



ADIYAMAN UNIVERSITY

Journal of Educational Sciences

AUJES

Volume:11

Issue:1

June 2021



ADİYAMAN UNIVERSITY JOURNAL OF EDUCATIONAL SCIENCES

An international refereed e-journal and publishes two issues per year.

Year: 2021 Issue: 1

Editor

Assoc. Prof. Dr. Ali ÜNİŞEN, Adıyaman University, Turkey

Editorial Board

Prof. Dr. Adnan BAKI, Karadeniz Technical University, Turkey

Prof. Dr. Feride BACANLI, Gazi University, Turkey

Prof. Dr. Ğlhan ERDEM, İnönü University, Turkey

Prof. Dr. Yüksel DEDE, Gazi University, Turkey

Prof. Dr. Süleyman Nihat ŞAD, İnönü University, Turkey

Prof. Dr. Elizabeth KING, University of Wisconsin – Whitewater, USA

Assoc. Prof. Dr. Muhammed Fatih DOĞAN, Adıyaman University, Turkey

Assoc. Prof. Dr. Seval KIZILDAĞ ŞAHİN, Adıyaman University, Turkey

Assoc. Prof. Dr. Amy ELLIS, University of Georgia-Athens, , USA

Assist. Prof. Dr. Bilal KALAKAN, Adıyaman University, Turkey

Assist. Prof. Dr. Reza Feyzi-BEHNAGH, University at Albany, SUNY, USA

Assist. Prof. Dr. Caro Williams-PIERCE, University Maryland, College Park, USA

Assist. Prof. Dr. Torrey KULOW, Portland State University, USA

Assist. Prof. Dr. Elise LOCKWOOD, Oregon State University, USA

Dr. Vahide YİĞİT GENÇTEN, Adıyaman University, Turkey

Dr. MEHMET GÜLTEKİN, Adıyaman University, Turkey

Dr. Crystle MARTIN, University of California, Irvine, USA

Contact Information

Assoc. Prof. Dr. Ali ÜNİŞEN

Adıyaman University Faculty of Education, Adıyaman, Turkey

aunisen@adiyaman.edu.tr

The ideas published in the journal belong(s) to the author(s).

Adıyaman University Journal of Educational Sciences, is an international refereed (peer- reviewed) journal and published two issues per year by Adıyaman University
© 2019 ADYU. All rights are reserved.

Table of Contents

Research Article

Learning English as a Second Language through Translanguaging in Early Years
1-8
İskender Gelir
Determination of Cognitive Structures and Misconceptions of Pre-service Science Teachers' Regarding the Concept of "Energy"
9-25
Filiz Avcı
Measurement and Assessment Literacy Levels of Teachers in Terms of Some Variables
26-35
Ayşe Özlem Ergül, Sevda Çetin
Effects of Task Complexity on Text Easibility and Coherence of EFL Learners' Narrative Writing
36-47
Mine Yıldız, Savaş Yeşilyurt
Sample Size Determination and Optimal Design of Randomized/Non-equivalent Pretest-posttest Control-group Designs
48-69
Metin buluş

REVIWER LİST OF THIS İSSUE

Dr. Bekir CANLI

Dr. Cemile DOĐAN

Dr. Gökhan GÜVEN

Dr. Esra AÇIKGÜL FIRAT

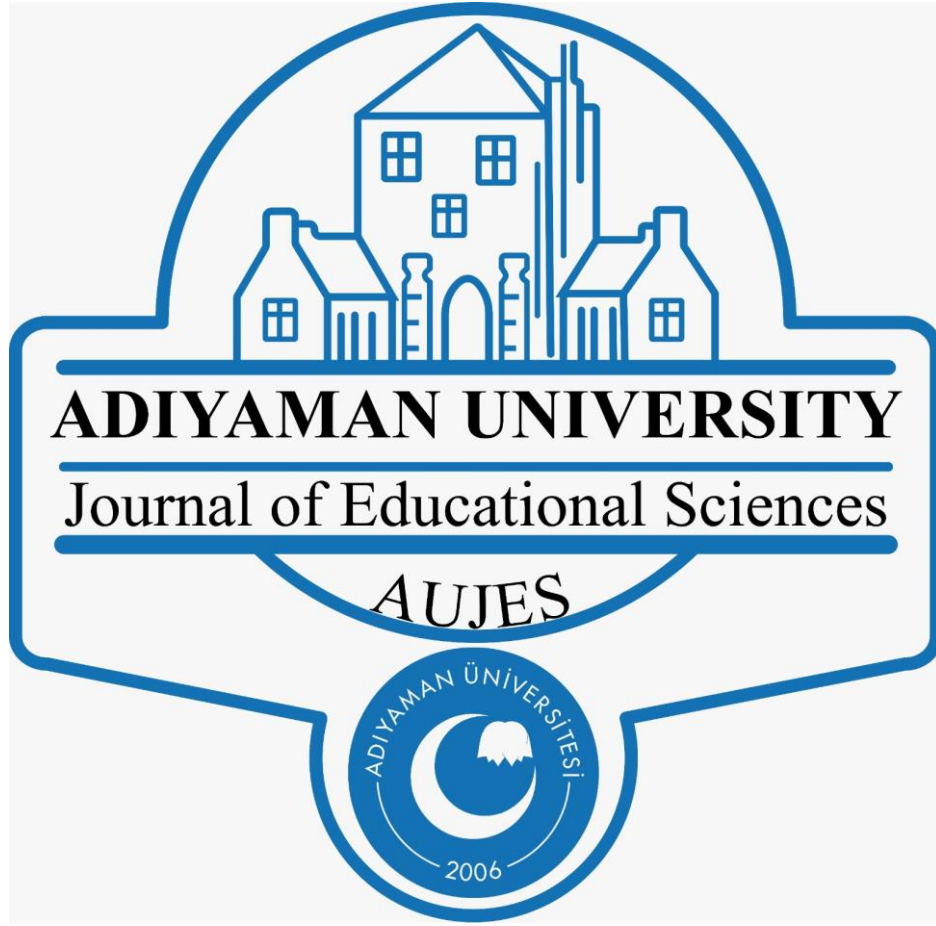
Dr. İlhan KOYUNCU

Dr. Neşe GÜLER

Dr. Özge GÜMÜŞ

Dr. Sedat ŞEN

Dr. Burak AYDIN



Article History

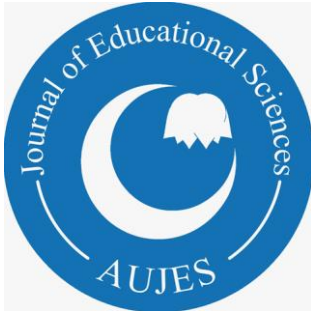
Received: 08.03.2020

Received in revised form: 01.04.2021

Accepted: 02.04.2021

Available online: 29.06.2021

Article Type: Research Article




ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

<https://dergipark.org.tr/tr/pub/adyuebd>

Learning English as a Second Language through Translanguaging in Early Years*

İskender Gelir¹

¹Early Childhood Education, Siirt University/ Early Childhood Education, Sultan Qaboos University, Oman 

To cite this article:

Gelir, İ. (2021). Learning English as a second language through translanguaging in early years. *Adiyaman University Journal of Educational Sciences*, 11(1), 1-8.

* A version of this work was presented at VII. International Eurasian Educational Research Congress.

Learning English as a Second Language Through Translanguaging in Early Years

Iskender Gelir^{1*},

¹ Early Childhood Education, Siirt University, Turkey/Early Childhood Education, Sultan Qaboos University, Oman

Abstract

This study examines preschool children's learning English as a second language in early years education. It is qualitative and uses ethnography as the data collection and analysis method. It was conducted between October 2019 and January 2020 in a private nursery in a city in Turkey. The participant children are 5 and 6 years old. In this study, different bilingual models are discussed, and the construct of translanguaging is used to examine children's second language and literacy learning, and their interactions with the teacher and peers in the classroom. The findings show that preschool children learn language and literacy through a flexible language teaching method. The study demonstrates that their expressive skills, vocabulary learning, and math develop in classroom activities. The findings also indicate that the model enables the participant children to improve their first language (Turkish) skills.

Key words: English, second language learning, models, preschool, translanguaging

Introduction

In early childhood education, English is widely taught and spoken as a second/foreign language in both private and state funded schools in many countries (Lugossy, 2018; Prošic- Santovac and Radovič, 2018). This provides English to have prestigious in many societies. Thus, parents and teachers have appetite for their children and students to learn English from early years. In spite of this, Koru and Åkesson found that the importance of English is not understood considerably and that there is a gap between students' competence in English countries such as Turkey and Brazil. In Turkey, parents increasingly want their children to learn English in early years education. However, in Turkey learning English as a second language is mainly taught in private nurseries (Sarica, 2019). In state-funded nurseries, English is taught in kids club that are arranged after daily activities have finished in Turkish (Official Gazette, 2014, article no:83).). But kids club are not arranged in every nursery as they are paid for and based on parents' choices (Gelir, 2020). In kids club, it is considered that there is language separation as the English language teacher teaches English (mainly grammar and speaking skills) after the preschool teacher has finished daily activities (Gelir, 2020). The language separation is mainly associated with the immersion method.

According to Çetintaş and Yazıcı (2016), immersion method is mainly used as a language teaching model in early years education in Turkey. The immersion method is criticised for separating languages (Schwartz and Asli, 2014). In recent years, however, translanguaging model is applied for teaching children two or more languages flexibly. Translanguaging is also used as a theoretical construct.

This study will indicate how translanguaging can be used as a teaching model in early years education. However, it seems that there is limited research on translanguaging both as a teaching model and construct in the relevant literature in Turkey. For example, this model is used in a few nurseries, and it has been used to examine English language teachers' perceptions of teaching English as second language in different levels of education such as primary and secondary (Yuvayapan, 2019) and Syrian refugee children in Turkey (Baytas and Seyma, 2019). Therefore, the current study aims to contribute to this gap by using the construct of translanguaging to investigate preschool children's learning English as a second language in a private nursery. This study addresses the following research question:

- How do preschool children learn English language and literacy?

Theoretical Framework

There are two main approaches to second language learning: bottom up and top down. Bottom- up approaches suggested that learning a second language (including reading) starts from small unit of meaning

* Corresponding Author: *Iskender Gelir*, campus97@hotmail.com

such as individual sound and phoneme to general knowledge of structure and language (syntactic knowledge). This means that sounds and words play important roles in understanding of language and structure as 'decoding sounds and pronouncing words is seen as a means to gain understanding' (Gregory, 2008, p.109). In contrast, top-down approaches claim that learning a second language begins with understanding of general knowledge to particular. These approaches argue that experience is crucial to a second language learning as a learner builds meaning through experience. In other words, the recognition of printed and written symbols is seen as a stimulus for remembering meaning (Gregory, 2008).

There are also theories of second language learning. In this paper, cognitive, sociocultural and sociolinguistic theories will be discussed with a focus on sociolinguistic theories. Cognitive approaches argued that there are systems that function without awareness and that each individual can access to. Cognitive approaches argued that a second language learning is integrated into the cognitive mechanism that is already established through the first language. These approaches claimed that the same learning mechanism is used to understand structures and patterns form a second language. Cognitive approaches also viewed memory, sentence processing, information processing and attention as important in second language learning (Mitchell, Myles and Marsden, 2013). On the other hand, sociocultural theories considered that a second language is learned in interactions with more knowledgeable adults or peers in social contexts. Sociocultural theories are based on Vygotsky's works (1978, 1986). Vygotsky argued that the child first learns in interactions with people around her/him, and then s/he internalises his learning. That is, this theory viewed a second language learning as a social practice. Although sociocultural theory claimed that the child neurobiology is crucial to human mental functioning, her or his cognitive system develop in interaction with people and the social context (Lantolf and Thorne, 2006; Lantolf, Thorne & Poehner, 2015). For sociocultural theories, language is an important cultural tool that mediates an individual connection to the social context and people around.

Similarly, sociolinguistic theory claimed that a second language is learned in a social context that affects children language use and development (Tarone, 2007). This means that this theory investigated how a social context and learner's participation affect the rate and direction of second language relearning and outcome (Mitchell, Myles and Marsden, 2013). Tarone (2007) argued that a second language is not learned in vacuum, rather, it is learned from and with people. The relationships between the learner and social context have effects on cognitive development. Scholars advocating sociolinguistic theories also emphasised the relationship between culture and language use. Researchers such as Bayyurt (2013) suggested that cultural elements such as accent can be included in teaching a second/ foreign language. According to Bayyurt, this provides positive attitudes towards a second language learning. I situate my study within sociolinguistic, which emphasised the role of social and cultural contexts in learning a second language.

Second Language Learning Models and Translanguaging

Researchers suggested different models of teaching children English as a second language in early years. They based their models either on language separation (e.g., by time, activity and teacher) or language integration. In either case, the goal was to improve children's second language learning in early years. Prošić-Santovac and Radović (2018) examined the language separation model (one teacher-one language) applied in a Serbian- English bilingual kindergarten. In this model, instructions were given in both Serbian and English. The authors found that the applied model had advantages and disadvantages. Their results showed that language separation during instruction had positive affect on children's receptive language skills. But expressive skills (e.g., communication) were not improved as much as those of receptive. Likewise, Lugossy (20018) explored immersion the use of the immersion model in teaching English as a second language in two private preschool settings in Hungarian. The author observed that the English language teachers were in the classrooms during different times of the day. For example, one of the teachers was available during mornings, and the other one was in the classroom all day. The author also observed children's language use that English was mainly used during mealtime (e.g., breakfast and lunch) (Lugossy, 20018).

In the immersion bilingual education model, children learn the second language that can be socially dominant and prestigious. This bilingual education model was first applied in Canada (Baker, 2007). In addition, Bayyurt (2012) suggested a content-based instruction for learning English as a second language. In this model, concepts first are introduced to children in their first language (e.g., Turkish), and then a week later these concepts are introduced in English. In other words, learning concepts in English follow learning them in Turkish. It could be argued that learning English is a repetition of what they have learned in Turkish. In the content-based model, learning English is not considered a situated activity (e.g., second language learning taking place in classroom interactions and is used for different purposes). Instead, English is learned through repetition and translation of content and concepts from Turkish.

Schwartz and Asli (2014) criticised these models for keeping language discrete and separate. In other words, these models do not allow children to use languages flexibly. The authors suggested that flexible

language use supported children bilingualism, not ‘double monolingualism’ in classrooms (Schwartz and Asli, 2014, p.22). This referred to the concept of translanguaging that considered that children can use their full linguistic resources to maximise their understanding and developing their second language learning. In recent years, scholarship focused on this construct to understand and analyse children’s language use in classrooms’ interactions. Translanguaging allows children to use their multiple discursive practices (García, 2009) and to move between languages (García and Wei, 2014). Researchers suggested that teachers can develop children’s second language learning (e.g., English) in early childhood education (Mifsud and Vella, 2018; Schwartz and Asli, 2014; Ting and Jintang, 2020). A recent study by Ting and Jintang (2020), which examined preschool children’s English language learning in Malaysia, indicated that teachers used translanguaging to develop children’s competence in English by providing cues and supporting children’s expressive skills. It is worth highlighting that there are mainly two types of the use of translanguaging: pedagogy and practice. Translanguaging as pedagogy is mainly supported and practiced by schools (Creese and Blackledge, 2015). But translanguaging as practice is used by individual teachers (Mary and Young, 2017). In other words, translanguaging as practice is not officially supported by the curriculum. The private nursery in which this research was conducted used translanguaging as pedagogy.

Method

This study was conducted between October 2019 and January 2020. Its method was ethnography, which had a qualitative approach to data collection and analysis. Ethnography requires a researcher to observe children in their social settings such as school on a long-term basis (Gregory, 2005). This method allowed the researcher to observe how young children learn a second language (English) in classroom activities. Ethnography was also chosen to document children’s learning as the time progresses. In other words, the goal was to document how children’s English language learning changed during the process. The researcher wanted to observe children’s learning in classrooms activities and interactions with each other and the teacher. In addition, this paper drew on Copland and Creese’s (2015) concept of linguistic ethnography, which linked learning to social contexts and aimed to find how children use language (Copland and Creese, 2015).

In this research, participant observation, and audio and video recordings were used as data collection methods. Audio and video recorders were also used to record children’s interactions with the English language teacher in classroom activities such as speaking and math. I visited the classroom one day per week during the fieldwork, and (each visit lasted around one hour). The researcher observed children’s participation in English activities and how the children respond to the teacher during interactions. The researcher put the phone on the top of a cupboard to record interactions while taking fieldnotes.

Setting and Participants

In Turkey, there are state funded and private nurseries. The state funded nurseries follow a unified curriculum, meaning that every state-funded nursery applies the same programme although teachers can adapt it to their local context in terms of activities. English as a second language can be taught only in a few state-funded nurseries. However, private nurseries can teach young children English as a second language as these preschools have their own programme. The data in this study was drawn from a study investigating children learning English as a second language early years in a city in Turkey. The school had more than 80 children at the time of the study. The participants children were 5 and 6 years old, and from socio-economically advantaged families as the nursery was private and their family paid fees. The majority of the parents were mainly from different cities and appointed by the government in different in state sectors.

In the school, there were four classrooms, of which had two teachers: the English language teacher and preschool teacher. The English teacher (Ayşe) was graduated from English Language and Literature department. She had the postgraduate certificate in education (PGCE). Based on the teacher’s self-reports, she was not trained to teach young children English. But she had in-service training to young children English. Each teacher was responsible for each language in the classroom. The English teacher organised activities for English language. The teachers defined their language model as flexible and activity-based. It was observed that their model can be defined as co-teaching (Schwartz and Gorbatt, 2018). Because both teachers were in the classroom except during music activities, which were given by a different teacher, and they helped each other to develop the children’s learning. This paper focused on interactions taking place during English language activities.

Data Analysis

In this study, the collected data were given codes to make interactions understandable (Gibbs, 2007), and then the codes were put into categories. The purpose of giving a code was to identify what took place in an interaction. For example, “learning the nose” was considered a code to suggest that the children learn the lexical item nose in English. This code was subsumed under the category *expressive skills*. This study used linguistic ethnography that suggested an interpretive approach to and a bottom-up approach to data analysis (Copland and

Creese, 2015). This means that a researcher works from data to theory. As the data were collected and analysed, the researcher searched for and examined the literature on learning English as a second language. In other words, the data and the researcher's participant observations in the classroom guided the researcher to choose the relevant literature.

Results and Discussion

This section will analyse the main categories that were emerged from the data. Three main categories were identified: *developing expressive skills*, *vocabulary learning* and *math learning*. The teacher sometimes nominates children to take on the role of a teacher to practice their learning. In such activities, the children sit on chairs and the child teacher sits on the front of the other children. In the activity below, the teacher nominates Emre to be a child teacher to ask his friends for the names of the body parts showed on the flashcards. The interaction shows how the teacher translanguages to develop the child's expressive skills.

Turkish is *italic*, and English is regular throughout.

Developing expressive skills

Excerpt 1: Learning the nose

- 1 Teacher: Emre, you are going to ask your friends questions.
- 2 Emre: What is this? [holds flashcard showing a nose]
- 3 Children: This is a nose.
- 4 Teacher: Emre, *bir daha sor* (ask one more time) [the other children do not say it loudly]
- 5 Emre: What is this?
- 6 Children: This is a nose [loudly] (Video recording, 25/11/2019)

In this excerpt, the teacher guides Emre to practise expressive skills in English (Turn 1). Emre, asks his friends for answering what is shown on the flash card (Turn 2). The other children respond to him by saying that is a nose (Turn 3), but it seems that the teacher is not satisfied with their response as she asks them for saying loudly quietly (Turn 4). The teacher encourages Emre to ask for his friend again by saying "*bir daha sor*" in Turkish (Turn 4). In other words, the teacher uses Turkish to guide Emre for repeating what he has said. He asks his friends, and they respond to him with a higher tone (Turn 6). This excerpt shows that the teacher uses the languages flexibly in order to develop children's expressive skills. The children are emerging bilingual. Because they are in the early stage of learning English as a second language. This interaction shows that the children's expressive language skills develop as they take on the role of the teacher (Gregory, 2001).

In the next interaction, the teacher asks the children where they come from. The purpose of examining this activity is to show how the teacher develops the children's English language learning. In this interaction, the teacher nominates each child to say their hometown.

Fieldnotes 2: Defining their hometown

- 1 Teacher: Where are you from, Can?
- 2 Can: I am from Ankara.
- 3 Teacher: Where are you from, Ayla?
- 4 Ayla: ... (silent)
- 5 Teacher: *Söyle, nereli olduğunu.* (say, where you are from)
- 6 Ayla: I am from Kastamonu (Fieldnotes, 01/11/2019).

In this excerpt, the teacher relates the activity to the children's daily life via using both languages. The teacher nominates Can and asks him to say his hometown (Turn 1). Can responds to her with a grammatically sentence (Turn 2). The teacher asks another child to say where she comes from (Turn 3). Ayla does not respond to the teacher (Turn 4). It seems that Ayla does not understand the question in English. Because Ayla responds the question in Turkish (Turn 6). when the teacher asks her the question in Turkish (Turn 5).

The teacher supports the child's language learning by using Turkish. This is an example of using the languages flexibly in order to develop children's English learning. In this interaction the children contextualise their second language learning by saying the names of their hometown in English. In contrast to the strict language separation model, the participant teacher follows and applies a language learning model that enables her to help the children where necessary. Through translanguaging in this activity, the teacher scaffolds the child to practise her new language (Wood et al., 1976).

