value. The rater's main effect is due to inconsistency between raters' ratings. Therefore, it is desirable that this effect is low. The variance component estimated from the G study for the main effect of criterion (c) explains 0.02% of the total variance. The criterion's main effect shows the degree of differentiation of the difficulty level of each measurement unit (item) in the rubric. According to the results obtained, it can be interpreted that the attainability levels of the criteria used to measure reading comprehension skills were similar to each other.

The activity x rater (ar) interaction effect explains 14.3% of the total variance, and this value is the second highest value. The activity and rater interaction refers to the inconsistency of the raters in terms of generosity-severity in scoring for some activities. In this case, it was concluded that although the raters gave generally consistent results, there were differences between their scoring in some activities. The activity x criterion (ac) interaction effect explains 9.8% of the total variance and is the third largest variance obtained. This shows that the relative status of certain activities differs from one criterion to another. In other words, it can be interpreted that the scores given to the criteria differ from activity to activity. Rater x criterion (rc) interaction effect explains 1.1% of the total variance. This value is the second smallest variance obtained. For this result, it can be interpreted that there is no difference between the raters according to the criteria. The variance component of the activity x rater x criterion (residual) interaction effect variance explains 8.6% of the total variance. A large residual variance is an indication of a large interaction of activity, rater, criterion, unmeasured sources of variability, and/or random errors. When the values obtained are examined, it is observed that the rate of random errors is low for the study.

G and Φ (Phi) coefficients were calculated for the reliability of scores in the axrxc completely crossed randomized design. In the rubric containing the criteria in the study, there are sixteen criteria in total and these criteria were scored by twelve raters. In this case, the G coefficient was 0.885, and the Phi coefficient was 0.863. It can be said that the measurements obtained from the measurement tool used are reliable.

Decision (D) studies are conducted using the variance values calculated over the data used in the generalizability study. D study allows the estimation of the coefficients G and Phi for the reliability values by decreasing and increasing the conditions of the facets in the universe G, respectively. Table 4 shows the values of G and Phi coefficients calculated by keeping the criterion facet constant and decreasing and increasing the number of raters, and the values of G and Phi coefficients calculated by keeping the rater facet constant and decreasing and increasing the number of criteria in the D study.

Table 3

axrxc Fully Crossed Randomized Design D Study Results

	Number of Rater	G coefficient	Φ coefficient
Number of Criteria: 16	5	0.782	0.743
	10	0.850	0.845
	15	0.886	0.864
	20	0.887	0.875
	25	0.889	0.877
	Number of Criteria		
Number of Rater: 12	5	0.740	0.726
	10	0.810	0.799
	15	0.882	0.861
	20	0.906	0.887
	25	0.914	0.909

In Table 3, G and Phi coefficients were calculated for the two different cases. In the first case, the number of criteria was kept fixed at 16, and the number of raters varied from 5 to 25. Also, in the second case, the number of raters was constant at 12, and the number of criteria varied from 5 to 25. When the number of criteria was kept fixed and the number of raters was changed, it was observed that the reliability value increased as the number of raters increased. However, it was observed that after 15 raters, the increase in reliability was significantly low for every 5-rater increase. Similarly, G and Phi coefficients were calculated when the number of raters was kept constant at 12, and the number of criteria was 5, 10, 15, 20, and 25. When the number of raters was kept constant and the number of criteria was changed, the highest reliability value was obtained from the scenario where the number of criteria was 25. It was observed that the reliability value increased as the number of criteria increased.

Discussion and Conclusion

The present study examined the design obtained by evaluating the activities for reading comprehension skills by using rubrics with the expert raters. The analysis revealed the criteria for more effective and efficient development of activities prepared to measure reading comprehension skills in terms of structure and content. An axrxc design was used in the study. An analysis of the axrxc design examined in the study regarding sources of variance shows that the largest source of variance is due to activity. Considering that the study involved activities designed for reading comprehension, these activities are diverse rather than being of a single type. Furthermore, the high value of this source of variance indicates that the effect of each activity to be used for reading comprehension is different. It can be concluded that students can better develop this skill if they interact more with reading comprehension activities. In addition, considering that the situation emphasized by each activity will be different, it is an important conclusion that the scoring keys should be arranged accordingly. Supporting this result, the study also revealed that for the axrxc design, the activity criterion interaction effect was also a significant source of variance. The interaction of activity and criterion indicates that the criteria are met in some activities and not in others. This result shows that not every activity met every criterion, so it can be stated that the criteria including the features that should be present in reading comprehension activities cannot be provided with only one activity alone, and it would be more accurate to use more than one activity. Considering the information obtained from these two findings, a small number of activities aimed at measuring reading comprehension skills will create a deficiency in terms of meeting the criteria. Therefore, a large number of activities will increase reading comprehension skills. Recent studies have emphasized the importance of using activities that serve this purpose in order to increase reading comprehension skills (Akyol & Ketenogluarter Kayabasi, Topuz 2018; Collins, et al. 2020; Siti & Mumu, 2022; Brilliananda & Wibowo, 2023).