In the activity below, the teacher introduces the children to occupations by using flashcards. The following excerpt indicates how the teacher supports children's vocabulary learning in a second language.

Vocabulary learning

Fieldnotes 3: Learning the names of the occupations

1 Teacher: This is a mechanic (holds the flashcard of a mechanic)

2 Children: Mechanic (only a few children repeat)

3 Teacher: Mechanic, *tamirci* (mechanic) [the teacher says it in both languages]

4 Children: Mechanic [all children] (Fieldnotes, 08/11/2019).

In this interaction, the teacher translanguages to teach the children the mechanic occupation. First, the teacher introduces the children to the occupation by showing the flashcard (Turn 1). The teacher wants the children to repeat after her. Only a few children repeat what the teacher has said (Turn 2). This time, she says the occupation in English and then in Turkish in order to enable the children to produce the vocabulary (Turn 3). All the children repeat after her only in English (Turn 4). In this activity, the teacher develops children's vocabulary learning in English by providing the meaning of mechanic in Turkish. In so doing, the teacher helps the children to understand the meaning of the mechanic occupation. This activity shows that the teacher encourages the children to learn new vocabulary by using both languages at the same time. It seems that the teacher's use of Turkish stimulates children's speaking skills (e.g., vocabulary learning) in English.

Math learning

The figure below was taken from a math activity in which the children developed their math skills. In such activities, as the figure shows, the children are given the instruction in both languages. Figure 1: Learning the geometric shapes.

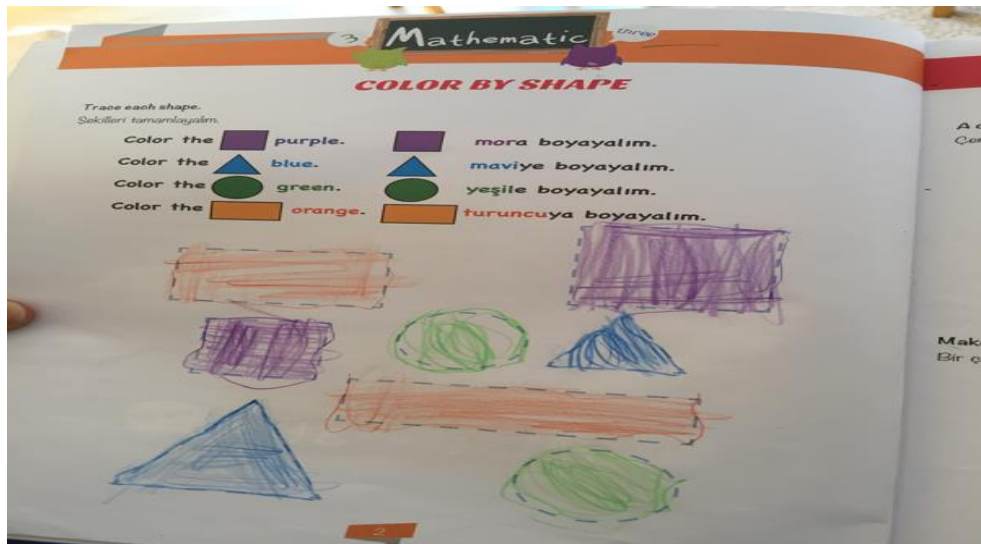


Figure 1. shows the instructions for the names of geometric shapes in Turkish and English.

This is an example that shows that the children's Turkish and English language develop through a flexible model that allows both languages to be used in the activity. In this activity, the children's language skills such as math and art developed by tracing dots and, painting and learning the names of geometrical shapes. In this activity, the children learn the names of geometric shapes and colours in English and Turkish as well. This activity also indicates that young children can learn two languages simultaneously (Kenner, 2004) without separating the languages.

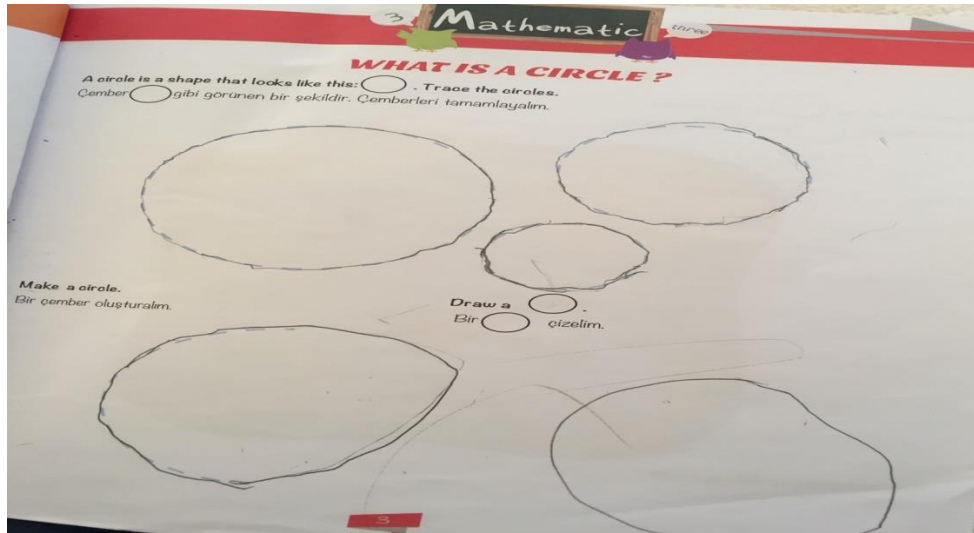


Figure 2. indicates children' learning circles.

For this activity sheet, the instructions are given in both English and Turkish. This activity develops children's math skills in both languages. The children learn the concept of the circle in math. They also learn the lexical items "draw" and "make" in English. They differentiate the lexical items from each other and practice the items by tracing the dots and making a new circle.

Conclusion

This paper has examined children's English language and literacy learning in a preschool setting. It highlighted the role of flexible language use in learning a second language. The findings of this study were vocabulary learning, expressive skills and math skills. This paper showed that the children's second language learning developed through participation in English language activities such as literacy and math. The excerpts showed that vocabulary learning plays an important role in learning a second language. The more the child knows vocabulary, the better s/he can develop her/his expressive skills. The children's expressive skills developed by taking on the role of the teacher. For example, in the excerpt 1 the teacher nominated the children to instruct their friends by taking on her role (Gregory, 2001). This study supported Protassova's (2018) study, which investigated Russian immigrant children learning Finnish in Finland. Protassova used the concept of translanguaging as a theoretical construct to examine children's second language learning in a setting where flexible language teaching model was used. The study demonstrated that the teachers organised activities that improved children's second language learning. She found that bilingualism had a positive effect on children's academic achievement. Their expressive language skills developed through translanguaging in the classrooms' activities. That is, the findings contrasted with Lugossy's (2018) study, which found that language separation by teacher was effective only in children receptive skills.

Strict dual language programmes can constrain language use in classrooms. But translanguaging can enable children to use their linguistic resources (Gort and Sembianti, 2015). In this study, the English language teacher allowed the children to use both languages in order to develop their learning and to familiarise them with literacy learning (e.g., vocabulary). This accorded with Gort and Sembianti' (2015) study, which examined children's language learning in a dual language programme. Gort and Sembianti (2015) found that flexible language use played an important role in children's participation in classroom's activities. The findings supported Schwartz and Palviainen's (2016) study that showed that two languages can be used or learned in combination. As Figure 1 indicated, the children were given instruction in both Turkish and English in a math activity. This enabled them to learn two languages simultaneously (Kenner, 2008). In so doing, the children made sense of their new language through scaffolding from the teacher (Wood et al., 1976).

The children developed their speaking (expressive) and writing skills (see Figure 1) through translanguaging. This study considered that when teachers teach a second language, they can have a holistic approach to teach it. This meant that they do not necessarily focus on vocabulary learning and expressive skills. They can also teach science and math in a second language. This study was at odds with Çetintaş and Yazıcı's (2016) study that examined the preschool teachers and English language teachers' views of teaching English in early years education. Defining immersion method one teacher-one language, Çetintaş and Yazıcı (2016) argued that this had advantages for teaching children English. But their study missed an important point that children's learning is more likely to be limited in the immersion method as children need to want to use one language at one time and the other language at another time. In other words, there can be a strict language use in the

classroom. However, Garcia (2009) pointed out that children can maximise their language skills (e.g., speaking and math) by using their new and first language at same time.

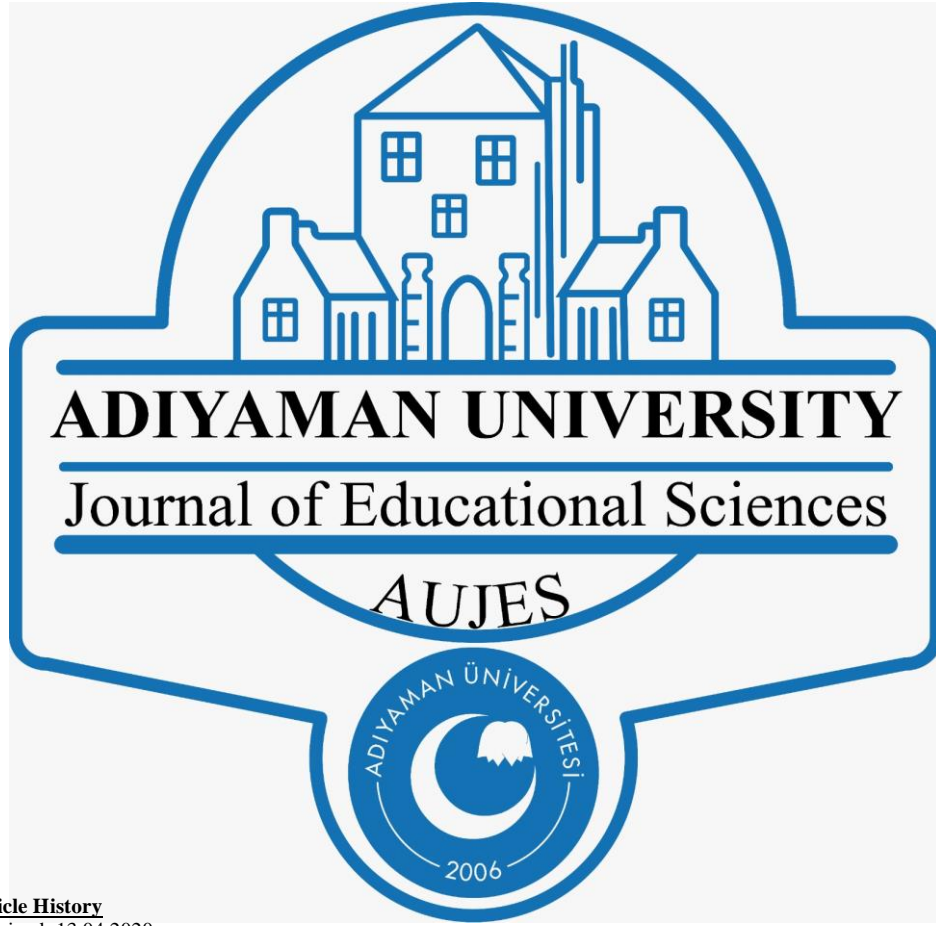
Recommendations

This study suggests that English should be taught along with Turkish in early years education. The study understands that the language of instruction is Turkish in Turkey, but arrangements can be made by policymakers to enable teachers in order to teach children English flexibly. The study had limitations. One of the limitations was that it was conducted in one nursery. Another limitation was the number of participants that could be considered short.

References

- Baker, C. (2007). Becoming bilingual through bilingual education. In P. Auer and L. Wei (eds) *Handbook of Multilingualism and Multilingual Communication* (pp.131-154). Berlin, Walter de Gruyter.
- Baytas, M. O. & Seyma, T. (2019). Translanguaging as a Pedagogical Approach with Syrian Refugee Learners in Turkey: Lessons Learned from a Collaborative Inquiry. AAAL Conference.
- Bayyurt, Y. (2012). Eğitim sisteminde erken yaşta yabancı dil eğitimi [Foreign language education within 4+4+4 education system4+4+4]. 1. Yabancı Dil Eğitimi Çalıştayı [1. Foreign Language Teaching Workshop]. Hacettepe Üniversitesi, Ankara
- Bayyurt, Y. (2013). Current perspectives on sociolinguistics and English language education. *The Journal of Language Teaching and Learning*, 1, 69-78.
- Çetintaş, B.G. & Yazici, Z. (2016). Teachers' opinions concerning bilingual education in early childhood: practice and experience in pre-school and nursery classes. *Mediterranean Journal of Humanities*, 173-187.
- Copland, F. & Creese, A. (2015). *Linguistic ethnography: collecting, analysing and presenting data*. London: SAGE
- Creese, A. & Blackledge, A. (2015). Translanguaging and Identity in Educational Settings. *Annual Review of Applied Linguistics*, 35, 20-35.
- García, O. (2009). Education, multilingualism and translanguaging in the 21st century. In: T. Skutnabb-Kangas, R. Phillipson, A.K. Mohanty and M. Panda (eds.), *Social Justice through Multilingual Education* (pp.140-158). Bristol: Multilingual Matters.
- García, O. & Li Wei. (2014). *Translanguaging: Language, bilingualism and education*. New York: Palgrave Macmillan.
- Gelir, İ. (2020). The investigation of co-teaching model in second language teaching in early years education. *ELT Research Journal*, 9(2), 135-145.
- Gibbs, G.R. (2007). *Analysing qualitative data*. London: SAGE Publications.
- Gort, M. & Sembante, S. F. (2015) Navigating hybridized language learning spaces through translanguaging pedagogy: Dual language preschool teachers' languaging practices in support of emergent bilingual children's performance of academic discourse. *International Multilingual Research Journal*, 9(1), 7-25.
- Gregory, E. (2001) Sisters and brothers as language and literacy teachers: synergy between siblings playing and working together. *Journal of Early Childhood Literacy*, 1(3), 301–322.
- Gregory, E. (2008) *Learning to read in a new language: Making sense of words and worlds*. London: Sage.
- Kenner, C. (2004). Community school pupils reinterpret their knowledge of Chinese and Arabic for primary school peers. In E. Gregory, S. Long and D. Volk (Eds.) *Many pathways to literacy: Young children learning with siblings, grandparents, peers and communities*. (pp.105-116). London: Routledge.
- Lantolf, J. P. & Thorne, S.L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford: Oxford University Press.
- Lantolf, J., Thorne, S. L., & Poehner, M. (2015). Sociocultural Theory and Second Language Development. In B. van Patten & J. Williams (Eds.), *Theories in Second Language Acquisition* (pp. 207-226). New York: Routledge.
- Lugossy, R. (2018). Whose challenge is it? Learners and teachers of English in Hungarian preschool contexts. In M. Schwartz (ed), *Preschool bilingual education: Agency in interactions between children, teachers, and parents* (pp.99-131). Switzerland: Springer
- Mary, L. & Young, A. S. (2017). From Silencing to Translanguaging: Turning the tide to support emergent bilinguals in transition from home to pre-school. In B. Paulsrud, J. Rosén, B. Straszer and Å. Wedin (eds), *New perspectives on translanguaging and education* (pp.108-128). Bristol, UK: Multilingual Matters.
- Mifsud, C. L. & Ann Vella, L. (2018). To mix languages or not? Preschool bilingual education in Malta. In M. Schwartz (ed), *Preschool bilingual education: Agency in interactions between children, teachers, and parents* (pp.57-98). Switzerland: Springer
- Mitchell, R., Myles, F. & Marsden, E. (2013). *Second Language Learning Theories* (3rd ed) London: Routledge.

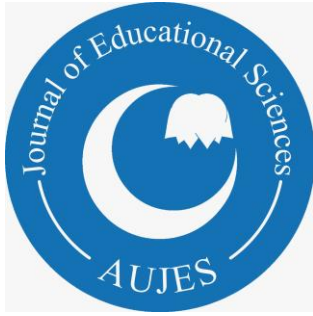
- Official Gazette (Ministry of Education). (2014). Regulations on preschool education and primary education institutions Retrieved December 05, 2020 from <http://dspace.ceid.org.tr/xmlui/bitstream/handle/1/299/ekutuphane3.2.3.6.pdf?sequence=1&isAllowed=y>
- Prošić- Santovac, D. & Radović, D. (2018). Separating the languages in a bilingual preschool: To do or not to do? In M. Schwartz (ed), *Preschool bilingual education: Agency in interactions between children, teachers, and parents* (pp.27-56). Switzerland: Springer
- Sarıca, E. (2019). Erken çocukluk döneminde ikinci dil eğitimi [Learning English as a second language in early years education]. [Unpublished Master's thesis]. Pamukkale Üniversitesi, Denizli.
- Schwartz, M. & Palviainen, Å. (2016). Twenty-first-century preschool bilingual education: facing advantages and challenges in cross-cultural contexts. *International Journal of Bilingual Education and Bilingualism*, 19(6), 603-613.
- Schwartz, M., & Asli, A. (2014). Bilingual teachers' language strategies: The case of an Arabic– Hebrew kindergarten in Israel. *Teaching and Teacher Education*, 38, 22–32.
- Schwarz, M. and Gorbatt, N. (2018). “Fortunately, I found a home here that allows me personal expression”: Co-teaching in the bilingual Hebrew-Arabic-speaking preschool in Israel. *Teaching and Teacher Education*, 71, 46-56.
- Tarone, E. (2007). Sociolinguistic approaches to second language acquisition research 1997–2007. *The Modern Language Journal*, 91 (1), 837-848.
- Ting, S.H. & Jintang, L. (2020). Teachers and students' translanguaging practices in a Malaysian preschool. *International Journal of Early Years Education*, <https://doi.org/10.1080/09669760.2020.1850429>
- Vygotsky, L. S. (1986). *Thought and language* (revised and edited by Alex Kozulin). Cambridge: MIT Press.
- Vygotsky, L.S. (1978) *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA, Harvard University Press.
- Wood, D., Bruner, J., and Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Child Psychiatry*, 17, 89-100.
- Yuvayapan, F. (2019). Translanguaging in EFL classrooms: teachers' perceptions and practices. *Journal of Language and Linguistic Studies*, 15 (2), 678–694.



Article History


Received: 13.04.2020
Received in revised form: 14.04.2021
Accepted: 21.05.2021
Available online: 29.06.2021
Article Type: Research Article

<https://dergipark.org.tr/tr/pub/adyuebd>



ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

**Determination of Cognitive Structures
and Misconceptions of Pre-service
Science Teachers' Regarding the
Concept of "Energy"**

Filiz Avcı¹,
¹Istanbul University, Mathematics and Science Education
Department, Istanbul, Turkey 

To cite this article:

Avcı, F. (2021). Determination of cognitive structures and misconceptions of pre-service science teachers' regarding the concept of "energy". *Adiyaman University Journal of Educational Sciences*, 11(1), 9-25.

Determination of Cognitive Structures and Misconceptions of Pre-service Science Teachers' Regarding the Concept of "Energy"

Filiz Avcı^{1*},

¹Istanbul University, Mathematics and Science Education Department, Istanbul, Turkey

Abstract

The concept of "Energy", which is one of the concepts directly related to vital activities increases in importance day by day with the increasing technology and industrialization. For this reason, the concept of "Energy" learned in primary, secondary and higher education programs has a privileged place in science education as being an interdisciplinary concept. The aim of this study is to reveal the cognitive structures of pre-service science teachers about the concept of "Energy" and to determine their misconceptions. In the study, a survey model was used. The research was carried out with 95 pre-service teachers studying in the 3rd and 4th grade of the faculty of education, science education program of a public university in 2018-2019. The data were collected using the word association test (KIT) and the drawing-writing technique. The data were categorized with the content analysis method. The model explaining the cognitive structures of pre-service science teachers about the concept of "Energy" was prepared with the Vue program. As a result of the word association test about the concept of "Energy", the cognitive structures of the teacher candidates; It has been determined that it focuses on the categories of "Energy types", "Scientific terms evoking the concept of energy" and "Energy sources". In the study, it was determined that pre-service teachers had various misconceptions about the concept of "Energy" in the data obtained from both the word association test and the drawings. It is recommended to create learning environments that will ensure correct and meaningful learning by emphasizing conceptual learning

Key words: Energy, Cognitive structure, Word association test, Drawing-writing technique, Misconception

Introduction

The aim of science education is to raise individuals who can understand and transfer scientific knowledge correctly, think, question, develop problem solving skills, keep up with technological developments, establish a cause and effect relationship, and have a contemporary scientific perspective (Balbağ, Leblebiciler, Karaer, Sarikahya & Erkan, 2016). Providing science education effectively is only possible with the complete and correct learning of the concepts. In this context, it is clear that establishing and revealing cognitive structures by establishing the correct relations between concepts plays a key role in the realization of conceptual learning.

Shavelson (1974) defines cognitive structure as theoretical structures that show the conceptual relationships in learners' long-term memories. Determining cognitive structures is very important in ensuring that science educators understand how learners receive and construct information (Tsai & Huang, 2002).

In order to determine the cognitive structures; Alternative measurement and evaluation tools such as word association test, branched tree, structured grid, concept maps, conceptual change texts, questionnaires, and interviews are used (Bahar, 2003). Among these tools, the most used tools are the word association test (Gussarsky & Gorodetsky, 1990; Işıklı, Taşdere & Göz, 2011; Köseoğlu & Bayır, 2011; Kurt, 2013) and the drawing-writing technique (Çetin, Özarslan, Işık & Eser, 2013; Patrick & Tunnicliffe, 2010; Smith & Metz, 1996). With the word association test, it can be determined whether the relationships between the concepts in the long-term memory are at a sufficient level while revealing the relationships between the cognitive structures of individuals related to concepts and the concepts (Balbağ, 2018; Cardellini & Bahar, 2000; Shavelson, 1974). In this context, misconceptions can also be detected with independent word association tests (Bahar & Özatlı, 2003; Ercan & Taşdere, 2010). Nakiboğlu (2008) stated that using traditional assessment and evaluation methods is not enough to reveal the cognitive structures of learners, and word association tests are a convenient method for revealing the conceptual change of learners as well as determining the reflections of cognitive structures. When the drawing and writing technique is used, it can be provided to reveal the visual images of

*Corresponding Author: Filiz Avcı, filizfen@istanbul.edu.tr

individuals related to their cognitive structures towards concepts (Özden, 2009; Patrick & Tunnicliffe, 2010; Reiss & Tunnicliffe, 2001). For this reason, the study carried out, it was aimed to obtain multidirectional data by using the word association test and the drawing-writing technique.

While learning is taking place, individuals may also embed unscientific concepts in their cognitive structures while constructing scientific and correct concepts. These non-scientific concepts are called "misconceptions" or "alternative concepts" (Nussbaum & Novick, 1982; Skelly & Hall, 1993). Misconceptions are expressed as behaviors resulting from false beliefs and experiences (Baki, 1999), and in a different definition, they are expressed as information that is incompatible with scientific facts and prevents the teaching and learning of concepts that have been proven by scientists (Chi & Roscoe, 2002; Çakır & Yürük, 1999). Studies in the literature show that learners have difficulty in forming their cognitive structures (Stavridou & Solomonidou, 1998; Tsai & Huang, 2002). While forming the cognitive structure, the fact that the information to be learned is abstract is one of the leading factors that affect learning negatively (Ekici & Kurt, 2014). Physics, chemistry, biology, and science courses in science education include not only concrete concepts depending on the nature of each subject but also many abstract concepts. Abstract concepts cause the formation of wrong / incomplete or misconceptions in the cognitive structure of the learners and this situation creates a problem for both educators and learners (Yağbasan & Gülçiçek, 2003).

The concept of "Energy", which is one of the abstract concepts directly related to vital activities; emerges as one of the most important issues and problems that human beings have emphasized from past to present. Today, the increase in technology and industrialization creates a need for more energy use. For this reason, the importance of learning and teaching the subjects related to energy sources, energy types, energy conversion, energy transmission and energy conservation increases more and more. With the increase in environmental problems in recent years, the concept of energy has started to come to the fore in the educational dimension. In the studies carried out for the concept of "Energy" at all education levels; it is seen that thoughts on the concept of energy are taken (Ayaz, Karakaş & Sarıkaya, 2016; Güven & Sülün, 2018; Kurnaz, 2007; Trumper, 1990; Lin & Reping, 2003; Yürümezoğlu, Ayaz & Çökelez, 2009), different learning-teaching methods are used to teach the concept of energy (Aydın & Balım, 2005; Marulcu & Höbek, 2014; Sarıca & Çetin, 2012) and misconceptions are detected (Ayaz, Karakaş & Sarıkaya, 2016; Chabalengula, Sanders & Mumba, 2012; Solomon, 1982). The fact that "energy" can be associated with many different subjects is one of the main factors that make it attractive to work on.

One of the most important features of the concept of "energy" is that it is an interdisciplinary concept besides being an abstract concept. For this reason, it is learned in secondary education and higher education programs starting from primary education and has a privileged place in science education (Kılıç & Cerit Berber, 2018; Sağlam Arslan, 2010; Töman & Odabaşı, 2012). With science education, students have the knowledge that they can apply in their lives by making the basic science concepts concrete (Çoban, Aktamış & Ergin, 2007). While the concept of "Energy" has various forms such as kinetic, potential, electricity, heat, light, chemical, sound and geothermal, the concept of "Energy" is defined by highlighting different forms in different disciplines. "Energy" is expressed as in chemistry lesson; the heat required to break the bonds between atoms while chemical reactions take place and the heat released when bonds are formed (Karaca and Gökten, 2007: p.77) in biology lesson; it is a concept that is needed for living creatures to survive and the sun is the main source (Sağdıç, Bulut, Korkmaz, Börü, Öztürk, & Cavak, 2007: p.38) and in physics lesson; the ability to do business (Trefil & Hazen, 2004). With a different expression, the "Energy" concept; is defined as a quantity with types such as kinetic, potential, electricity and nuclear energy (Şahan & Tekin, 2007).

The concept of "energy" is included as subtopics in subjects belonging to various disciplines. For this reason, it is stated that students construct the concept differently and encounter difficulties in inter-subject association (Ayas et al., 2002). It is stated in the literature that the concept is considered difficult by students from different disciplines and levels due to its abstract nature (Stylianidou, 2002; Opitz, Harms, Neumann, Kowalzik, & Frank, 2015; Yuenyong & Yuenyong, 2007; Yürümezoğlu, et al., 2009). Solomon (1982) found that because the concept of energy is abstract, learners memorize energy concepts without thinking at all. In studies conducted on the concept of "Energy" at primary, secondary and undergraduate levels, it was determined that students have misconceptions about energy sources, energy conversion, energy transmission and energy conservation (Opitz, Blankenstein, & Harms, 2017; Solomon, 1982; Toman, Karatas, and Çimer, 2013; Trumper, 1990; Trumper, 1998). Kruger (1990), as a result of the study of 20 primary school teachers' thoughts on the concept of energy, found that teachers had misconceptions such as "energy is about movement", "energy is stored force", "energy is not conserved". Köse, Bağ, Sürücü and Uçak (2006) as a result of the study in which they aimed to determine the misconceptions of science teacher candidates regarding energy and energy resources, found that plants and animals have misconceptions about where they get energy. In the same study, it was stated that most of the pre-service science teachers focused on the concept of energy in physics. Ünal Çoban Aktamış and Ergin (2007) emphasized in their study that there are difficulties in understanding the concept of

“Energy” because it is an interdisciplinary subject and that it should be handled with its physical, chemical and biological dimensions in order to overcome difficulties.

When the studies in the literature are examined; It is seen that there are studies in which cognitive structures for the concept of "Energy" are put forward and the level of conceptual understanding is examined by using different methods and techniques at different learning levels. In the studies, the questionnaire consisting mostly of open-ended questions, multiple choice test and interview technique (Duit, 1984; Kruger, 1990; Kurnaz, 2007; Opitz, Harms, Neumann, Kowalzik, & Frank, 2015; Toman & Çimen, 2011; Yıldırım, Önal, & Büyük, 2019; Yürümezoğlu et al., 2009), it was determined that a limited number of word association tests (Çardak, 2009; Uyduran, 2019) were used. A study conducted by Güven and Sülün (2018) with pre-service teachers in which the word association test and drawing and writing technique were used was found. However, in the literature review, there is no study in which the word association test and the drawing and writing technique for the concept of "energy" were used together and misconceptions were detected. In the literature, it is seen that there are misconceptions in students at all levels regarding the concept of "Energy". It is also very difficult to reveal the cognitive structures that will enable us to determine what misconceptions are. In this context, determining the opinions of individuals on the concepts is one of the prominent issues. With the word association test, these thoughts can be explained with words and can be explained visually with the drawing and writing technique. Considering that there are students who express their thoughts in different ways; this situation creates an opportunity to express the concepts in a multifaceted way. Thus, versatile data can be obtained to detect errors in students. In addition, students learn the concept of "Energy" in different lessons starting from primary school. "Science lesson" is the first lesson in which students encounter the concept of "Energy". Considering that they will see them in different courses in the future, they should be learned in the most conceptually correct way. For this reason, it is very important to determine whether the pre-service teachers who will teach in the secondary school science course know the concepts correctly. In this context, misconceptions were investigated by using the word association test and the drawing and writing technique together in the study. It is thought that the results of the study will contribute to the relevant literature by providing a different perspective.

The aim of this study is to determine the cognitive structures and misconceptions of pre-service science teachers about the concept of "Energy" by using the word association test and the drawing-writing technique.

In line with this main purpose, the following questions were sought:

- What are the cognitive structures of pre-service science teachers towards the concept of "Energy"?
- What are the misconceptions of pre-service science teachers about the concept of "Energy"?

Method

Research Model

A Survey model was used in this research. A Survey model is a research approach that aims to describe a situation that has existed in the past or still exists (Karasar, 1999). In the study, the data about the cognitive structures of pre-service science teachers for the concept of "Energy" was investigated in detail.

Research Group

The research was carried out with third and fourth grade students studying in the Science Teaching Department of a state university in the Marmara Region in the 2018-2019 academic year. The selection of the study group was influenced by the fact that the students studying in these classes had taken "General Chemistry, General Physics and General Biology" courses in previous years and were close to the teaching profession. This situation is suitable for purposeful sample selection. The purposeful sampling allows for the detailed investigation of situations that are thought to have versatile information (Yıldırım & Şimşek, 2016). In this context, 95 volunteer students participated in the research.

Data Collection Tools

Word association test (WAT) and drawing-writing technique were used in order to reveal the cognitive structures of pre-service teachers regarding the concept of "Energy" and to determine their misconceptions. Before starting the application, explanations were made to the students about the WAT and the application of the drawing-writing technique, and application examples were given.

Word Association Test

Word association test consists of two parts. In the first part, the concept of "Energy" has been included, leaving gaps for students to write what they think about the concept. In the word association test, pre-service teachers should write the first ten words brought to mind by the main concept within 40 seconds (Gussarsky &

Gorodetsky, 1990). For this reason, pre-service teachers were given 40 seconds. The main concept is written on a single page. The reason for this situation; to prevent the possibility of chain response. If the pre-service teacher does not return to the main concept in every word writing, he can write the words that he wrote as an answer instead of the main concept (Bahar & Özatlı, 2003).

In the second part, pre-service teachers were asked to write a sentence about the concept of "Energy" within 20 seconds. The reason for this is the possibility that the words related to the main concept can only be at the recall level or words that do not have a meaningful relationship with the main concept (Nartgün, 2006). In addition, since the "related sentence" written by pre-service teachers will have a more complex and higher level structure compared to a single word, whether the sentence is scientific or not and whether it contains misconceptions will affect the evaluation (Ercan & Taşdere, 2010).

The WAT for the concept of "Energy" is organized as follows:

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Energy

Related Sentence

Data Analysis

The data collected with the measurement tool were analyzed by content analysis. Content analysis; It is a scientific method that enables objective and systematic evaluation of verbal, written and other materials (Tavşancıl & Aslan, 2001). With content analysis, similar data are organized by gathering under certain concepts and categories (Lichtman, 2010). Content analysis; consists of processing data and extracting codes, creating themes, arranging codes and themes by associating them, defining the findings and making comments (Yıldırım & Şimşek, 2016). All the data of the study were analyzed following these steps.

In the word association test, a frequency table expressing the frequency of repetition of the concepts used by pre-service science teachers concerning the concept of "Energy" was prepared. The words that were repeated only once by the participating pre-service science teachers and not related to other words and the subject were not taken into consideration (Kostova & Radoynovska, 2008; Kurt, 2013). The reason for this; It is difficult to create a concept network consisting of meaningful categories from these words used in large numbers. In the analysis process, words were categorized using semantic relationship criteria. There are many studies in the literature showing that this type of data analysis technique gives reliable results (Kostova & Radoynovska, 2008; Kurt, 2013). While analyzing the data, the model of the pre-service science teachers' cognitive structures related to the concept of "Energy" was prepared with the Vue program.

The sentences expressed by pre-service science teachers related to the concept of "Energy" were categorized as sentences containing scientific information, sentences containing non-scientific and superficial information, and sentences containing misconceptions, using the table developed by Ercan, Taşdere, and Ercan (2010).

If the related sentence is scientifically correct and related to the concept, the sentence containing scientific information; If it is not scientific, reflects feelings and thoughts used in daily life, unscientific and superficial sentence; If expressing the concepts using different and incorrect expressions, it is determined as a sentence containing misconception.

Validity and Reliability Study

Validity of research results; coding of data, detailed explanation of the analysis process (how to reach the conceptual category) (Hruschka, Schwartz, St. John, Picone-Decaro, Jenkins & Carey, 2004) and it was provided by giving the opinions of the pre-service science teachers in the findings section. After the data of the

study were coded separately by two science education experts, the code and category list were finalized. Reliability of data analysis; It was calculated using the formula $[\text{Agreement} / (\text{Agreement} + \text{Disagreement}) \times 100]$ (Miles & Huberman, 1994). The average reliability between coders for the research performed was found to be 90%.

Results

In this part of the research, the findings obtained by analyzing the data collected by the word association test and the drawing and writing technique are included.

Findings Obtained According to the Word Association Test Data

As a result of the analysis of the data showing the cognitive structures of the pre-service science teachers regarding the concept of "Energy", it was determined that they associated 80 concepts with the main concept of "Energy". These 80 concepts have been repeated 936 times in total. Among all concepts, 67 concepts that have been repeated 881 times were collected in 7 main categories. The category with the highest frequency among the specified categories is the category of "Energy types". This category is followed by "Scientific terms evoking the concept of energy", "Energy sources", "Situations that provides energy formation", "Properties of energy", "Contribution of energy to daily life" and "Affective effect of energy" categories. If these words are repeated once, if they are not meaningful and not related to the concept, they are not combined with other answer words (Kostova & Radoynovska, 2008; Kurt, 2013). For this reason, 17% (13 words) of the words given as answers were not included in the categories. Charge (7), universe (7), vibration (6), Chernobyl (6), particle (6), brain (5), effect (4), formula (4), acceleration (4), organism (2), explosion (2), reaction (1) and cell (1) the concepts that do not fall into any category were repeated 55 times in total. In this context, the distribution of the cognitive structures of the pre-service science teachers obtained by the word association test on the concept of "Energy" by categories is given in Table 1.

Table 1. The distribution of the cognitive structures of the pre-service science teachers obtained by the Word Association test on the Concept of "Energy" by categories

Categories	Included in the Categories Concepts and Frequencies	Total Frequencies
Energy types	Potential Energy(58)	317
	Kinetic Energy (55)	
	Heat Energy(37)	
	Mechanical Energy(35)	
	Motion Energy (32)	
	Electrical Energy (31)	
	Nuclear Energy (20)	
	Chemical Energy (20)	
	Light energy (20)	
	Sound Energy(5)	
Scientific terms evoking the concept of energy	Geothermal Energy(4)	208
	Power (22)	
	Joule(21)	
	Mass (19)	
	Speed (15)	
	Work (15)	
	Temperature (13)	
	Calori (10)	
	Atom(10)	
	Force (10)	
	Einstein (6)	
	Energy pyramid (6)	
	Relativity (6)	
	Matter (5)	
	Newton(5)	
Tesla(5)		
Essence (5)		
$E=mc^2$ (5)		
Science (5)		
Thermodynamics (5)		

	Enthalpy (5) Entropy (4) Exothermic reaction (4) Endothermic reaction (4) Quantum(3)	
Energy sources	Sun (45) Renewable energy sources (24) Wind (21) Non-renewable energy sources (10) Battery (10) Water (5) Molecule breakdown (5)	120
Situation that provides energy formation	Eating (25) Collision of subatomic particles (13) Human body (10) Power Plants (8) Mitochondria (6) Friction (6) Breathing (5) Eating chocolate (5) Fission-fusion (4)	82
Properties of energy	Conservation (27) Transformation (20) Produced (15) Consumable (5) Savings can be achieved (5) Shopping (5)	77
Contribution of energy to daily life	Allows to live (19) Physical activity (13) Technology (5) Photosynthesis (4)	41
Affective effect of energy	Positive energy(13) Negative energy (5) Love (5) Metaphysics (5) Dance pleasure (4) Coffee enjoyment (4)	36
	67	881

As seen in Table 1, according to the answers given by pre-service science teachers to the word association test, "Energy types" for the main concept of "Energy" was determined as the first category (f = 317). Under this category, 11 words were determined. The first 6 words that are frequently repeated among these words are "potential energy", "kinetic energy", "heat energy", "mechanical energy", "motion energy" and "electrical energy". It was determined that other pre-service science teachers focused on the concepts of "nuclear energy", "chemical energy", "light energy", "sound" and "geothermal energy".

According to the answers given by pre-service science teachers to the word association test, "Scientific terms evoking the concept of energy" for the main concept of "Energy" was determined as the second category (f = 208). Under this category, 24 words were determined. The first 6 words that are most frequently repeated among these words are "power", "joule", "mass", "speed", "work", "temperature". It has been determined that other preservice teachers focus on the concepts of "calori", "atom", "force", "Einstein", "energy pyramid", "relativity", "matter", "Newton", "Tesla", "essence", "E=mc²", "science", "thermodynamics", "enthalpy", "entropy", "exothermic reaction", "endothermic reaction" and "quantum".

According to the answers given by the pre-service science teachers to the word association test, for the main concept of "Energy", "energy sources" was determined as the third category (f = 120). Under this category, 7 words were determined. The first 4 words that are most frequently repeated among these words are "Sun",

"renewable energy sources", "wind" and "non-renewable energy sources". It was determined that other pre-service science teachers focused on the concepts of "battery", "water" and "molecule breakdown".

According to the answers given by pre-service science teachers to the word association test, for the main concept of "Energy", "Situations that provides energy formation" was determined as the fourth category ($f = 82$). Under this category, 9 words were determined. The first 3 words that are most frequently repeated among these words are "eating", "collision of subatomic particles" and "human body". It was determined that other pre-service science teachers focused on the concepts of "power plants", "mitochondria", "friction", "breathing", "eating chocolate" and "fission-fusion".

According to the answers given by the pre-service science teachers to the word association test, for the main concept of "Energy", the "Properties of energy" was determined as the fifth category ($f = 77$). Under this category, 6 words were determined. The first 3 words that are most frequently repeated among these words are "conservation", "transformation" and "produced". It was determined that other pre-service science teachers focused on the concepts of "consumable", "savings can be achieved" and "shopping".

According to the answers given by pre-service science teachers to the word association test, for the main concept of "Energy", "Contribution of energy to daily life" was determined as the sixth category ($f = 41$). Under this category, 4 words were determined. The word that is most frequently repeated among these words is the word "allows to live". It was determined that other pre-service science teachers focused on the concepts of "physical activity", "technology" and "photosynthesis".

According to the answers given by pre-service science teachers to the word association test, for the main concept of "Energy", "Affective effect of energy" was determined as the seventh category ($f = 36$). Under this category, 6 words were determined. The most frequently repeated of these words is the word "positive energy". It was determined that other pre-service science teachers focused on the concepts of "negative energy", "love", "metaphysics", "dance pleasure", "coffee enjoyment".

Findings Regarding the Sentences That Pre-Service Science Teachers Made About the Concept of "Energy"

In order to reveal the knowledge structures of the pre-service science teachers about the concept of "Energy" in detail, the sentences expressed concerning the concept of "Energy" were analyzed depending on their relationship with the concept and divided into categories according to their meanings. At this stage, while it was determined that some of the pre-service science teachers did not write sentences, most of the pre-service science teachers have defined the concept of energy and its types, explained the properties of energy, emphasized energy sources, and explained its usage areas in daily life.

When the sentences are evaluated under the specified categories;

The pre-service science teachers', according to the category of "Energy types"; it has been determined that have misconceptions such as "There are 3 types of energy: potential, kinetic and mechanical".

The pre-service science teachers', according to the category of "Scientific terms evoking the concept of energy"; has been determined that they have statements that contain scientific information such as "The unit of energy is Joule." and they have misconceptions such as "It is the power required for an object to do work.", "It is the force that exists in matter in the universe." and "It is what is necessary for living things to move".

The pre-service science teachers', according to the category of "Energy resources"; it has been determined that they have statements that contain scientific information such as "The sun is the most important source of energy.", "There are renewable and non-renewable energy sources." and "Wind is a renewable energy source."

The pre-service science teachers', according to the category of "Situations that provides energy formation"; it has been determined that they have statements that contain scientific information such as "As a result of breathing with oxygen, 38 ATP energy is obtained."

The pre-service science teachers', according to the category of "Properties of energy"; it has been determined that they have statements that contain scientific information such as "Energy cannot be created out of nothing and the existing energy cannot be destroyed.", "Energy can be converted." and "Energy is always conserved in the universe." and besides this, they have misconceptions such as "The energy is running out." and "Only living things have energy."

The pre-service science teachers', according to the category of "Contribution of energy to daily life"; it has been determined that they have statements that contain unscientific and superficial sentences such as "Energy provides living."

The pre-service science teachers', according to the category of "Contribution of energy to daily life"; it has been determined that they have statements that contain unscientific and superficial sentences such as "Energy provides living."

The pre-service science teachers', according to the category of "Affective effect of energy"; it has been determined that they have statements that contain unscientific and superficial sentence such as "Life is beautiful thanks to the positive energies.", "I have no energy today." and "Energy is enjoying coffee."