The necessity of increasing the number of activities and using a rubric in evaluating these activities has been emphasized in many studies in the literature because it reveals the learning objectives clearly and understandably, reduces the errors involved in the evaluation, and provides an opportunity to complete missing learning (Arter, 2002; Dunbar, Brooks, & Miller, 2006; Hall & Salmon, 2003; Oaklef, 2009; Wolf and Steven, 2007). Another question that comes to mind is the number of criteria in the rubrics used in the evaluation. In the decision studies conducted in the study, it was observed that the reliability of the study increased as the number of criteria increased. In the study, the maximum value for the number of criteria was 25, and the highest reliability was obtained from this value. Similarly, when the number of criteria was reduced to five, a value around 0.70 was obtained. In this respect, it can be concluded that the number of criteria should be at least five and that there is no upper limit. One of the most valid ways to ensure objectivity and inter-rater reliability in multi-rater measurements is to use rubrics (Jonsson,& Svingby,2007). The reliability and validity of rubrics have been examined from various perspectives. While some researchers have focused on the objectivity of rubrics (Rezaei, & Lovorn,2010 ; Spandel, 2006; Wolfe, 1997), others have critiqued them as being overly reductive (Kohn, 2006; Mabry, 1999). However, the interaction result in all the studies mentioned is that using rubrics is a more reliable way than not using them. In this case, what to consider when using rubrics is another important issue that increases reliability. In the literature, it has been emphasized that the number of items should be increased as well as different factors (Henson, R, & Thompson, 2002; Hellman, Fuqua & Worley, 2006). At this point, when the studies in the literature and the results obtained from this study are interpreted together, using rubrics increases reliability and it can be said that for a more reliable measurement, the criteria in the rubrics should be at least 5 and reliability will increase as the number of criteria increases.

Another important source of variance for the axrxc design is the event rater interaction effect. The activity and rater interaction refer to the inconsistency of the raters in terms of generosity-rigor in scoring for some activities. In this case, it was concluded that although the raters gave generally consistent results, there were differences between their scoring in some activities. This result revealed that the activities may have different effects on different raters, leading to a scoring disadvantage. Within this perspective, it can be claimed that scoring the activities to measure reading comprehension skills by a single rater would create deficiencies, whereas evaluating the activities by more than one rater would increase the efficiency of the activities. This result is consistent with the literature in terms of both rater reliability and reducing the error rate of the process (Alkan & Doğan, 2023; Kim, 2020; Kim et al. 2021). This result obtained from the present study and other studies in the literature raises the question of the required number of raters. The finding obtained through the decision study conducted in the present study has been a relevant answer in this context. Based on the decision studies conducted to answer the question of how much the number of raters should be increased, it was concluded that the reliability of the study increased as the number of raters increased, but the increase in the reliability value was not very high if the number of raters was above 15, so it would not be practical to increase the number of raters above 15.

The findings of the study necessitate separate discussions for classroom assessments and large-scale assessments. According to the results, when the number of raters is five, the reliability coefficient exceeds the interactionally accepted threshold of 0.70 for Cronbach's Alpha. Considering that for classroom assessments with multiple raters, an acceptable reliability coefficient can be as low as 0.60 (DeVellis & Thorpe, 2021), it can be stated that even with fewer than five raters, acceptable reliability can still be achieved. Thus, in classroom assessments, having multiple raters invariably yields more reliable results compared to assessments conducted with a single rater. On the other hand, Cizek (2009) highlights that there should be procedural distinctions between classroom and large-scale assessments. For large-scale examinations, the acceptable threshold for reliability is higher than that for classroom assessments. When evaluated in the context of large-scale examinations, the finding that 15 raters represent an upper limit is both significant and practical. At both national and international levels, large-scale examinations often involve open-ended questions that require multiple raters. The number of raters required for evaluating these exams becomes a critical factor in managing the assessment process. In Turkey, for instance, the pilot implementation and the first official administration of the "four-skill Turkish language exams"—which consist of both open-ended and multiple-choice questions—were conducted in 2024. These exams were administered to approximately 10,000 students across 4th, 7th, and 11th grades. For the writing and speaking skills components, open-ended assessments were used, and multiple raters were involved in the evaluation process for each grade level. In the pilot study, which involved approximately 2,000 participants, it was reported that five raters were sufficient for reliable scoring (MoNE, 2020). However, the significant discrepancy between the number of participants in the pilot study and the actual implementation (approximately 10,000) indicated the need for an increased number of raters. Based on the results of this study, it can be concluded that 15 raters are sufficient for the four-skill language examinations. On the other hand, it is worth noting that working with 15 raters is not easy. In this context, it is recommended that the selection of raters and the harmony processes between the obtained scores should be carried out with scientific steps.