In this context, it can be said that pre-service science teachers have scientific knowledge about the concept of "Energy" but at the same time have quite a number of misconceptions. As a result of the analysis of the obtained data, a model was created that visualizes the cognitive structures of pre-service science teachers towards the concept of "Energy" (Figure 1).

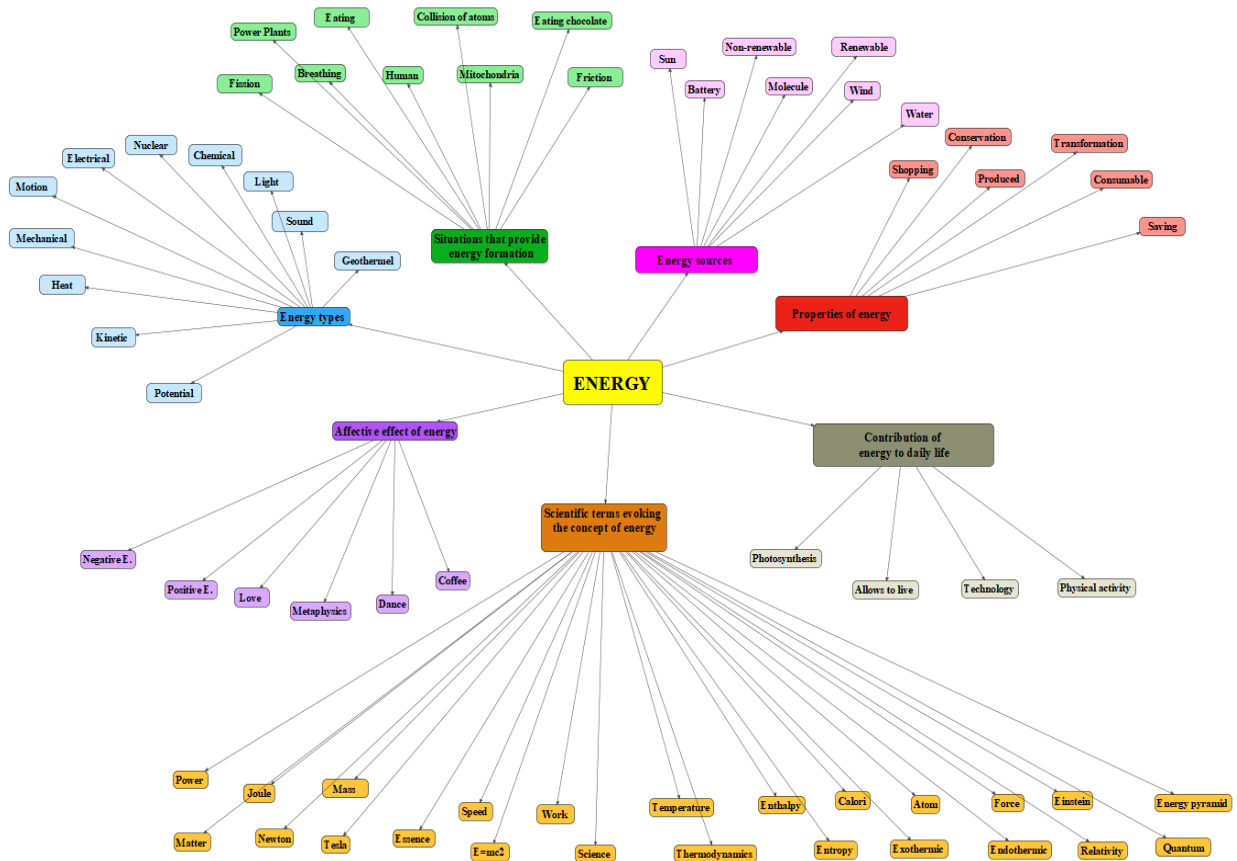


Figure 1. The cognitive structure of pre-service science teachers determined by the word association test on the concept of "Energy"

As seen in Figure 1, the conceptual structures of pre-service science teachers related to the concept of "Energy" have emerged concerning 7 categories.

Findings Obtained According to the Drawing-Writing Technique Data

The conceptual structures created by the drawing-writing technique that support the cognitive model (Figure 1) associated with the concept of "Energy" of pre-service science teachers are shown in Figure 2. It was determined that the conceptual structures of the pre-service science teachers, for the concept of "Energy" are related to three categories.

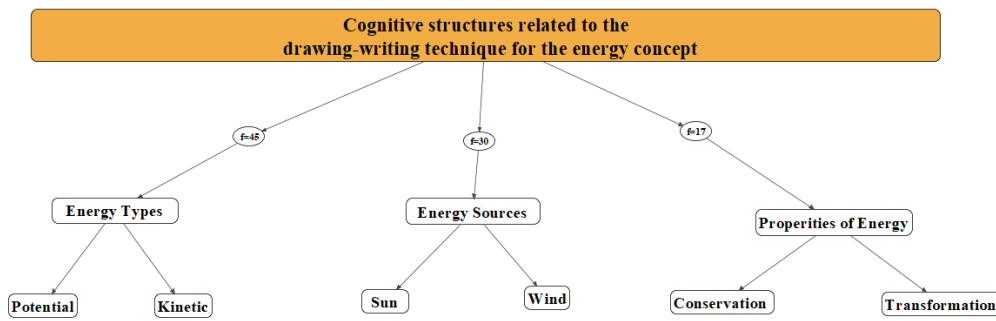


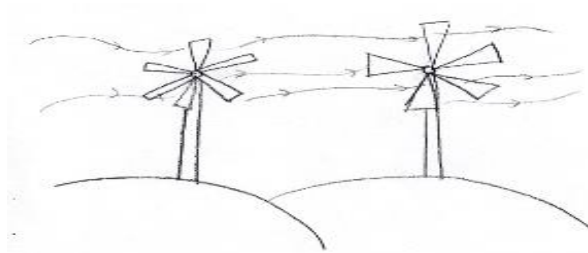
Figure 2. cognitive structure of pre-service science teachers determined by the drawing-writing technique related to the concept of "Energy"

The drawings of the pre-service science teachers on the concept of "Energy" were collected under 3 categories. These categories are; "Energy types (45)", "Energy sources (30)" and "Energy properties (17)". Examples of the drawings of the pre-service science teachers, related to the concept of "Energy", their categories and frequency values are shown in Table 2.

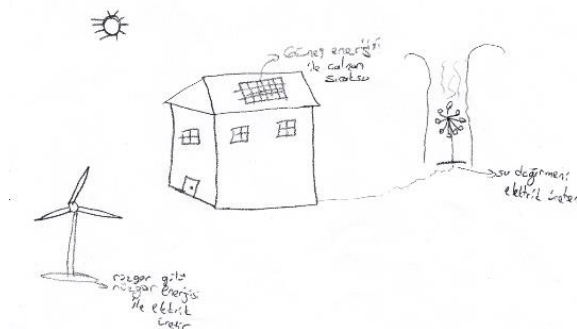
Table 2. Distribution of preservice science teachers' cognitive structure determined by the drawing-writing technique related to the concept of "Energy" and drawing samples

Categories	Drawing concep	Drawing samples	f
Energy types	Potential Energy		445
	Kinetic Energy		
Energy sources	Sun		30

Wind

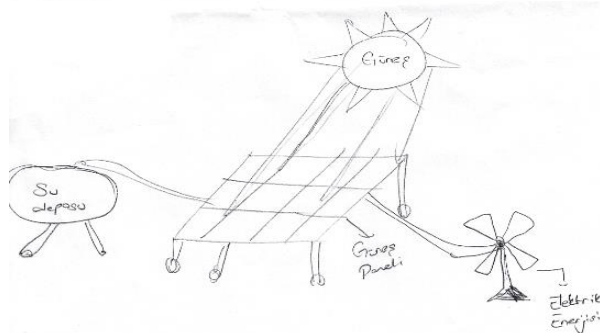


Conservation



Properties of energy

Transformation



17

From the drawings of the pre-service science teachers, about the concept of "Energy"; it has been determined that they drew heavily thinking about energy types and in such a way that the concepts of potential and kinetic energy are dominant. These are followed by drawings related to energy resources, primarily wind and solar energy and it has been observed that they express the properties of energy at a low rate in a way to relate the conservation and transformation of energy.

Discussion-Conclusion

In this study, which aims to reveal the cognitive structures of pre-service science teachers related to the concept of "Energy" through word association test and drawing-writing technique, and to determine their misconceptions, multifaceted data supporting each other was obtained. In this context, the data obtained through the word association test are collected in 7 categories in total ("Types of energy", "Scientific terms evoking the concept of energy", "Energy sources", "Situations that provides energy generation", "Properties of energy", "Contribution of energy to daily life", "Affective effect of energy"), 3 categories were determined with the drawing and writing technique ("Types of energy", "Energy sources", "Properties of energy"). When the categories were examined, it was determined that there were more categories obtained by the word association test. It has been determined that the categories of "Scientific terms evoking the concept of energy", "Situations that provides energy formation", "Contribution of energy to daily life" and "Affective effect of energy" did not appear in the drawing-writing technique. In the light of the answers received as a result of the application of both techniques; it has been determined that the pre-service science teachers, frequently concentrate on potential and kinetic energy types among energy types, on the sun, renewable energy sources, non-renewable energy

sources and wind among energy sources, on conservation, transformation and production concepts regarding the properties of energy.

The obtained results show once again the importance of using various measurement tools in academic studies in a way that supports each other in order to obtain qualified data. In the study, while the students explained the word association test for the concept of "Energy" in writing, they also had the opportunity to express their thoughts by drawing in the drawing and writing test. Although the results support each other, the fact that some concepts that are not obtained in one technique can be obtained with another technique shows the richness of the data obtained from the study.

The pre-service science teachers' from the data obtained using the word association test and drawings about the concept of "Energy"; it has been determined that they have misconceptions and incomplete information such as "There are 3 types of energy, potential, kinetic and mechanical.", "It is the power required for an object to do work.", "It is the force that exists in the matter of the universe.", "The energy is running out." and "Only living things have energy.". In parallel with these results obtained in the study, it was found that there were similar misconceptions in the results of the studies on the concept of "Energy" carried out with pre-service science teachers and students had difficulties in understanding the concept of energy (Kruger, 1990; Köse et al., 2006; Kurnaz, 2007; Töman and Çimen, 2011; Trumper, 1996; Trumper, 1998). It can be thought that the reason for this is that the concept is abstract and therefore they cannot form the different sub-concepts related to the concept in their mind in a holistic way or they cannot learn the subjects completely and correctly.

It was determined that pre-service science teachers focused on the concepts of potential energy and kinetic energy under the category of "Energy Types". It is seen that the number of students talking about other types of energy such as heat, mechanics, motion, and electricity is less. This result shows that students do not have enough information about energy types. Similar results were obtained when the drawings of the pre-service science teachers were examined. Parallel to this result obtained from the research, Kurnaz (2007) conducted a study conducted to determine the teaching and learning situations of the concept of "Energy" at the first grade of the university; he found that most of the students do not know the types of energy and that the meaningful relationship between kinetic and potential energy cannot be established with mechanical energy. It can be thought that the reason for this result may be because the pre-service science teachers learned the concept of energy with the contents containing a high level of theoretical knowledge in different courses and because of the limited lecture environments where they can learn by doing different types of energy. When the sentences written by the pre-service science teachers are examined, regarding the category of "Energy types"; "There are 3 types of energy: potential, kinetic and mechanical."; it has been determined that they have misconceptions in the form. Similar to the results of the study, Gülçiçek and Yağbasan (2004) concluded that the students did not know that the mechanical energy value of the system was the sum of the kinetic and potential energy values.

Under the category of "Scientific terms evoking the concept of energy", the pre-service teachers primarily emphasized the concepts of power, joule, speed, work, temperature, mass, calorie, atom and force. This result is an indication that pre-service teachers have information about the unit of energy, but they have confusion about concepts regarding the definition of energy. When the drawings of the pre-service science teachers were examined, it was found that no drawing was made in this category. When the sentences written by the pre-service science teachers are examined, in parallel with the results obtained from the concepts in the word association test, the "Unit of energy is Joule." while it was determined that they could write expressions containing scientific information in the form of however, it has been determined that they have misconceptions such as "Energy is the power required for an object to do work.", "Energy is the force that exists in matter in the universe." and "Energy is what is necessary for living things to move.". The reason for this situation is thought to be because the differences between the concept of "Energy" and concepts (such as power, force) at primary education, secondary education and even undergraduate level cannot be fully distinguished in the teaching process. The fact that these concepts are used interchangeably in daily life is an indicator of this. The findings showing that the pre-service science teachers confused the concept of energy with the concept of force and power, obtained in the study, support the findings obtained by different researchers (Kruger, 1990; Trumper, 1998; Tomanve Çimer, 2011; Watts; 1983). Similarly, the result that the science teacher candidates obtained in the study associated the concept of "energy" with movement was also found in Watts's (1983) study. Also, in a different study, it was found that there was a misconception among pre-service science teachers that "People gain energy by moving" (Toman & Çimer, 2011). Trumper (1998) at the end of a 4-year study investigating the perceptions of pre-service teachers on the concept of "Energy"; it was determined that pre-service teachers' perceptions increased, but they had misconceptions about energy. Trumper found that pre-service teachers perceive energy as a tangible entity and confuse the concepts of energy and force. Bayram, Şahin, and Gürdal (1999), in their study with pre-service physics, chemistry and biology teachers who were preparing for primary education, found that the teacher candidates could not establish an inter-conceptual relationship on "Energy".

The pre-service science teachers frequently focused on the concepts of sun, renewable energy sources, wind, non-renewable energy sources under the category of "Energy Resources". Concepts such as non-renewable energy resources and water have been expressed very little. It is noteworthy that there is no mention of many different energy sources such as wave, hydrogen, hydroelectric, biomass, geothermal energy. This result shows that pre-service teachers do not have a sufficient level of perception on this issue. It has been determined that the concepts of wind and sun are predominantly used in the drawings of the teacher candidates. Similar to the results obtained in the study, Saraç and Bedir (2014) found that 10 classroom teachers had a lack of knowledge about renewable energy sources as a result of their research conducted to determine the perceptions of classroom teachers about renewable energy sources. In the same study, it was determined that some of the teachers confuse renewable and non-renewable energy sources. As a result of the study, it was concluded that there is a need for educational trips, materials and seminars related to the teaching of energy resources. Yıldırım et al. (2019), in their study with 8th grade students, found that although there are different energy sources, the students do not include other sources than Sun, wind and water in their drawings about these sources. This situation is an indication that the knowledge about energy resources is not at a sufficient level in secondary school and as seen at the end of the study, there is not much change at the undergraduate level. One might think that the reason for this is that the lessons are taught with the emphasis on the concepts of wind and Sun in all education level lessons on "Energy resources". When the sentences written by the pre-service science teachers are examined, regarding the "Energy resources" category; "The Sun is the most important energy source.", "There are renewable and non-renewable energy sources." and "Wind is a renewable energy source."; it was determined that they made sentences containing scientific information and did not make any sentences showing that they had misconceptions. This result shows that the pre-service science teachers have limited knowledge about the subject, but they do not have any errors.

It was determined that pre-service science teachers primarily focused on the concepts of eating, the collision of subatomic particles, breathing and friction under the category of "Situations that provides energy formation". When the drawings were examined, no drawing was found for this category. When the sentences written by pre-service science teachers are examined, "38 ATP energy is obtained as a result of breathing with oxygen."; it was determined that they made sentences containing scientific information in the form of. This result shows that the pre-service teachers have limited knowledge about the subject, but they do not have any errors. It was determined that the pre-service teachers focused primarily on the concepts of conservation, transformation and producible under the category of "Properties of energy". It is seen that a smaller number of students have expressed the concepts of consuming, saving and shopping. This situation shows that the pre-service teachers have ideas about some properties of energy but they have misconceptions about some of its properties. When the drawings were examined, it was determined that they made drawings related to the concepts of conservation and transformation. The findings obtained support each other. When the sentences written by the pre-service science teachers are examined, in parallel with the results obtained from the concepts in the word association test, it has been determined that they wrote statements containing scientific information such as "Energy cannot be created from nothing and existing energy cannot be destroyed.", "Energy is converted" and "Energy is always preserved in the universe" and besides this, it has been determined that they have misconceptions in the form such as "Energy is exhausted." and "Only living things have the energy.". It is thought that the reason for this may be that the pre-service teachers did not learn the subject of energy in learning environments where they can see that energy can be transformed, not exhausted, and can make experimental practices. This situation is parallel to the findings obtained in existing studies (Kruger, 1990; Liu, Ebenezer, & Fraser, 2002; Watts, 1983). Kruger (1990), in a study conducted with 20 elementary school teachers, researched the teachers' understanding of the concept of "Energy". It has been determined that they have misconceptions such as "Energy is about movement", "Energy is consumed", "Energy is not conserved". Watts (1983) found similar misconceptions in the study that "Only living organisms have energy".

It has been determined that pre-service teachers primarily focus on the concepts of physical activity and provide living under the category of "Contribution of energy to daily life". Fewer students used the concepts of technology and photosynthesis. When the drawings and sentences of the preservice teachers were examined, no data related to this category was found. This situation can be considered as an indication that students cannot relate the concepts they learn in lessons with daily life. In parallel with the results obtained, different studies conclude that students experience confusion of concepts between the scientific knowledge they have learned at school and the concept of energy they use in daily life (Toman & Çimer, 2011).

It was determined that pre-service science teachers primarily focused on the concept of positive energy under the category of "Affective effect of energy". It was determined that other pre-service science teachers focused on the concepts of negative energy, love, metaphysics, dance pleasure, and coffee enjoyment. The result shows that the cognitive structures of pre-service teachers for the concept of energy, which is a versatile concept, have concepts that create a considerable affective effect. The preservice teachers did not draw for this category. When the sentences written by the science teacher candidates are examined, parallel to the results

obtained from the concepts in the word association test, It was determined that they express non-scientific sentences containing superficial information such as "Life is beautiful thanks to positive energies.", "I have no energy today." "Energy is enjoying coffee." This situation is an indication that pre-service science teachers cannot correctly associate scientific concepts with their life practices. Similar to the results obtained from the study, it is stated in the studies that students have difficulties in perceiving these concepts and in structuring them by associating them with daily life because some subjects and the concepts related to these subjects are mainly abstract and the theoretical part is mainly taught (Anagün, Ağır & Kaynaş, 2010).

When all categories are examined, it is seen that the students use concepts related to physics subjects intensively about the concept of energy. However, the subject of energy is a common concept in all the subjects of physics, chemistry and biology. It has been determined that the concepts related to biology and chemistry course contents are quite limited in the cognitive structures of pre-service teachers. Köse, Bağ, and Sürücü (2006) obtained similar results in the study. In a different study, Carr and Kirkwood (1988) observed 3 teachers teaching activities related to the concept of "Energy" in biology, chemistry and physics lessons for 3 years in order to investigate teachers' perceptions about the concept of "Energy", and as a result of the research, they found that teachers teach energy limited to their fields. The results obtained show that the concept of "Energy" should be given together with the integration of physics, chemistry and biology at all levels from primary school to higher education.

The results of the study indicate that the cognitive structures of pre-service science teachers related to the concept of "Energy" should be developed more consciously and purposefully. Regarding the correct perception and use of the concept of "Energy", it can be said that the reason for the inadequacies in the cognitive structure of pre-service science teachers is the negativities they have encountered in their education process at primary and secondary education level as well as the negativities they experience during their university education. This situation is very important for prospective teachers who will become teachers. For this reason, it is extremely important to ensure correct and meaningful learning in pre-service teachers by bringing conceptual learning to the forefront at all educational levels starting from primary school. In this context, in order to give students conceptual understanding; it is suggested that primarily the science education curriculum, as well as the curriculum of physics, chemistry and biology courses, which include subjects related to the concept of energy, should be arranged in a way that enables conceptual learning. In addition, it will be effective to make sense of the concept by associating it with other lessons, providing learning environments that enable students to actively participate in the lesson and make experimental applications. As a result, qualified generations are the work of qualified teachers. For this reason, it should not be forgotten that teacher education is an issue that should be emphasized. It is thought that the results obtained from the study will contribute to similar studies in which cognitive structures are analyzed and misconceptions are determined in the future.