The most significant finding of this study, which involved the scoring of reading comprehension activities by different raters based on specific criteria, is that these activities exhibit a high level of variance. Accordingly, it can be stated that the activities differentiate in terms of assessing students' reading comprehension skills. This suggests, indirectly, that students need to encounter a wide variety of activities in order to develop their reading comprehension skills. In this regard, it is recommended that teachers, school administrators, and educational policymakers emphasize the importance of numerous reading activities to enhance students' reading comprehension abilities.

In scoring using rubrics, the difference between the raters decreases and compliance increases. Therefore, it is recommended that the activities for reading comprehension be scored using a rubric to determine the reliability of the rater and obtain more reliable results. A 16-item rubric was used in this study. The findings indicate that as the number of criteria increases, the reliability of the raters also improves. Therefore, it is recommended to increase the number of criteria in the scoring rubric used to assess reading comprehension skills as much as possible. On the other hand, according to the results of this study, it is considered important for the reliability of the rubric that the number of criteria should not be fewer than five. Additionally, it was observed that after 20 criteria, increasing the number to 25 did not result in a sharp improvement. In this context, it is suggested that the rubric should include at least five criteria, and considering usability and practicality, there is no need to exceed 25 criteria.

Similar to the present study, in which it was found that a high number of raters increased reliability, different raters could be used in scoring, and the number of raters could be increased up to 15. It is recommended to keep this number around 15, especially in large-scale exams, as increasing the number of raters above 15 will not make a big difference in the results.

The present study, intended to determine the effectiveness level of reading skills activities and their deficiencies concerning structure and content, can also be applied to writing, speaking, and listening skills, which are among the basic language skills, and their rater reliability can be examined. Most of the raters who contributed to this study are experts and experienced in their fields. It could be taken into consideration that experienced raters make more accurate interpretations and judgments than less experienced raters (Jorgenson, 1975), and similar studies could be conducted by grouping raters according to their experience. This study examined the activities for reading comprehension skills prepared by the researcher and used in the MoNE. Similar studies can be conducted by utilizing different types, content, and grade-level activities to determine reading comprehension skills.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval for the study was received from Sakarya University, Educational Sciences Ethics Committee dated 15.02.2023 numbered E-61923333-050.99-222299

References

Akyol, H. (2005). *Turkish primary reading and writing teaching*, Ankara: PegemA.

Akyol, H., & Ketenoğlu Kayabaşı, Z. E. (2018). Improving the Reading Skills of a Students with Reading Difficulties: An Action Research. *Education and Science, 43*(193). https://doi.org/10.15390/EB.2018.7240

Alkan, M., & Doğan, N. (2023). A Comparison of Different Designs in Scoring of PISA 2009 Reading Open Ended Items According to Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology, 14*(2), 106-117. https://doi.org/10.21031/epod.1210917

Arter, J. (2002). Rubrics, scoring guides, and performance criteria. In C. Boston (Ed.), *Understanding Scoring Rubrics a Guide for Teachers* (p. 21-31). Office of Educational Research and Improvement.

Başpınar Yörük, N. (2013). *An investigation on the use of creativity development methods in 6th grade Turkish course reading activities.* (Master thesis), University of Necmettin Erbakan Üniversitesi, Konya. Accessed from YOK Thesis Center database (Dissertation No: 348744).

Baştuğ, M., Hiğde, A., Çam, E., Örs, E., & Efe, P. (2019). *Strategies, techniques, practices to improve reading comprehension skills.* Ankara: PegemA.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer Verlag.

Brilliananda, C., & Wibowo, S. E. (2023). Reading Strategies for Post-Pandemic Students' Reading Comprehension Skills. *International Journal of Elementary Education, 7*(2).

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory into practice, 48*(1), 63-71.