References

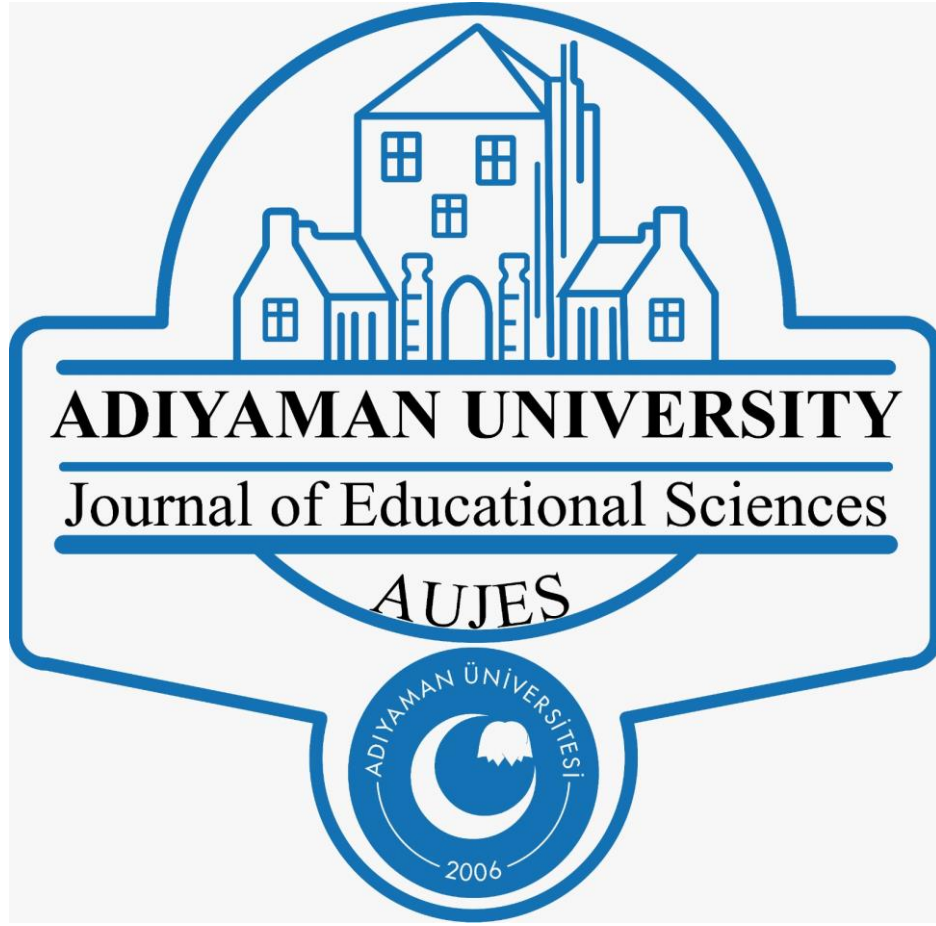
- Anagün, Ş. S., Ağır, O., Kaynaş, E. (2010). İlköğretim öğrencilerinin fen ve teknoloji dersinde öğrendiklerini günlük yaşamlarında kullanım düzeyleri. [Primary school students' usage levels of science and technology course knowledge in their daily lives]. 9. Ulusal Sınıf Öğretmenliği Eğitimi Sempozyumu. Elazığ: Fırat Üniversitesi Eğitim Fakültesi.
- Ayas, A., Karamustafaoğlu, S., Cerrah, L. & Karamustafaoğlu, O. (2002). Fen bilimlerinde öğrencilerdeki kavram anlama seviyelerinde ve yanlışlarını belirleme yöntemleri üzerine bir inceleme.[An investigation on the level of conceptual understanding of students in science and the methods of determining their mistakes]. X. Ulusal Eğitim Bilimleri Kongresi, Abant İzzet Baysal Üniversitesi, Bolu.
- Ayaz, E., Karakaş, H., & Sarıkaya, R. (2016). Sınıf öğretmeni adaylarının nükleer enerji kavramına yönelik düşünceleri: bağımsız kelime ilişkilendirme örneği. [Class teacher candidates' opinions on the concept of nuclear power: the sample of independent word association test]. *Cumhuriyet Üniversitesi Fen-Edebiyat Fakültesi Fen Bilimleri Dergisi*, 37, 42-54.
- Aydın G, Balım A. G. (2005). Yapılandırmacı yaklaşıma göre modellendirilmiş disiplinler arası uygulama: enerji konularının öğretimi. [An interdisciplinary application based on constructivist approach: teaching of energy topics]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 38(2), 145 - 166.
- Bahar, M. (2003). Biyoloji eğitiminde kavram yanlışları ve kavram değişim stratejileri. [Misconceptions in biology education and conceptual change strategies]. *Kuram ve Uygulamada Eğitim Bilimleri*, 3(1), 27-64.
- Bahar, M. & Özatlı, S. (2003). Kelime iletişim testi yöntemi ile lise 1. sınıf öğrencilerinin canlıların temel bileşenleri konusundaki bilişsel yapılarının araştırılması.[Investigation of cognitive structures of first grade high school students about basic components of living things with word communication test method]. *Balıkesir Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, (5), 75-85.

- Baki, A. (1999). Cebirle ilgili işlem yanlışlarının değerlendirilmesi. [Evaluation of processing errors related to algebra]. In *III. Fen Bilimleri Eğitimi Sempozyumu*. M.E.B. ÖYGM (pp: 46–49).
- Balbağ, M. Z. (2018). Fen Bilgisi öğretmen adaylarının hız ve sürat kavramlarına ilişkin bilişsel yapıları: Kelime ilişkilendirme testi (KİT) uygulaması. [Cognitive constructs related to velocity and speed concepts of science teacher candidates: Application of word association test (WAT)]. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*. Issue(33),38-48.
- Balbağ, M. Z., Leblebiciler, K., Karaer, G., Sarıkahya, E., & Erkan, Ö.(2016). Türkiye’de fen eğitimi ve öğretimi sorunları. [Science education and training issues in Turkey]. *Eğitim ve Öğretim Araştırmaları Dergisi*, 5(3), 12-23.
- Bayram, H., Şahin, F., & Gürdal, A. (1999). İlköğretim öğretmen adaylarının enerji konusunda bütünlüğü sağlama ve ilişki kurma düzeyleri üzerine bir araştırma. [A research on primary school teacher candidates' levels of ensuring integrity and relationship in energy]. *Buca Eğitim Fakültesi Dergisi, Özel Sayı 10*.
- Carr, M. & Kirkwood, V. (1988). Teaching and learning about energy in New Zealand secondary school junior science classrooms, *Physics Education*, 23 (2), 86- 91.
- Cardellini, L. & Bahar, M. (2000). Monitoring the learning of chemistry through word association tests. *Australian Chemistry Research Book*, 19, 59- 69.
- Chabalengula, V. M., Sanders, M., & Mumba, F. (2012). Diagnosing students’ understanding of energy and its related concepts in biological context. *International Journal of Science and Mathematics Education*, 10(2), 241-266.
- Chi, M. & Roscoe, R. (2002). The processes and challenges of conceptual change. [Electronic version]. *Behavioral Science*. 2, 3-27.
- Çakır, S.Ö. & Yürük, N. (1999). Oksijenli ve oksijensiz solunum konusunda kavram yanlışları teşhis testinin geliştirilmesi ve uygulanması. [Development and application of misconceptions about oxygenated and oxygen-free breathing]. *III. Fen Bilimleri Eğitimi Sempozyumu*. M.E.B. ÖYGM.
- Çardak, O. (2009). The determination of the knowledge level of science students on energy flow through a word association test. *Energy Education Science and Technology Part B: Social and Educational Studies*, 1(3): 139-155.
- Çetin, G., Özarslan, M., Isık, E., & Eser, H. (2013). Students’ views about health concept by drawing and writing technique. *Energy Education Science and Technology, Part B*, 5 (1), 597-606.
- Çoban, G. Ü., Aktamış, H., & Ergin, Ö. (2007). İlköğretim sekizinci sınıf öğrencilerinin enerjiyle ilgili görüşleri. [The views of 8th grade students about energy]. *Kastamonu Eğitim Dergisi*, 15(1), 175-184.
- Duit, R. (1984). Learning the energy concept in school-empirical results from the Philippines and West Germany. *Physics Education*, 19, 59–66.
- Ekici, G. & Kurt, H. (2014). Öğretmen adaylarının “Aids” kavramı konusundaki bilişsel yapıları: Bağımsız kelime ilişkilendirme testi örneği. [Student teachers’ cognitive structure on the concept of “aids”: the sample of free word association test]. *Türkiye Sosyal Araştırmalar Dergisi*, 3,267-306.
- Ercan, F. & Taşdere, A. (2010). Kelime ilişkilendirme testi aracılığıyla bilişsel yapının ve kavramsal değişimin gözlenmesi. [Observing the cognitive structure and conceptual change through the word association test]. *Türk Fen Eğitimi Dergisi (TUFED)*, 7(2).
- Güven, G & Sülün,Y.(2018). Investigation of the effect of the interdisciplinary instructional approach on pre-service science teachers’ cognitive structure about the concept of energy. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi (EFMED)* 12, (1) 249-281.
- Gussarsky, E., & Gorodetsky, M. (1990). On the concept “chemical equilibrium: the associative framework. *Journal of Research in Science Teaching*, 27 (3), 197204.
- Gülçipek, Ç. & Yağbasan, R. (2004). Basit sarkaç sisteminde mekanik enerjinin korunumu konusunda öğrencilerin kavram yanlışları. [Students’ misconceptions about conservation of mechanical energy in simple pendulum system]. *Gazi Eğitim Fakültesi Dergisi*, 24(3), 23-38.
- Hruschka, D.J., Schwartz, D., St.John, D.C., Picone-Decaro, E., Jenkins, R.A., & Carey, J.W. (2004). Reliability in coding open-ended data: *Lessons learned from HIV behavioral research*. *Field Methods*, 16 (3), 307-331.
- Karaca, G., & Gökten, S.Ö. (2007). *Ortaöğretim kimya 10 ders kitabı*. [Secondary school chemistry 10 textbooks]. Paşa Yayıncılık.
- Karasar, N. (1999). *Bilimsel araştırma yöntemi*. [Scientific research method]. Nobel Yayınları.
- Kaya, E. (2017). Biyoloji öğretmen adaylarının “enzim” konusundaki bilişsel yapıları (Erzurum Örneği). [Cognitive structures of teacher candidates on “enzyme” (Erzurum Example)]. *Ekev Akademi Dergisi*, 21 (72).
- Kılıç, S. & Cerit Berber, N. (2018). Fotoğraf derleme yoluyla kavram haritalama yöntemi kullanılarak öğrencilerin enerji algılarının gündelik yaşam bağlamında araştırılması. [Researching students’

- perception of energy in terms of daily life through the concept mapping method by collecting photographs]. *Millî Eğitim Dergisi*, 218.
- Kostova, Z., & Radoynovska, B. (2008). Word association test for studying conceptual structures of teachers and students. *Bulgarian Journal of Science and Education Policy*, 2 (2), 209-231.
- Köse, S., Bağ, H., Sürücü, A. & Uçak, E. (2006). Prospective science teacher' about energy, *International Journal of Environmental and Science Education*, 1,(2) 141-152.
- Köseoğlu, F. & Bayır, E. (2011). Kelime ilişkilendirme test yöntemiyle kimya öğretmen adaylarının gravimetrik analize ilişkin bilişsel yapılarının incelenmesi. [Investigation of cognitive structures of pre-service chemistry teachers related to gravimetric analysis with word association test method]. *Trakya Üniversitesi Eğitim Fakültesi Dergisi* 1 (1), 107-125.
- Kruger, C. (1990). Some primary teachers' side as about energy. *Physics Education* 25, 86-91.
- Kurnaz, M. A. (2007). Enerji kavramının üniversite 1. sınıf seviyesinde öğrenim durumlarının analizi. [Analysis of the education situation of the energy concept at the first year level of the university]. Yüksek lisans tezi. Karadeniz Teknik Üniversitesi, Eğitim Bilimleri Enstitüsü: Trabzon.
- Kurt, H. (2013). Biology student teachers' cognitive structure about "Living thing". *Educational Research and Reviews*, 8 (12), 871-880.
- Kurt, H., Ekici, G., Aktaş, M. & Aksu, Ö.(2013).Determining biology student teachers cognitive structure on the concept of "diffusion" through the free word association test and the drawing-writing Technique. *International Education Studies*, 6(9),187-206.
- Lichtman, M. (2010). *Qualitative research in education*. Los Angeles: Sage Publications, Inc.
- Lin, Chen-Yung & Reping Hu. (2003) "Students' understanding of energy flow and matter cycling in the context of the food chain, photosynthesis, and respiration." *Int. J. Sci. Educ.* 25.12 (2003): 1529-1544.
- Liu, X., Ebenezer, J. &Fraser, D. M. (2002). Structural characteristics of university engineering students' conceptions of energy, *Journal of Research in Science Teaching*, 39, (5), 423-441.
- Marulcu, İ. & Höbek, K. M. (2014). "8. Sınıflara alternatif enerji kaynaklarının mühendislik dizayn metodu ile öğretimi". [Teaching alternate energy sources to 8 th grades students by engineering design method]. *Middle Eastern and African Journal of Educational Research*, 9, 41- 58.
- Miles, M.B., & Huberman, A. M. (1994). *Qualitative data analysis*. Thousand Oaks, CA: Sage.
- Nakiboglu, C. (2008). Using word associations for assessing non-major science students' knowledge structure before and after general chemistry instruction: the case of atomic structure. *Chemistry Education Research and Practice*, 9 (4), 309-322.
- Nartgün, Z. (2006). Fen ve teknoloji öğretiminde ölçme ve değerlendirme. [Measurement and evaluation in science and technology teaching]. Bahar, M. (Ed), Fen ve Teknoloji Öğretimi. Pegema Yayıncılık.
- Nussbaum, J. & Novick, S. (1982). Alternative frameworks, conceptual conflict and accommodation: Toward a principled teaching strategy. *Instructional science*.
- Nyachwayaa, J.M., Mohameda, A.R., Roehriga, G.H., Woodb, N.B., Kernc, A.L. & Schneiderd, J.L.(2011).The development of an open ended drawing tool: An alternative diagnostic tool for assessing students' understanding of the particulate nature of matter. *Chemistry Education Research and Practice*, 12(2),121-132.
- Opitz, S. T., Harms, U., Neumann, K., Kowalzik, K. & Frank, A. (2015). Students' energy concepts at the transition between primary and secondary school. *Research in Science Education*, 45(5), 691-715.
- Opitz, S. T., Blankenstein, A., & Harms, U. (2017). Student conceptions about energy in biological contexts. *Journal of Biological Education*, 51(4), 427-440.
- Özden, M. (2009). Primary student teachers' ideas of atoms and molecules: *Using drawings*. *Education*, 129(4), 635-642.
- Patrick, P. G. & Tunnicliffe, S. D. (2010). Science teachers' drawings of what is inside the human body. *Journal of Biological Education*, 44(2), 81-87.
- Pluhar, Z. F., Piko, B. F., Kovacs, S. & Uzzoli, A. (2009). Air pollution is bad for my health: Hungarian children's knowledge of the role of environment in health and disease. *Health & Place*, 15, 239-246.
- Reiss, M. J., & Tunnicliffe, S.D. (2001). Students' understandings of human organs and organ systems. *Research in Science Education*, 31, 383-399.
- Roberts, P., & Priest, H. (2006). Reliability and validity in research. *Nursing Standard*, 20, 41-45.
- Sağlam Arslan, A. (2010). Cross-grade comparison of students' understanding of energy concepts. *Journal of Science Educational Technology*, 19, 303-313.
- Sağdıç, D., Bulut, Ö., Korkmaz, S., Börü, S., Öztürk, E., & Cavak, Ş. (2007). *Ortaöğretim 10. sınıf biyoloji*. (2. Baskı) [Secondary school 10th grade biology. (2nd Edition)]. MEB. Yayınları.
- Saraç, E., & Bedir, H. (2014). Sınıf öğretmenlerinin yenilenebilir enerji kaynakları ile ilgili algıları üzerine nitel bir çalışma. [Primary school teachers related to perceptions of renewable energy sources on the qualitative research]. *Kara Harp Okulu Bilim Dergisi*, 24(1), 19-45.

- Sarıca, R. & Çetin, B. (2012). Öğretimde kavram haritaları kullanımının öğrencilerin akademik başarısına ve kalıcılığa etkisi. [The effects of using concept maps on achievement and retention in teaching science lessons]. *İlköğretim Online*, 11(2), 306-318.
- Shavelson, R. J. (1974). Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching*, 11(3), 231-249.
- Skelly, K.M., & Hall, D. (1993). The development and validation of categorization of sources of misconceptions in chemistry. Paper presented at the Third *International Seminar on Misconceptions and Educational Strategies in science and Mathematics*, Ithaca.
- Smith, K.J. & Metz, P.A. (1996). Evaluating student understanding of solution chemistry through microscopic representations. *Journal of Chemical Education*, 73 (3), 233-235.
- Solomon, J. (1982). How children learn about energy or do the first law come first? *School Science Review*, 63(224), 415-422.
- Stavridou, H., Solomonidou, C. (1998). Conceptual reorganization and the construction of the chemical reaction concept during secondary education. *International journal of science*.
- Stylianidou, F. (2002). Analysis of science textbook pictures about energy and pupils' readings of them. *International Journal of Science Education*.
- Şahan, B.Y. & Tekin, L. (2007). *Ortaöğretim 10. sınıf fizik ders kitabı*. [Secondary school 10th grade physics textbook]. Zambak Yayınları.
- Şimşek, M. (2013). Sosyal bilgiler öğretmen adaylarının coğrafi bilgi sistemleri (CBS) konusundaki bilişsel yapılarının ve alternatif kavramlarının kelime ilişkilendirmesi testi ile belirlenmesi. [Determining the cognitive structures and alternative concepts of social studies teacher candidates about geographical information systems by word association test]. 4. Ulusal ilköğretim bölümleri öğrenci kongresi, 8-9 Kasım 2013 Nevşehir Üniversitesi, Nevşehir.
- Tavşancıl, E., & Aslan, E. (2001). *İçerik analizi ve uygulama örnekleri*. [Content analysis and application examples]. Epsilon Yayınları.
- Thomas, G. V., & Silk, A. M. J. (1990). An introduction to the psychology of children's drawings. Hemel Hemstead, UK: Harvester Wheatsheaf.
- Töman, U. & Cimer, O. S. (2011). Enerji kavramının farklı öğrenim seviyelerinde öğrenilme durumunun araştırılması. [An investigation into the conceptions of energy at different educational levels]. *Bayburt Üniversitesi Eğitim Fakültesi Dergisi* 6, (I-II), 23.
- Töman, U. & Odabaşı, O. S. (2012). Enerji dönüşümü kavramının farklı öğrenim seviyelerinde öğrenilme durumunun araştırılması. [An investigation into the conception energy conversion at different educational levels]. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi*, 14 (2), 289-312.
- Toman, U., Karatas, O. F. & Cimer, O. S. (2016). Development and implementation of a standard test to diagnose misconceptions about energy and related concepts: The beginning. *Bayburt Eğitim Fakültesi Dergisi*, 8(1), 116-134.
- Trumper, R. A. (1990). Being constructive: An alternative approach to the teaching of the energy concept-part one. *International journal of science education*, 12(4), 343-354.
- Trumper, R. (1996). Survey of Israeli physics students' conceptions of energy in pre-service training for high school teachers. *Research in Science and Technological Education* 14: 179-192.
- Trumper, R. A. (1998). A longitudinal study of physics students' conceptions on energy in pre-service training for high school teachers. *Journal of Science Education Technology*, 7(4), 311-318.
- Trefil, J., & Hazen, R.M. (2004). Physics matters an introduction to conceptual physics. Wiley, New York.
- Tsai, C. C., & Huang, C. M. (2002). Exploring students' cognitive structures in learning science: a review of relevant methods. *Journal of Biological Education*, 36(4), 163-169.
- Uyduran, G. (2019). Ortaokul öğrencilerinin "enerji" konusundaki bilişsel yapılarının kelime ilişkilendirme testi (kit) yoluyla incelenmesi. [Investigation of middle school students' cognitive structures about "energy" through word association test (WAT)]. Yüksek Lisans Tezi. Ömer Halisdemir Üniversitesi Eğitim Bilimleri Enstitüsü. Niğde.
- Ünal Çoban G., Aktamış H. & Ergin Ö. (2007) İlköğretim 8. sınıf öğrencilerinin enerjiyle ilgili görüşleri. [The views of 8th grade students about energy.] *G.Ü. Kastamonu Eğitim Dergisi*, 15(1), 175-184.
- Watt, D., M.(1983). Some alternative views of energy. *Physics Education*, 18, 213-217.
- Yağbasan, R. & Gülçiçek, Ç. (2003). Fen öğretiminde kavram yanlışlarının karakteristiklerinin tanımlanması. [Describing the characteristics of misconceptions in science teaching]. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, (1) 13.
- Yayla, R. G., & Eyceyurt, G. (2011). Mental models of pre-service science teachers about basic concepts in chemistry. *Western Anatolia Journal of Educational Sciences*, 285-294
- Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri*. [Qualitative research methods in the social sciences]. Seçkin Yayıncılık.

- Yıldırım, T., Önal, N., Büyük, U. (2019). Sekizinci sınıf öğrencilerinin yenilenebilir enerji kaynaklarına ilişkin algılarının bilim karikatürleri aracılığıyla incelenmesi. [Investigation of eighth grade students' renewable energy resources perceptions by science cartoons]. *Kuramsal Eğitim Bilim Dergisi*, 12(1), 342-368.
- Yuenyong, C., & Yuenyong, J. (2007). Grade 1 to 6 Thai Students' existing ideas about energy. *Science Education International*, 18(4), 289-298.
- Yürümezoğlu, K., Ayaz, S. & Çökelez, A. (2009). Grade 7-9 students' perceptions of energy and related concepts. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 3(2), 52-73.



Article History

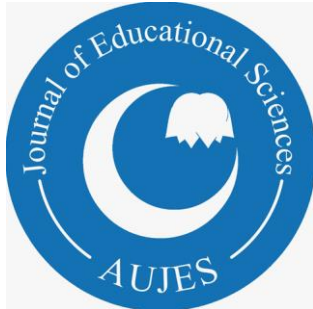
Received: 21.06.2020

Received in revised form: 23.05.2021

Accepted: 29.05.2021

Available online: 29.06.2021

Article Type: Research Article





ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

<https://dergipark.org.tr/tr/pub/adyuebd>

**Measurement and Assessment Literacy
Levels of Teachers in Terms of Some
Variables**

Ayşe Özlem Ergül¹, Sevda Çetin²

¹Ministry of National Education, Van, Turkey 

²Hacettepe University, Department of Educational Sciences,
Ankara, Turkey 

To cite this article:

Ergül, A. Ö. & Çetin, S. (2021). Measurement and assessment literacy levels of teachers in terms of some variables. *Adiyaman University Journal of Educational Sciences*, 11(1), 26-35.

Measurement and Assessment Literacy Levels of Teachers in Terms of Some Variables

Ayşe Özlem Ergül¹, Sevda Çetin^{2*}

¹ Ministry of National Education, Van, Turkey

² Hacettepe University, Department of Educational Sciences, Ankara, Turkey

Abstract

Measurement and assessment literacy of teachers is important not only to measure student performance but also to evaluate the functioning of the education system and to decide whether the education given by teachers is qualified. In this study, assessment literacy inventory was applied to 189 secondary school teachers from different branches to determine teachers' measurement and assessment literacy levels. It was discovered that the measurement and assessment literacy level of the teachers was low. It was also found that measurement and assessment literacy levels of the teachers significantly differ by professional seniority and branch. Also, in the study item examination skill test was applied to secondary school mathematics teachers to determine the relationship between secondary school mathematics teachers' measurement and assessment literacy levels and their skill to examine items appropriate to attainment and grade level. A moderately significant positive correlation was found between the teachers' skill to examine items appropriate to educational attainment and grade level, and their measurement and assessment literacy levels. Due to the low measurement and assessment literacy level of secondary school teachers, studies such as seminars or training on this subject can be conducted. Practical activities can also be conducted along with the theoretical information on the subject.

Keywords: Measurement and Assessment Literacy, Teacher Competencies, Item Examination Skills

Introduction

The general aim of education is to equip individuals with the knowledge and skills required by the era and make them ideal individuals for their society. Individuals should be trained in a way to be able to adapt and make a contribution to adjustments and trends in the world and society. This can be ensured by qualified teachers along with a good education system (Dilaver, 1996). Education is a process and this process has three basic components: teacher, student, and curriculum. Raising individuals with the desired characteristics and ensuring quality, effective and efficient education is straight associated with the tightness of the hyper between these three primary elements. The most important role among these components belongs to the teacher (Arslan & Özpınar, 2008; Bulut, 2009; Kavas & Bugay, 2009; Kuş & Çelikkaya, 2010).

Teacher competence refers to the knowledge, skills, and attitudes that teachers will need to have to be able to find a way to meet the education career successfully and efficiently. The fact that teachers have these competencies is very important in increasing students' success and developing student personality (MEB, 2017a). In other words, the development of teachers' professional competencies increases the quality of education (Aybek, 2017).

Newfields (2006) explained the importance of measurement and assessment as follows. Measurement and assessment are a common part of all education systems in the world. It helps to understand how education programs work and enables teachers to see their performance. The more convenient and efficient the measurement and assessment applications are used, the more the student's learning performance will increase (Mertler & Campbell, 2005). Measurement and assessment help to determine and interpret student's readiness and to correct the deficiencies with the results obtained, so the student's learning quality is improved (Black & Williams, 1998). According to the studies, teachers spend 50% of their time with activities that include measurement and assessment (Plake, 1993). This very importance of measurement and assessment has brought along many pieces of research examining the competencies of teachers in measurement and assessment.

The concept of literacy is generally defined as the ability to read and write, but other than that, it is also used in the sense of knowledge and competence of individuals in a particular subject area (Koh, Burke, Luke, Gong, & Tan, 2017). Measurement and assessment literacy is the knowledge and skill of right management to

* Corresponding Author: Sevda Çetin, tsevda@hacettepe.edu.tr

detect the effectiveness of the curriculum and to evaluate students' success by selecting, developing, applying, scoring, managing, informing, and transmitting the results of the measurement tools in line with ethical rules and principles (Koh, et al., 2017).

The concept of measurement and assessment literacy was first introduced in 1991 by Richard Stiggins. According to Stiggins, measurement and assessment literate educators comprehend what they measure, why they measure, how they measure, what are the possible problems related to measurement, and the way to forestall these problems. It is also argued that educators are acquainted with the miserable penalties of improper and insufficient assessment. (Stiggins, 1991).

Measurement and assessment literacy consists of understanding basic measurement and assessment concepts and procedures that affect an individual's decisions about education (Popham, 2018). Fundamental concepts of measurement and assessment such as validity, reliability, and fairness refer to the methods and procedures used when creating or evaluating a test. Measurement and assessment literacy is the ability of a teacher to measure, interpret what the students learned, and use the measurement and assessment results obtained to improve student's learning and improve the quality of education provided (Webb, 2002).

A measurement and assessment literate teacher should be able to choose the most appropriate measurement tool to realize teaching achievements (Gottheiner & Siegel, 2012). They need to have the ability to appreciate the reliability of this measurement and assessment tool, know the concepts such as reliability and validity, and be aware that these concepts are effective in making educational decisions (Popham, 2011). A teacher who is measurement and assessment illiterate falls into a systematic error because s/he cannot ensure the reliability and structural validity of the measurement tool to be used, and this endangers the education system by making false assessments and taking false decisions (Lai Waltman, 2008).

In a study conducted in 2006, Newfields explained the importance of measurement and assessment literacy for three persuasive reasons. First, measurement and assessment are common features of many education systems. Teachers spend greater than half their time on measurement and assessment activities, and most school budgets and time are spent on standardized tests. Second, it provides an understanding of the literature on education. Understanding the basic statistical concepts provides a critical approach to a piece of research, otherwise, the research moves away from the reality of science and unfounded knowledge emerges. Finally, being a measurement and assessment literate teacher allows conveying results about the general condition of the class to others. In this way, the teacher shares his research with other colleagues and results that encourage learning.

Measurement and assessment competencies are the knowledge and skills that a teacher should have as an educator. The inadequacy of teachers in measuring and evaluating student development has revealed the need to develop measurement and assessment competence standards (AFT, NCME & NEA, 1990). The first study on measurement and assessment standards was conducted in 1987 in collaboration with the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA). This project, carried out by the committees, was completed in 1990. In time, many committees and researchers made efforts to develop standards similar to those established by this committee. One of these is the 11-item measurement and assessment standards developed by Brookhart in 2011 by improving the 1990-standards. In Turkey, the first studies on teacher competencies started in 1998. The current version is given by examining the teacher competence documents of organizations such as the International Labor Organization (ILO), the Organization for Economic Development and Cooperation (OECD), the United Nations Educational, Scientific and Cultural Organization (UNESCO), the European council and countries such as Finland, England, Canada, and Singapore. (MEB, 2017). Table1 presents the measurement and assessment standards prepared by this committee and individuals.

Table 1. Measurement and Assessment Standards

AFT, NCME, and NEA (1990)	Brookhart (2011)	MoNE (Ministry of National Education) (2017)
To have the ability to select measurement and assessment methods suitable for teaching decisions	To have the knowledge of subject area related to the field	To prepare and use measurement and assessment tools suitable for students' developmental features
To have the ability to develop measurement and assessment methods suitable for teaching decisions	To reveal the situations that are compatible with the content and depth of thought determined by the curriculum objectives and standards during the assessment	To use process and result-oriented methods in measurement and assessment
To have the ability to interpret,	To have a strategy to communicate	To make measurement and

score and manage the results of measurement methods	with students about their success	assessment objectively and fairly
To have the ability to use measurement and assessment results when making decisions about education	To know the philosophy, purpose, advantages and disadvantages of the assessment methods preferred	To provide accurate and constructive feedback to students and others according to the results of measurement and assessment
To have the ability to develop a valid grading system to be used in student assessment	To make item analysis of questions, to know performance assessment content for thought skills and special information	To rearrange the teaching and learning processes according to the results of measurement and assessment
To have the ability to transmit measurement results to students, parents, educators and individuals	To have the ability to provide feedback that is useful and effective in activities of the students	
To be aware of ethical and illegal practices	To have the ability to create a scoring key for student success assessment	
	To have the ability to interpret the results of decisions related to students, class, school and regions and to manage external assessments	
	To be able to explain and interpret the decisions taken according to the results of the assessment to the people they serve with related reasons	
	To ask students for help in using assessment information to give correct education decisions	
	To know the responsibilities required for the assessment process to be legal and ethical	
AFT, NCME, and NEA (1990)	Brookhart (2011)	MoNE (2017)

As these statements point out, measurement and assessment are an integral part of education. Measurement and evaluation literacy has become necessary to determine the functioning of the education system, to establish if the education given by teachers is eligible and to improve the success of pupils. This study measures the teachers' measurement and assessment literacy levels with improving, regulatory and regenerative data, and investigates the relationship between the measurement and assessment literacy levels of secondary school teachers with respect to various variables.

Purpose and Importance of the Research

The study aims to determine the measurement and assessment literacy levels of secondary school teachers, to examine which teacher competence is deficient, and to reveal the relationship between measurement and assessment literacy and various variables (professional seniority, branch, and item examination skills).

The more appropriate and efficient the measurement and evaluation applications are used, the more the student's learning performance will increase. The student's readiness is determined, interpreted and the student's deficiencies are eliminated with the assessment and evaluation, thus the quality of the student's learning is improved. Also, considering the measurement tools that are not suitable for the purpose (acquisition), whose reliability and validity are not investigated, and the teachers' inability to score, interpret and manage the results of assessment and evaluation, it is very difficult to interpret successful or unsuccessful in educational activities, even to say that this situation is more harmful than the benefit of education.

The competencies of teachers in the field of measurement and assessment are the determiner of both the education and the future of the student. This study is important to measure the teachers' measurement and assessment literacy levels, to determine deficiencies in this field, to reveal the relationship between measurement and assessment literacy and the variables of professional seniority, branch and item examination skills, and to take reformative, regulatory and renovator precautions in line with the analysis of the obtained data, in other words, results.

Research Problem

How are the measurement and assessment literacy levels of secondary school teachers with respect to various variables?

The sub-problems of the study can be listed as follows.

1. How are the measurement and assessment literacy levels of secondary school teachers?
2. Do the measurement and assessment literacy levels of secondary school teachers differ significantly in terms of professional seniority and branch?
3. What is the correlation between secondary school mathematics teachers' measurement and assessment literacy level and their ability to examine items appropriate to educational attainment and grade level?

Method

Survey research which is one of the quantitative research types was used in the study. Survey research is a study that includes the use of a questionnaire to collect data from a sample of elements drawn from a population. In this type of research, the relationships between the variables measured in the study can also be looked at (Büyüköztürk, Kılıç, Akgün, Karadeniz & Demirel, 2016). In this current study assessment literacy inventory was applied to secondary school teachers to determine teachers' measurement and assessment literacy levels. Also, the item examination skill test was applied to secondary school mathematics teachers to determine the relationship between secondary school mathematics teachers' measurement and assessment literacy levels and their skill to examine items appropriate to attainment and grade level.

Study Group of the Research

The study group of this current study has consisted of 189 secondary school teachers working in different schools in a big-scale city in the Eastern Anatolia region of Turkey. The convenience sampling method was used to select teachers.

Table 2. Demographic Features of Study Group

		Frequency	Percentage
Gender	Female	92	48.7
	Male	97	51.3
Professional Seniority	0-4	75	39.7
	5-9	64	33.9
	10 and above	50	26.5
Education level	Undergraduate	171	90.5
	Master	18	9.5
Branch	Mathematics	54	28.6
	Turkish language	35	18.5
	Physical sciences	20	10.6
	Social studies	18	9.5
	English language	24	12.7
	Education of religion	18	9.5
	Other	20	10.6
MoNE In-Service Training	Yes	56	29.6
	No	133	70.4
	Total	189	100

As can be seen in Table 2, 48.7% (n = 92) of the teachers who participated in the study are female; 51.3% (n = 97) are male teachers. 39.7% of the participants (n = 75) were teachers whose professional seniority varied between 0-4 years, 33.9% (n = 64) were teachers whose professional seniority varied between 5-9 years and 26.5% (n = 50) are teachers with professional seniority of 10 years or more. 28.6% of the participants (n = 54) were mathematics teachers and 70.4% (n = 133) of the teachers who participated in the study did not receive in-service training on measurement and assessment in MoNE.

Data Collection Tools

The measurement tool used for the study consists of three parts: personal information form, measurement and assessment literacy inventory, and item examination skill test.

Personal Information Form: In the form, questions are included regarding the teachers' personal information, gender, education level, professional seniority, branch, and in-service training on measurement and assessment in MoNE.

The Assessment Literacy Inventory (ALI): Test developed by Mertler and Campell in 2005 was adapted into Turkish by Bütüner et al. in 2010. The Turkish form of the inventory was used in this study. This inventory, which allows not only to determine the assessment literacy levels of teachers but also to detect which teachers have deficiencies in certain competence areas, consists of five scenarios each containing six questions. ALI has been prepared in parallel with the teacher competence standards required in the educational assessment of students. Table 3 shows which items in the inventory provide information about related standards.

Table 3. Distribution of Items in Assessment Literacy Inventory by Standards

Teacher Competence Standards	Items
1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.	1, 7, 13, 19, 25
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.	2, 8, 14, 20, 26
3. The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.	3, 9, 15, 21, 27
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.	4, 10, 16, 22, 28
5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.	5, 11, 17, 23, 29
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.	6, 12, 18, 24, 30

The Kuder-Richardson Reliability Coefficient (KR20) of the inventory containing 30 items in total was calculated as 0.86.

The Item Examination Skill Test: The test was prepared by the researchers. The form contains 7 multiple choice math items that are suitable for different grade levels (5, 6, 7, and 8th grade) of the secondary school. While determining the items to be put in the test, the questions of the common exams held by MoNE, whose level was determined according to class level, attainment, and Bloom taxonomy, were used. Respondents were asked to find out the grade level, learning outcome, and Bloom taxonomy level of these items.

When preparing an achievement test, items with medium difficulty, high discrimination power, and suitable for attainment and student-level should be placed in the measurement tool. At the same time, it is important to prepare the questions to be used in these tests in different forms according to Bloom Taxonomy (Linn & Gronlund, 1995). Otherwise, an improper measurement tool will result in incorrect evaluation. In this case, it leads to wrong decisions about education. According to teacher competencies, a teacher should have the ability to select and develop appropriate assessment and evaluation methods (AFT, NCME & NEA, 1990).

The test, which includes 21 items in total, includes 3 items related to the above-mentioned concepts of each of the 7 multiple-choice math items. To determine the reliability of the test, it was applied to 224 secondary school mathematics teachers. The results were analyzed using TAP (Test Analysis Program) and the reliability of the 21-item test was calculated as 0.57. Keohe (1995) stated that the reliability coefficient is acceptable around 0.50 in short tests with 10-15 items, and above 0.80 in tests with 50 or more items. Besides, in the literature, the mean of the correlations between items is ideal between 0.20 and 0.40. The values determined indicate that the items are homogeneous enough and contain the original variance. In the developed test, the correlation value between the mean items was 0.21. These values show that the test is reliable (Tabachnick & Fidell, 2007).

Results

Results on the First Sub-Problem

Results regarding the measurement and assessment literacy levels of teachers, the first sub-problem of the study, are presented in Table 4.

Table 4. Measurement and assessment Literacy Levels of Teachers

Standard	\bar{X}^*	SS
1. Choosing appropriate measurement and assessment methods	3.03	1.08
2. Developing appropriate measurement and assessment methods	2.10	1.08
3. Managing, scoring and interpreting measurement and assessment results	2.62	1.20
4. Using measurement and assessment results while making decisions about students, education planning, curriculum development and school development.	1.76	1.02
5. Developing a valid grading system (rubric) to be used to evaluate students	1.39	1.44
6. Communicating measurement and assessment results to others	2.04	1.11
Total	12.94	3.66

* The highest score is 5 and the lowest score is 0 for each standard.

Assessment literacy inventory was applied to 189 secondary school teachers to determine teachers' measurement and assessment literacy levels. Secondary school teachers correctly answered approximately 13 (43%) out of 30 questions in the inventory. According to this finding, it can be suggested that measurement and assessment literacy levels of teachers are quite low. Considering the performances of the secondary school teachers by the assessment competence standards, it was found that the standard at which the teachers were the best was 'choosing assessment methods appropriate for instructional decisions' ($\bar{X} = 3.03$), while the standard at which they were the worst was 'developing valid grading procedures (rubric) which use student assessments' ($\bar{X} = 1.39$). According to the results of the inventory applied to the secondary school teachers in the study group, the teachers correctly answered approximately 13 of the 30 questions on average, or 43%, on average.

Results on the second sub-problem

Examination of the relationship between measurement and assessment literacy level and professional seniority. One-way ANOVA (Variance Analysis) was applied to analyze whether there is a significant difference between secondary education teachers' measurement and assessment literacy levels and the professional seniority variable. The results are presented in Table 5.

Table 5. ANOVA Results by the Professional Seniority Variable of Measurement and assessment Literacy Levels of Secondary School Teachers

Source of Variance	Total Sum of Squares	sd	Average of Squares	F	p	Significant Difference	(η^2)
Inter-group	171.85	2	85.93	6.79	0.01	0-4 and above 10 5-9 and above 10	0.07
Intra-groups	2352.51	186	12.65				
Total	2524.36	188					

According to the results of the analysis, a statistically significant difference was found between teachers' measurement and assessment and literacy level and their professional seniority ($F_{(2,186)} = 6.79$; $p < 0.05$). Measurement and assessment literacy levels of teachers vary significantly depending on professional seniority. The effect size was calculated as eta-square (η^2) = 0.07. The effect size was calculated as eta-square (η^2) = 0.07. η^2 is interpreted as the proportion of variance of the dependent variable that is related to the factor. η^2 of .01, .06, and .14 are, by convention, interpreted as small, medium, and large effect sizes, respectively (Green and Salkind, 2005). Accordingly, it can be suggested that the professional seniority variable has a moderate effect on measurement and assessment literacy. At the same time, it can be said that only 7% of the variance observed in the scores obtained from the assessment literacy inventory depends on professional seniority. As can be seen in the table, there is a statistically significant difference between the teacher who has worked for at least 10 years ($\bar{X}=11.10$) and others who worked for 4 years ($\bar{X}=13.89$) and 5-9 years ($\bar{X}=12.95$). This difference is against teachers who have worked for at least 10 years. Based on this result, it can be said that teachers who have worked for at least 10 years have lower measurement and assessment literacy levels than those who have worked for less than 10 years.

Investigation of the relationship between measurement and assessment literacy level and branch. Table 6 presents the results of one-way ANOVA conducted to determine whether there is a significant difference between the measurement and assessment literacy levels of secondary school teachers and the branch variable.

Table 6. ANOVA Results of Secondary School Teachers' Measurement and assessment Literacy Levels by the Branch Variable

Source of Variance	Total Sum of Squares	sd	Average of Squares	F	p	Significant Difference	(η^2)
Inter-group	171.55	6	28.59	2.21	0.04*	Math. - Other	0.07
Intra-groups	2352.81	182	12.93			E. of Religion – Math. E. of Religion - Physical Sci.	
Total	2524.36	188				E. of Religion - English L.	

* p < 0.05

According to the results of the analysis, a statistically significant difference was found between teachers' measurement and assessment literacy level and branch ($F_{(6,182)} = 2.21$; $p < 0.05$). This result shows that the branch variable has a moderate effect on measurement and assessment literacy. At the same time, it can be said that only 7% of the variance observed in the scores of the assessment literacy inventory depends on the branch.

As presented in the table, there is a significant difference between mathematics teachers ($\bar{X}=13.85$) and others ($\bar{X}=11.90$) (visual arts, technology design, music, information technologies, guidance counselor) and education of religion teachers ($\bar{X}=10.78$) in favor of mathematics teachers. Accordingly, it can be suggested that mathematics teachers' measurement and assessment literacy levels are higher than others and education of religion teachers. There is a significant difference between measurement and assessment literacy levels of physical sciences ($\bar{X}=13.55$) and education of religion teachers ($\bar{X}=10.78$) in favor of physical sciences teachers. Similarly, there is a significant difference between English language teachers ($\bar{X}=13.54$) and education of religion teachers ($\bar{X}=10.78$) in favor of English language teachers.

Results on the Third Sub-Problem

For the third sub-problem, the relationship between secondary school mathematics teachers' measurement and assessment literacy levels and their skill to examine items appropriate to attainment and grade level was analyzed.

Item examination skill test was applied to 54 secondary school mathematics teachers. In the assessment literacy inventory, approximately 14 of 30 questions (46% of the questions in the inventory) were answered correctly by these participants. These teachers have the highest mean ($\bar{X}=13.85$) among teachers in other branches in the assessment literacy inventory. For the item examination skill test these participants correctly answered around 14 questions out of 21 questions in the item examination skill test, in other words, participants answered 64% of the questions correctly.

Since the literacy level and item examination skill are continuous variables and they are distributed normally together, the Pearson moments product correlation coefficient was calculated to determine the direction and amount of the relationship between the variables. The analysis results are presented in Table 7.