Clarke, P. J., Snowling, M. J., Truelove, E. & Hulme, C. (2010). Ameliorating Children's Reading-Comprehension Difficulties: A Randomized Controlled Trial. Psychological Science, 21(8), 1106–1116. https://doi.org/10.1177/0956797610375449

Collins, A. A., Compton, D. L., Lindström, E. R., & Gilbert, J. K. (2020). Performance variations across reading comprehension assessments: Examining the unique contributions of text, activity, and reader. *Reading and Writing, 33*(3), 605-634.

Çepni, S. (2010). *Introduction to research and project work.* Trabzon: Celepler.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications.* Sage publications.

Dunbar, N. E., Brooks, C. F. & Miller, T. K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education, 31*(2), 2006, 115-128.

Floris, F. D., & Divina, M. (2015). A study on the reading skills of EFL university students. *Teflin Journal, 20*(1), 37–47.

Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Generalizability theory.* Ankara: PegemA.

Güneş, F. (2017). Activity approach in teaching Turkish, *Journal of Native Language Education, 5*(1), 48-64. https://doi.org/10.16916/aded.286415

Güvendir, M. A. (2014). Öğrenci başarılarının belirlenmesi sınavında öğrenci ve okul özelliklerinin Türkçe başarısı ile ilişkisi. *Eğitim ve Bilim, 39*(172).

Hall, E. K. & Salmon, S. J. (2003). Chocolate chip cookies and rubrics helping students understand rubrics in inclusive settings. *Teaching Exceptional Children, 35*(4), 8-11.

Hellman, C. M., Fuqua, D. R., & Worley, J. (2006). A reliability generalization study on the survey of perceived organizational support: The effects of mean age and number of items on score reliability. *Educational and psychological measurement, 66*(4), 631-642.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development,* 35, 113-126.

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language, 17*(1), 23–59.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review, 2*(2), 130-144.

Jorgenson, G. W. (1975). An analysis of teacher judgments of reading level. American Educational Research Journal, 12 (1), 67-75. https://doi.org/10.2307/1162581

Karasar, N. (2010). *Scientific research method.* Ankara: Nobel.

Kaya Uyanık, G., & Güler, N. (2016). Examining the reliability of concept map scores: An example of a crossover mixed design in generalizability theory. *Hacettepe University Faculty of Education Journal 31*(1). 97-11. http://doi.org/10.16986/HUJE.2015014136

Kim, J. S., Relyea, J. E., Burkhauser, M. A., Scherer, E., & Rich, P. (2021). Improving elementary grade students' science and social studies vocabulary knowledge depth, reading comprehension, and argumentative writing: A conceptual replication. *Educational Psychology Review*, 1-30.

Kim, Y. S. G. (2020). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology, 112*(4), 667.

Kohn, A. (2006). The trouble with rubrics. *English Journal*, 95(4), 12–15.

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. Thinking Skills and Creativity, 15, 13-25.

Mabry, L. (1999). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan*, 80(9), 673–679.

MoNE. (2020). *Turkish Language Exam in Four Skills: Pilot Study Results.* https://www.meb.gov.tr/meb_iys_dosyalar/2020_01/20094146_Dort_Beceride_Turkce_Dil_Sinavi_Ocak_2020.pdf

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement, 4*(4), 386-422.

Nalbantoğlu F., & Gelbal S. (2011). Comparison of different designs with generalizability theory at the communication skills station scale, *Hacettepe University Faculty of Education Journal, 41*, 509-518. https://dergipark.org.tr/tr/download/article-file/87423

Oaklef, M. (2009). Using rubrics to assess information literacy: An examination of methodology and ınterrater reliability. *Journal of the American Society for Information Science and Technology, 60*(5), 969-983. https://doi.org/10.1002/asi.21030

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing*, 15(1), 18-39. https://doi.org/10.1016/j.asw.2010.01.003

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage. http://doi.org/10.1002/9781118445112.stat00068

Siti, M., & Mumu, M. (2022). The effect of critical multiliteracy learning model on students' reading comprehension. *International Journal of Educational Qualitative Quantitative Research (IJE-QQR), 1*(1), 28-33.

Smith, R., Snow, P., Serry, T., & Hammond, L. (2021). The role of background knowledge in reading comprehension: A critical review. *Reading Psychology, 42*(3), 214-240.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders, 25*(1), 33-50.

Spandel, V. (2006). In defense of rubrics. *English Journal*, 96 (1), 19–22. https://doi.org/10.58680/ej20065683

Şata, M., & Karakaya, İ. (2021). Investigating the Effect of Rater Training on Differential Rater Function in Assessing Academic Writing Skills of Higher Education Students. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(2), 163-181. https://doi.org/10.21031/epod.842094

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150-173.

Wolf, K. & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, *7*(1), 3-14.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4(1), 83–106.