Table 7. Correlation between Measurement and Assessment Literacy Level and Item Analysis Skill

Variable	N	r	p
Literacy Level – Item examination Skill	54	0.40	0.00*

* p < 0.05

As presented in Table 7, a statistically positive and moderately significant relationship was found between the variables obtained from the assessment literacy inventory ($r = 0.40$; $p < 0.05$). Accordingly, as the literacy level of the teachers increases, the item examination skill also increases. When the determination coefficient (square of the correlation coefficient) ($r^2 = 0.16$) is examined, it can be said that 16% of the total variance at the literacy level is caused by the item examination skill.

Table 8. Correlation Between Item Examination Skill and Teacher Competency Standards

Variable	N	r	p
Standard 1 - Item Examination Skill	54	0.34	0.01*

Standard 2 - Item Examination Skill	54	0.23	0.10
Standard 3 - Item Examination Skill	54	0.21	0.13
Standard 4 - Item Examination Skill	54	-0.02	0.89
Standard 5 - Item Examination Skill	54	0.18	0.20
Standard 6 - Item Examination Skill	54	0.37	0.01*

* $p < 0.05$

In Table 8, the relationship between the item examination skill and the standards involved in the assessment literacy inventory is presented. As presented in the table, there is a significant relationship only between the first and sixth standards and the item examination skills ($p < 0.05$). There is a positive and moderately significant relationship between the ability to choose suitable assessment methods for instruction (standard 1) and the item examination skill ($r = 0.34$; $p < 0.05$). It was also found that there is a moderate positive relationship between the ability to communicate measurement and assessment results to the educators (standard 6) and the question analysis skill ($r = 0.37$; $p < 0.05$). As seen in results as the literacy level of the teachers increases, the skill to analyze questions also increases. It is possible to say the opposite.

Discussion, Conclusions and Recommendations

This study was carried out to determine the measurement and assessment literacy levels of secondary school teachers, to reveal which deficiencies they have, and to determine the relationship between the measurement and assessment literacy levels and some variables. The study also serves to reveal the ability of teachers to conduct the item examination at the required difficulty and level in line with the attainment and grade level in the curriculum.

The current study has found that secondary school teachers in the study group correctly answered approximately 13 of the 30 questions on average, or 43%, on average. Compared to other studies in the literature, measurement and assessment literacy level were similarly found to be insufficient. According to the results of the research carried out in other countries, teachers correctly answered 23 of the 35 items (66%) in the study of Plake and Impara (1993). Pre-service teachers correctly answered 21 of the 35 questions (60%) in the study of Campbell et al. (2002) and approximately 19 (54%) of 35 items in the study of Mertler (2003). In the study of Davidheiser (2013) with 180 high school teachers, the participants answered approximately 24 (68%) of 35 questions correctly. Regarding the results of the study conducted in Turkey, Gül (2011) determined the measurement and assessment level of 180 pre-service teachers who correctly answered approximately 18 (50%) of 35 questions. In another study carried out by Karaman and Şahin (2014), it was reported that fourth-grade pre-service teachers correctly answered approximately 16 (51%) of 30 questions. In a similar study, Azrak (2017) revealed that social studies pre-service teachers correctly answered approximately 10 questions (33%) of 30 questions. The aforementioned studies indicate that secondary school teachers' measurement and assessment literacy levels are low.

Additionally, in this current study measurement and assessment literacy levels of secondary school teachers were examined according to each competence area in the inventory. The most competent area (standard) in which the teachers performed the highest was found to be choosing appropriate measurement and assessment methods ($\bar{X} = 3.03$), and the least competent area was found as developing a valid grading system ($\bar{X} = 1.39$) which use student assessments. Campbell et al. (2002), Mertler (2003) (for pre-service teachers), Gül (2011), and Karaman and Şahin (2014) also found the competence area to select appropriate measurement and assessment methods as the most highest-performance area. Plake and Impara (1993) and Mertler (2003) (for teachers) found that the highest-performance competence area was found as managing, scoring, and interpreting measurement and assessment results. Consistent with the results of Mertler (2003) (for teachers) and Karaman and Şahin (2014), the lowest-performing competence area is choosing a valid pupil grading system to be used in the assessment of students in the present study. The competence area for communicating measurement and assessment results is the lowest-performing competence area in the studies of Plake and Impara (1993), Campbell et al. (2002), Mertler (2003) (for pre-service teachers) and Gül (2011). In the studies carried out by Gelbal and Kelecioğlu (2007) and Erdoğan and Kurt (2012), teachers have been reported to be insufficient in the field of measurement and assessment, and more education is needed.

The current study has determined that there is a significant difference between the professional seniority of secondary school teachers and the level of measurement and assessment literacy. As the professional seniority of the teachers increased, the level of measurement and assessment literacy decreased. Accordingly, it can be said that the professional seniority variable has a moderate effect on measurement and assessment literacy. According to this study, it can be said that teachers who have worked for at least 10 years

have lower measurement and assessment literacy levels than those who have worked for less than 10 years. This particular finding was not compatible with other studies. Erdost (2018) detected a linear relationship between experience and measurement and assessment literacy level. Likewise, in the study of Plake and Impara (1993), it was found that experienced teachers' measurement and assessment literacy levels were higher than those of less experienced teachers. This finding may result from regional differences, interpersonal differences, attitudes towards measurement and assessment.

Another finding of the study is that the branch variable has a moderate effect on measurement and assessment literacy. A statistically significant difference was found between the branch of secondary school teachers and their level of measurement and assessment. There is a significant difference between mathematics teachers and others (visual arts, technology design, music, physical education, information technologies, guidance counselor) and education of religion teachers in favor of mathematics teachers. Accordingly, it can be suggested that mathematics teachers' measurement and assessment literacy levels are higher than others and education of religion teachers. It is also found that there is a significant difference between measurement and assessment literacy levels of physical sciences and education of religion teachers in favor of physical sciences teachers. Similarly, there is a significant difference between English language teachers and the education of religion teachers in favor of English language teachers. In the literature, Karaman and Şahin (2014) found a significant difference between the branch and measurement and assessment literacy levels. This finding is compatible with the study.

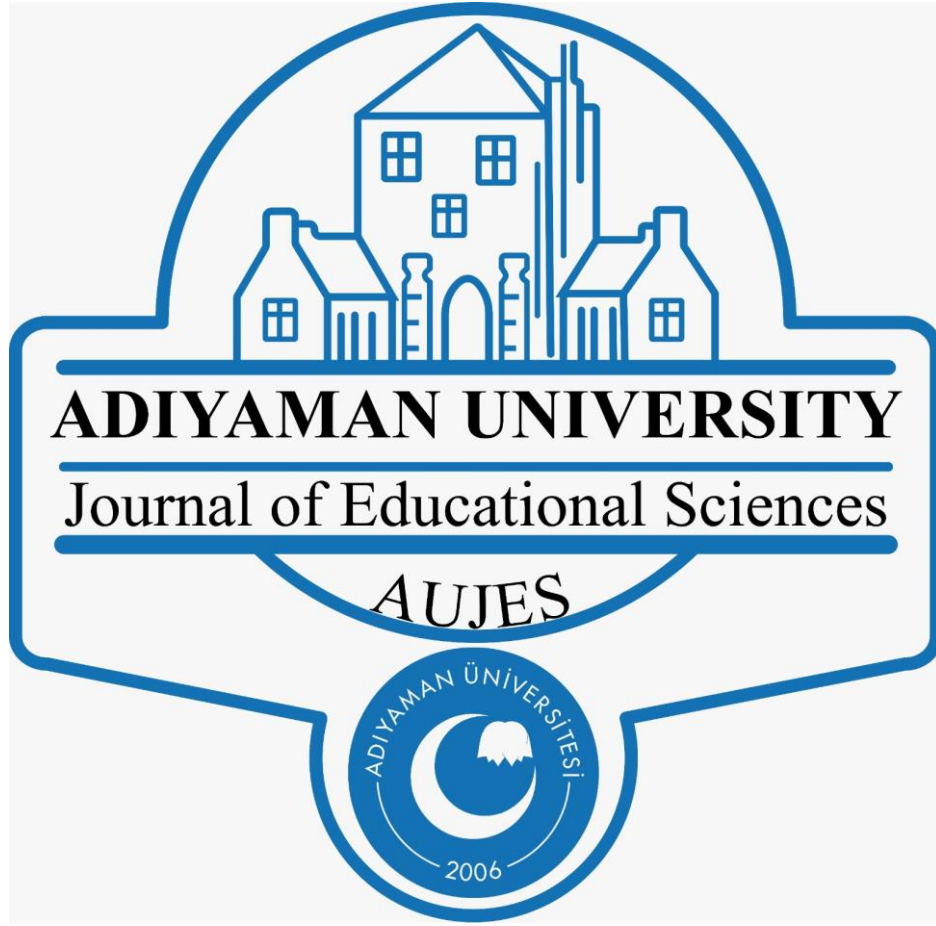
Lastly, the current study has determined that secondary school mathematics teachers answered correctly around 14 questions on average out of 28 questions in the item examination skill test, that is, the participants answered 64% of the questions correctly. In the assessment literacy inventory, this sample group answered correctly around 14 questions out of 30 questions, in short, the participants answered 46% of the questions in the inventory correctly. A statistically positive and moderately significant relationship was found between the measurement and assessment literacy level of secondary school mathematics teachers and the skill to examine items in the desired way in terms of the level of attainment and grade level. Accordingly, as the literacy level of the teachers increases, the skill to analyze questions also increases. It is possible to say the opposite.

Although this study has important findings, it has some limitations: the data of this study, for instance, were obtained from secondary school teachers. Further studies can be conducted with teachers from different levels of education such as primary school, secondary school, and high school, and it can be investigated whether there is a significant difference between the education level that the teacher is at service and the measurement and assessment literacy. Also, the "Item Examination Skill Test" prepared by the researchers can be prepared and developed not only for the mathematics lesson but also for other lessons and the difference can be calculated according to the branch. Also, due to the low measurement and assessment literacy level of secondary school teachers, studies such as seminars or training on this subject can be conducted. Practical activities can also be conducted along with the theoretical information on the subject.

References

- American Federation of Teachers, National Council on Measurement in Education, National Education Association (AFT, NCME, NEA). (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, 9(4), 30-32.
- Arslan, S. & Özpınar, İ. (2008). Öğretmen nitelikleri: İlköğretim programlarının beklentileri ve eğitim fakültelerinin kazandırdıkları. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, 2(1), 38-63.
- Atanur Başkan, G. (2001). *Öğretmenlik mesleği ve öğretmen yetiştirmede yeniden yapılanma*. Denge Yayıncılık.
- Aybek, Ş. (2017). Öğretmen olmak bir yaşam biçimidir. *Hürriyet Gazetesi* <http://www.hurriyet.com.tr/egitim/ogretmen-olmak-bir-yasam-bicimidir-40349588>
- Azrak, Y. (2017). *Sosyal bilgiler öğretmen adaylarının ölçme-değerlendirme okuryazarlık düzeylerinin çeşitli değişkenler açısından incelenmesi*. (Unpublished master's thesis), Ömer Halisdemir University, Niğde.
- Black, P. & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan International*, 80 (2), 139-148.
- Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, 30(1), 3-12.
- Bulut, İ. (2009). Öğretmen adaylarının öğretmenlik mesleğine ilişkin tutumlarının değerlendirilmesi (Dicle ve Fırat Üniversitesi örneği). *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 14, 13-24.
- Bütüner, S.Ö., Yiğit, N. & Çimer, S.O. (2010). Ölçme değerlendirme okuryazarlığı envanterinin Türkçeye uyarlanması. *E-Journal of New World Sciences Academy*, 5 (3), 792-809.

- Büyüköztürk, Ş., Kılıç, E., Akgün, Ö. E., Karadeniz, Ş., Demirel, F. (2016). *Bilimsel araştırma yöntemleri*, Pegem Akademi.
- Davidheiser, A. S. (2013). *Identifying areas for high school teacher development: a study of assessment literacy in the central bucks school district*. (Unpublished doctorate dissertation), Drexel University School of Education, Philadelphia.
- Dilaver, H. (1996). Türkiye’de öğretmen istihdamının dünü, bugünü ve yarını, eğitimimize bakışlar. *İstanbul: Kültür Koleji Vakfı Yayınları*, 1, 119.
- Erdost, A. (2018). *Türk İngilizce okutmanlarının yabancı dilde ölçme ve değerlendirme okuryazarlığı: çoklu bir durum araştırması*. (Unpublished doctorate dissertation), Atatürk University, Erzurum.
- Gelbal, S. & Kelecioğlu, H. (2007). Öğretmenlerin ölçme ve değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 33, 135-145.
- Gottheiner, D. M. & Siegel M. A. (2012). Experienced middle school science teachers’ assessment literacy: investigating knowledge of students’ conceptions in genetics and ways to shape instruction. *The Association for Science Teacher Education*, 23, 531-557.
- Green, S. B., & Salkind, N. J. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data*. Prentice-Hall.
- Gül, E. (2011). *İlköğretim öğretmen adaylarının ölçme-değerlendirme okuryazarlığı ve ölçme-değerlendirmeye ilişkin tutumlarının belirlenmesi*. (Unpublished master’s thesis), Fırat University, Elazığ.
- Karaman, P., & Şahin, Ç. (2014). Öğretmen adaylarının ölçme değerlendirme okuryazarlıklarının belirlenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD)*, 15 (2), 175-189.
- Kavas, A. B. & Bugay, A. (2009). Öğretmen adaylarının hizmet öncesi eğitimlerinde gördükleri eksiklikler ve çözüm önerileri. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 25, 13-21.
- Koh, K., Burke, L., Luke, A., Gong, W. & Tan, C. (2017). Developing the assessment literacy of teachers in Chinese language classroom: A focus on assessment task design. *Language Teaching Research*, 22(3), 264-288.
- Kuş, Z. & Çelikkaya, T. (2010). Sosyal bilgiler öğretimi için sosyal bilgiler öğretmenlerinin beklentileri. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, 7(2), 69-91.
- Lai, E. R. & Waltman, K. (2008). Test preparation: Examining teacher perception and practices. *Educational Measurement: Issue and Practice*, 27(2), 28-42.
- Linn, R. I. & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th edition). Prentice-Hall.
- MEB, (2017a). *Öğretmenlik mesleği genel yeterlikleri*. http://oygm.meb.gov.tr/meb_iys_dosyalar/2017_12/11115355_YYRETMENLYK_MESLEYY_GENE_L_YETERLYKLERY.pdf
- MEB, (2017b). 2016-2017 Eğitim öğretim yılı ii. dönem merkezi ortak sinavi test ve madde istatistikleri. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_06/12171001_2017_2.doYnem_Merkezi_Ortak_SYnavY_genel_bilgiler_raporu_12.06.2017.pdf
- Mertler, C. A. & Campbell, C. (2005). Measuring teachers’ knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory. *Paper presented at the annual meeting of the American Educational Research Association*, Montréal, Quebec, Canada.
- Newfields, T. (2006). *Teacher development and assessment literacy*. Paper presented Proceeding of the 5th Annual JALT Pan-SIG Conference. Shizuoka, University College of Marine Science, Tokai.
- Plake, B. S. (1993). Teacher assessment literacy: Teachers’ competencies in the educational assessment of students. *Mid-Western Educational Researcher*, 6 (1), 21-27.
- Plake, B.S., & Impara, J.C. (1993). *Teacher assessment literacy questionnaire*. University of Nebraska-Lincoln. In cooperation with the National Council on Measurement in Education and the W.K. Kellogg Foundation.
- Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator’s confession. *The Teacher Educator*, 46(4), 265-273.
- Popham, W. J. (2018). *Assessment literacy for educators in a hurry*. ASCD.
- Shunk, D. H. (1996). Goal and self-evaluative influences during children’s cognitive skill learning. *American Educational Research Journal*, 33, 359-382.
- Stiggins, R. (1991). Assessment literacy. *Phi Delta Kappan*, 72, 534-539.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson Education
- Tan, Ş. & Erdoğan, A. (2004). *Öğretimi planlama ve değerlendirme*. Pegem A Publishing.
- Webb, N. (2002). Assessment literacy in a standards-based urban education setting. *Paper presented at the annual meeting of the American Educational Research Association*, New Orleans.



Article History

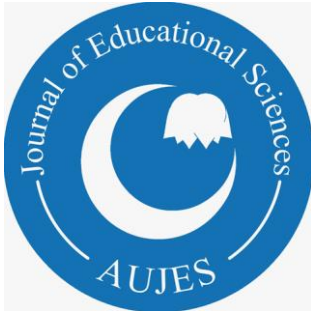
Received: 11.12.2020

Received in revised form: 02.01.2021

Accepted: 15.01.2021

Available online: 29.06.2021

Article Type: Research Article



ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

<https://dergipark.org.tr/tr/pub/adyuebd>

**Effects of Task Complexity on Text
Easibility and Coherence of EFL
Learners' Narrative Writing**

Mine Yıldız¹, Savaş Yeşilyurt²

¹Atatürk University, Department of Foreign Languages
Education, Erzurum, Turkey 

²Atatürk University Department of Foreign Languages
Education, Erzurum, Turkey 

To cite this article:

Yıldız, M. & Yeşilyurt, S. (2021). Effects of task complexity on text easibility and coherence of EFL learners' narrative writing. *Adiyaman Univesity Journal of Educational Sciences*, 11(1), 36-47.

Effects of Task Complexity on Text Easibility and Coherence of EFL Learners' Narrative Writing

Mine Yıldız^{1*}, Savaş Yeşilyurt²

¹ Atatürk University Department of Foreign Languages Education, Erzurum, Turkey

² Atatürk University Department of Foreign Languages Education, Erzurum, Turkey

Abstract

This study was carried out to examine the effects of task complexity on text easibility and coherence in narrative writing of EFL learners. Data were collected from 41 Turkish EFL learners during a writing course. Task complexity was operationalized at two levels as a complex and simple task based on the resource-dispersing variables of Robinson's the Triadic Componential Framework, +/- task structure. Accordingly, a colorful picture was first illustrated on the board, and students were asked to examine the picture for five minutes (complex task /-TS). They were then asked to write a story based on the picture they had seen (simple task /+TS). Two weeks later, they were given a sheet involving 16 pictures designed in an order and asked to narrate a story based on these pictures. Their essays were analyzed by the researcher and another rater in terms of coherence through an analytic rubric. An automated program was used to evaluate the essays for text easibility indices involving the indices of narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. The results analyzed with a Wilcoxon signed-rank test showed that complexity of the task had a statistically significant effect on some indices of text easibility such as word concreteness, and referential cohesion, whereas other indices, narrativity, syntactic simplicity, and deep cohesion, and coherence in their writing production were not affected by the complexity of the task at a significant level. However, it can be concluded that students' texts produced in simple tasks are easier to comprehend.

Keywords: Task-based language teaching, Task complexity, Coherence, Text-easibility, Coh-Metrix

Introduction

Task-based language teaching (TBLT) aims to enhance language learning based on the tasks involving meaningful, pragmatic, and communicative activities in all processes of learning such as planning, instruction, and assessment (Larsen-Freeman & Anderson, 2011). TBLT, one of the examples of a 'strong version' of the communicative approach, has received great attention of not only second language acquisition researchers but also researchers of second language teaching as it is primarily motivated by a theory of learning (Richards & Rodgers, 2001) and poses several advantages over PPP (present-practice-produce) paradigm that is claimed to be an over-simplified approach (Kırkgöz, 2014). In other words, the learners in TBLT are provided with the opportunity to learn the language through authentic scenarios involving meaningful, intentional, pragmatic, and surely communicative activities in which they are required to use their linguistic resources to perform the task. (Arslanyılmaz, 2013; Willis, 1996). Hence, they are able to use the language as a "vehicle for attending task goals" (Willis, 1996, p. 25) in a meaningful and natural atmosphere inside the classroom.

The main focus of a task is on meaning and the primary aim is to achieve an outcome as a result of the process using language (Skehan, 1998). Similarly, although Ellis (2009) states that "there is no single way of doing TBLT" (p. 224), he proposes three phases involving the pre-task phase, the main phase (the only one to be obligatory), and the post-task phase. Willis (1996), likewise, presents three elements but shows a difference in their features as pre-task, task cycle, and language focus (Willis, 1996, p. 135). Accordingly, a topic or task is introduced at the pre-task phase through clear and insightful instructions; the learners are required to conduct, plan how to report the outcome of their performance, and then to produce something at the phase of task cycle, and lastly, they analyze the recording of their reports and practice the phrases, words or structures at the phase of language focus. Therefore, writing activities can be described as a task since they meet all the requirements of a task. Their focus is on meaning, they involve a goal to be achieved by learners, and the learners obtain an outcome as a result of their performance. Furthermore, the writing activities are conducted at three phases as pre-writing, while-writing, and post-writing. Thus, students who participated in this study were asked to produce narrative essays for their task performance.

* Corresponding Author: Mine Yıldız, mine.yazici@atauni.edu.tr

Studies on Task Complexity

The studies reviewed in this study mostly are based on the two influential frameworks, Robinson's Cognition Hypothesis (Robinson, 2001, 2003, 2005; Robinson & Gilabert, 2007) and Skehan's Limited Attentional Capacity Model (Skehan & Foster, 1999, 2001), in terms of either dependent or independent variables, or both of them (Ellis & Yuan, 2004; Jackson & Suethanapornkul, 2013; Kawauchi, 2005; Kim, 2020a, 2020b; Ruiz-Funes, 2015; Tavakoli & Skehan, 2005; Yang, 2014; Yang, Lu, & Weigle, 2015; Yıldız & Yeşilyurt, 2017; Yuan & Ellis, 2003). Moreover, task complexity is generally identified along with the variables of these two competing models. Limited Attentional Capacity Model (Skehan, 2003, 2014; Skehan & Foster, 1999, 2001), which proposes that while performing a complex task people use more attentional resources due to a limited capacity they have in order to process information, deals with the task complexity under three dimensions such as code complexity, cognitive complexity, and communicative stress. While the focus of the code complexity is on linguistic demands the task requires, the content of the task and structuring of the material used in the task is seen within the concern of the cognitive complexity which is also divided into two main categories as *cognitive familiarity* and *cognitive processing*. The third dimension, communicative stress, basically focuses on the conditions and components of task performance such as participants, presentation, text, and time. Furthermore, it suggests that manipulations regarding the task require the learners to use more cognitive demand and attentional resources and thus reach production with trade-off effects among the three basic constructs complexity, accuracy, and fluency. Due to the limited attentional capacity learners have, they are not capable of paying simultaneous attention to those dimensions of language; that is, while paying attention to one dimension, they fail to focus on the others. In other words, focusing on producing complex performance probably leads to trade-off effect between accuracy and fluency since over-attention to complexity probably leads to lacks in accuracy or fluency of the production, or vice versa (Skehan, 2009).

The other model on which task complexity studies are based, Robinson's Triadic Componential Framework, similarly deals with the task complexity under three dimensions: task complexity, task conditions and task difficulty. This framework considers task complexity within the frameworks of information-processing demands required for a pedagogic task for memory, attention, and reasoning (Robinson, 2001). It presents two categories for the cognitive task features as resource-directing and resource-depleting variables, which was seen with a new name in the expanded version of Cognition Hypothesis as resource-dispersing variables (Robinson & Gilabert, 2007). The tasks manipulated along the variables of this framework are assumed to have language production in various ways. Whether there are few elements to be compared (+/- few elements) or not, whether the events occur in the past or present, or things are far or near (+/- here-and-now), whether some reasoning demands are provided for the learner (+/- reasoning) or not is assumed to require *cognitive* and *conceptual* demands and involved in resource-directing variables. On the other hand, the dimension of resource-dispersing variables deals with other variables such as *performative* and *procedural* demands for learners manipulated along whether planning time is allocated to the learners or not (+/- planning), the task has a loose or tight structure (+/- single task), and learners have prior knowledge or not (+/- prior knowledge) in order to apply in their task performance (Robinson, 2001). Furthermore, in contrast to Skehan (1998) suggesting that learners have to prioritize between the three dimensions due to their capacity and attentional resources to process information, Robinson (2001) suggests that learners' performance can be enhanced in all three dimensions of complexity, accuracy, and fluency (CAF).

The studies investigating whether increasing complexity of the task had any effect on language production show differences from many perspectives such as in terms of both dependent and independent variables although complexity, accuracy, and fluency are commonly preferred as dependent variables. For instance, Arslanyılmaz (2013) contrasted two different teaching tools – the computer-assisted task-based instruction (CATBI) and computer-assisted form-focused language instruction (CAFFI) in order to investigate the role of the teaching approach in second language development in terms of accuracy, lexical complexity, and fluency. According to the results of the study, the students taught through CATBI produced better language than those taught through CAFFI; in particular, although no significant difference for lexical complexity was seen, the language of production of task-based instruction was more fluent and accurate.

Tavakoli and Foster (2008) examined how oral second language performance is affected by narrative structure (tight/loose) and narrative complexity (with or without background information) in terms of the most common measures of task complexity, complexity, accuracy, and fluency (CAF). In support of previous studies, they concluded that accuracy appeared to increase through tight task structure and also that narrative tasks with background information seemed to result in higher syntactic complexity. In another study investigating the effects of task design on L2 task performance in terms of accuracy, fluency, syntactic complexity, and lexical diversity, Tavakoli (2009) pointed out that syntactic complexity could be enhanced through more structured tasks – narratives with both foreground and background storylines and also that they yielded more accurate and

fluent performance in more structured tasks than they did in less structured ones. However, no clear result has been obtained for the effect of task structure on lexical diversity.

Based on the cognitive frameworks for TBLT, Révész (2011) conducted a study with the goal of exploring whether there is a relationship between task complexity and learners' use of form-meaning mappings in oral tasks and also whether individual differences have an impact on such a relationship. Speech production of the participants performing two versions of the same argumentative task – complex or simple - manipulated along the +/- reasoning and the +/- few elements dimensions were analyzed through some global and specific measures of oral performance. It was illustrated that although participants' speech in the complex task was more accurate and lexically diverse but lower syntactically complex speech, no significant effects of learners' individual differences were observed.

In their study, Kuiken and Vedder (2007) firstly aimed to compare the two most influential models of task complexity – Robinson's Cognition Hypothesis and Skehan's Limited Attentional Capacity Model – regarding the effect of task complexity on L2 writing performance in terms of three measures of linguistic complexity and accuracy. The learners of Italian and French were assigned two writing tasks manipulated along with cognitive complexity as a non-complex condition in which they were required to write a letter taking three requirements into consideration and a complex condition in which six requirements would be considered. Although in previous studies Kuiken et al., (2005), Kuiken and Vedder (2008) revealed an effect of task complexity on accuracy evaluating it through general measures, the study by Kuiken and Vedder (2007) utilized more specific measures of accuracy and lexical variation regarding error type and the most frequent words used to illustrate their role for such an effect. The results of the study confirmed that fewer errors were seen in the complex task which might explain the accuracy case in the complex task; in other words, the fact that complex tasks yield more accurate texts probably results from a decrease of lexical errors in such tasks. As for the frequency of words, while French participants used less frequent words in a complex task, the case for the Italian participants was the opposite. In light of the results, it was also pointed out that it seemed not possible to establish a relationship between task complexity and language proficiency level.

Operationalizing task difficulty as the storyline structure – loose or tight – Ahmadian et al., (2012) investigated the effect of task difficulty on self-repair behavior in L2 oral performance. While performing the structured task the participants were observed to mainly focus on producing error-free units in terms of lexicon, grammar, and phonology; on the contrary, in the unstructured task, they were primarily concerned with conceptualizing the oral production producing D- repairs (different information involving alteration of the content of the preverbal message) and A-repairs (appropriacy that includes changes in the content of the message in terms of inaccuracy, incoherence, ambiguity, and inappropriacy) regularly.

Similarly, Adams et al., (2015) investigated the role of task structure and language support in increasing the accuracy and linguistic complexity of writing via text chat. For their four experimental groups, they implemented two task variables – task structure (+/- TS) from Robinson's (2007) Triadic Componential framework and language support (+/- LS) utilizing pre-task to raise consciousness. Whereas learners in the +TS case were provided with detailed written instruction about task performance and also a worksheet guiding them, those in the condition of low task structure (-TS) were given just basic instructions but no worksheet. Similarly, learners of +LS condition were provided pre-task language support activities, but others did take no language support in their task that is therefore expected to be more complex. Analysis of the chat texts on the engineering simulation task revealed that although the learners performing more complex tasks (-TS and -LS) produced less accurate texts, making tasks more complex had no impact on the linguistic complexity.

In order to investigate whether cognitive task complexity influences lexical and syntactic complexity, Frear and Bitchener (2015) utilized resource-directing variables (Robinson, 2007) by manipulating the number of reasoning demands (+/- reasoning) and numbers of elements (+/- few elements). As a result of their analysis of letters by L2 writers of English in terms of lexical variety through a mean segmental type-token ratio and syntactic complexity by the ratio of dependent clauses to T-units, it was pointed out that an increase appeared on the lexical complexity as a result of increasing complexity of cognitive task. However, in contrast to the expectation of the Cognition Hypothesis (Robinson, 2001, 2007), in which it is assumed that increases in task complexity will lead to language development resulting in complex language performance, no significant change was seen in syntactic complexity among tasks.

Like many researchers based on the assumptions of Robinson's Cognition Hypothesis, Salimi et al., (2011) investigated the effects of tasks manipulated along with resource-directing factors on accuracy, fluency, and syntactic complexity. Using two versions of the same decision-making task, complex and simple, their findings on fluency and complexity confirmed the predictions of Cognition Hypothesis that complex tasks would lead to more fluent and syntactically complex texts; nevertheless, the case for accuracy was different. No significant difference was obtained between complex and simple tasks in accuracy.

The studies in literature scarcely investigate the writing task performances of learners in terms of task complexity (Jackson & Suethanapornkul, 2013). For instance, in their study reviewing the studies on task complexity, Salimi and Dadaspour (2012) revealed that many of the studies regarding the effects of task complexity mainly focus on L2 oral performance but just a few on the written performance of L2 learners. Furthermore, Ellis and Yuan (2004), Kormos (2011), Salimi et al. (2011) and Yang et al. (2015) drew attention to a limited number of studies on the effect of task complexity on L2 writing performance. Therefore, this study focuses on whether the narrative writing performance of EFL learners is affected by task complexity manipulated along with resource-dispersing variables (+/- task structure) by Robinson's Cognition Hypothesis. Furthermore, there is very limited research investigating the effects of task complexity on written production in the Turkish EFL context. The studies found in the Turkish context were the study of Genç (2012) that investigates the effects of strategic planning on the accuracy of EFL learners' both oral and written narrative task performances and that of Yıldız and Yeşilyurt (2017) examining whether task planning had an influence on the complexity and overall writing quality of EFL learners' writing performance. Therefore, this study aims at probing into the effects of cognitive task complexity on written production of EFL learners in Turkey.

Text Easibility And Coherence

The present study was carried out to see the effect of task complexity on L2 writing performance in terms of two dependent variables as text easibility and coherence. Therefore, the two most commonly interchangeably used terms, coherence, and cohesion should be explained in detail to show the distinction between these terms. Though being an important characteristic of effective writing in terms of connectedness that "refers to all of the links, both explicit and implicit, in a text that make it a unified whole" (Watson Todd et al., 2007), coherence is generally thought to be an abstract and fuzzy term to define exactly and make a distinction from other concepts in writing such as cohesion, unity, etc. Lee (2002) describes coherence as "the relationships that link the ideas in a text to create meaning for the readers" (p. 135). It is commonly misused with the term cohesion: whereas cohesion, in simple terms, regards implicit links, coherence refers to the opposite, explicit links (Watson Todd et al., 2007). In other words, whereas cohesion is described as the connection of ideas at sentence level or "the connectivity of ideas in discourse and sentences to one another in text, thus creating the flow of information in a unified way" (Hinkel, 2004, p. 279), being a broader term, coherence is the organization of ideas at discourse level with all elements.

Coherence is, in simple terms, what the reader grabs from the text while cohesion provides the reader with linguistic elements-cohesive devices- to make a connection between ideas (Crossley et al., 2016). As stated in their seminal work "Cohesion in English", regarded as a theoretical framework on textual cohesion, Halliday and Hassan (1976) describe cohesion as a semantic concept that illustrates "relations of meaning that exist within a text" (p. 4). Similarly, according to Harmer (2004), writers use two main elements to build cohesion in text-linguistic techniques and grammar structures; in other words, like Halliday and Hassan (1976), he also describes cohesion in two headings- lexical cohesion and grammatical cohesion. The coherence that enables the reader to catch both "the writer's purpose" and "the writer's line of thought" is far beyond the sentence level and achieved through sequencing information in order to meet the expectations of the discourse community that it is written for (Harmer, 2004, p. 22-25).

As in definitions of coherence and cohesion, the research shows also differences in the ways or measures to assess them. For example, one of the scales applied to assess coherence both in spoken and written discourse is a topic-based analysis which depends on identifying key terms in a text, finding the relationships between these terms, ranking these relationships, and then mapping the text along the hierarchy identified through the relationships (Todd et al, 2004). In their study, Todd et al. (2004) applied topic-based analysis because it meets the three criteria defined by the researchers to select an appropriate scale to evaluate coherence: it (1) is objective, (2) unequivocally measures coherence, and (3) focuses on propositional coherence that is predominant in written discourse rather than interactional coherence seen in informal spoken language. Similarly, Knoch (2007) reported that the previous scales developed to assess coherence are either too time-consuming or complicated. Therefore, in his study undertaken in three phases as (1) analysis of writing samples, (2) rating scale design, and (3) rating scale validation, he chose and adapted a topical structure analysis (TSA) scale with the aim of investigating whether the use of a TSA scale-an empirically-based scale- to evaluate coherence in the written production of students is more reliable and has greater discrimination compared to the more traditional measures. However, the results revealed that although raters using the TSA scale scored more accurately, the TSA scale was not less time-consuming than the previous scale; rather, it might require more labor to analyze a large number of written texts and thus not practical in some cases.

McNamara et al. (2009) used Coh-Metrix-an automated tool- to examine whether the quality of the essays- low or high- can be predicted through the three indices as syntactic complexity, lexical diversity, and word frequency. In contrast to the general notion that more cohesive and thus more coherent essays are produced by more proficient writers, their study using linguistic indices of cohesion from Coh-Metrix could not

provide any evidence about whether there is a significant difference between high- and low-proficiency essays in terms of coherence; that is, the essays scored high were not more coherent than those rated low (McNamara et al., 2009). According to McNamara et al. (2009), the Coh-Metrix cohesion indices validated by a number of studies are confidential to assess cohesion. In the light of the literature, they reached a conclusion that Coh-Metrix is “an extremely powerful text analysis tool, capable of assessing and differentiating an enormous variety of text types from the genre level to the sentence level” (p. 59). Therefore, in their study investigating the degree to which these indices have a role in predicting the quality of essays, they used 26 linguistic indices of cohesion from Coh-Metrix. Similarly, McNamara et al. (2010) point out that Coh-Metrix which is a tool presenting a great variety of linguistic indices for the automatic analysis of text comprehension uses lexicons, latent semantic analysis (LSA), and many other linguistic components and thus meets the needs of researchers who seek a computational linguistic analysis of texts to measure text cohesion and text difficulty in terms of various linguistic features such as word, sentence, paragraph, and discourse dimensions. Furthermore, their study comparing the outcomes of Coh-Metrix indices with two commonly used readability indices – Flesch Kincaid Grade Level and Flesch Reading Ease added evidence on validation of Coh-Metrix as a tool to assess cohesion.

Through the studies on cohesion, we reached the conclusion that we can measure text easibility indices through the automated tool “The Coh-Metrix Common Core Text Ease and Readability Assessor (T.E.R.A.)” which is less time-consuming and more practical. On the other hand, the coherence that is more subjective and exists in the mind of the reader will be best assessed using an analytical rubric that involves a specific dimension of coherence. This study which was carried out to see whether increasing complexity of writing task along with resource-directing variables as presence or absence (+/-) of few elements affect cohesion and coherence of EFL learners’ narrative writing will probably contribute to filling the research gaps in the effects of task complexity, particularly in terms of written task performance. As also pointed out by Jackson and Suethanapornkul (2013) in their synthesis and meta-analysis study of research on task complexity, writing task performance has been rarely investigated in terms of task complexity although writing production of learners is valuable to obtain more reliable and concrete results. In line with these aims, the findings of the study were presented and discussed to provide responses to the following research questions:

1. Does increasing the complexity of a task affect the text easibility of EFL learners’ narrative writing?
 - a. Does increasing the complexity of a task affect the text narrativity of EFL learners’ narrative writing?
 - b. Does increasing the complexity of a task affect the syntactic simplicity of EFL learners’ narrative writing?
 - c. Does increasing the complexity of a task affect word concreteness of EFL learners’ narrative writing?
 - d. Does increasing the complexity of a task affect referential cohesion of EFL learners’ narrative writing?
 - e. Does increasing the complexity of a task affect deep cohesion of EFL learners’ narrative writing?
2. Does increasing the complexity of a task affect coherence of EFL learners’ narrative writing?

Methodology

Research design

A one-group pretest- posttest design, one of the poor experimental designs, in which just a single group is “measured or observed not only after being exposed to a treatment of some sort but also before” was employed in this study (Frankel, Wallen & Hynun, 2012, p.269).

Participants

Forty-one freshmen (33 female and 8 male students) studying at the ELT department of a state university in Turkey and whose ages ranged between 19-28 years participated in the study. The proficiency levels of students were generally intermediate whereas there were also a few students at an advanced level, as also understood particularly from their scores of general writing achievement. Moreover, although 35 students were in their second year as they had prep-class (included writing course) in the previous year, 6 of the participants were newcomers who did not take a writing course before. Before collecting data, basic training for writing essays was provided within the scope of the writing course. Before performing the tasks involved in this study, students were asked to write paragraphs and essays to assess their levels and proficiency in writing. Based on the results, it was regarded that students were homogenous in terms of writing proficiency.

Each participant was involved in narrative writing tasks—simple and complex. In all, 82 essays were involved in the analysis process. All participants were requested to sign a consent form allowing their written production to be used for research purposes. Furthermore, ethics committee approval was also obtained to show that this study complied with ethical standards.

Operationalizations of task complexity

Task complexity was operationalized at two levels as simple and complex based on one of the resource-dispersing variables of Robinson's the Triadic Componential Framework, +/- task structure. Accordingly, the first writing task was identified as complex as a consequence of loose task structure which required more cognitive demand in performance of the task. The second task carried out two weeks later was described as simple since it involved a tight task structure and thus was assumed to need less cognitive demand to design and perform the writing process.

Data collection procedure

Data were collected during a writing course for students studying at the ELT department of a state university. Before starting to collect the data for the study, the instructor, also one of the researchers of this study, provided them with basic education about the writing process and asked them to write some sample paragraphs based on the knowledge they gained and one essay considering the features of an effective essay. After learning the features of narrative writing theoretically, the students were required to produce a narrative essay. Accordingly, they were first shown a picture and asked to analyze it for five minutes. They were then given an hour to create a story based on the picture and thus perform their complex writing task (-task structure/-TS). In another lesson, the students were given a sheet involving 16 related pictures in order and create a story following the order of the pictures for the simple task (+ task structure/ +TS).

Data analysis

Text easibility

The texts produced by the students were analyzed by the Coh-Metrix Common Core Text Ease and Readability Assessor (T.E.R.A.) which is designed to evaluate the easibility and readability of texts and thus provide valuable information about the characteristics of effective texts. Five components of text easibility used in this study were briefly explained below (McNamara et al., 2014, p. 85):

Narrativity: It deals with whether the text is story-like and involves familiar words, world knowledge, and conversation. Therefore, the higher a text is in narrativity, the easier it is.

Syntactic simplicity: It “reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar structures that are less challenging to process”. It is measured based on a variety of indices such as the number of clauses for sentences, the number of words in each sentence, and the number of words used before the main verb of the clause. Accordingly, the texts with fewer words or clauses in sentences are syntactically simple and thus easy to comprehend.

Word concreteness: It shows whether the words are concrete and meaningful which helps the reader to visualize and comprehend the text.

Referential cohesion: Referential cohesion in a text illustrates the extent of connections that link the ideas together to help the reader process easily.

Deep cohesion: A text with high deep cohesion contains causal and intentional cohesive devices where they are required to establish relationships among ideas, events, or actions since these devices enable the reader to understand the relationships more deeply and coherently.

Coherence

Since coherence is a subjective feature of writing and simply what the reader grabs from the text (Crossley et al., 2016), we analyzed it through the same analytic rubric also used for the analysis of the students' general writing achievement. The rubric we used included a separate section to evaluate coherence under the name of organization and coherence. As in other sections of the rubric, the scores ranged between 1 and 5. If a text takes the maximum score of 5 for coherence, it means that the text uses a logical structure regarding the purpose, audience, subject of the paper, utilizes true and enough transitions to build a clear connection between sentences, and lead the reader to comprehend the chain of reasoning or progression of ideas. On the contrary, if it is given the minimum score of 1, the text lacks organization, coherence, and transitions.

Statistical Analysis

In order to compare the results of the simple task and complex task performed by the same learners, data were firstly analyzed in terms of normality. Since they were not normally distributed, a Wilcoxon signed-rank test, a nonparametric test, was applied to assess the difference between the two tasks in terms of dependent variables, text easibility indices and coherence. When a statistically significant difference was obtained, its effect size was also calculated (Cohen, 1988)

Reliability and Validity

In the light of the literature, it was agreed to use T.E.R.A that provides an automated evaluation of a text. Furthermore, the validation of indices in this program was verified by McNamara et al. (2010) that those indices measure what is expected to be measured and can be compatible with all types of data regarding human performance.

The other dimensions in the current study, coherence not rated by the computer analysis was also evaluated by another rater to ensure the reliability of coding data. Firstly, three raters were trained about the analytical rubric used in the evaluation and then asked to evaluate 30 essays chosen at random to see whether there was consistency between their results. Reliability test results showed a high level of reliability (Cronbach's alpha coefficient=.87). Based on the results of the reliability test showing consistency in three raters' results, two of the raters completed rating the rest of the essays. The overall results of the two raters were again tested for reliability. Although a high level of inter-rater reliability (Cronbach's alpha coefficient=.88) was found, the essays for which raters had two or more-point-difference in rating scores were determined and discussed by the researchers. As a result of the reevaluation of these essays by the two raters, high-level inter-rater reliability (.89) with a Cronbach's alpha coefficient was reached.

Results

The results of a Wilcoxon signed-rank test conducted to see whether increasing the complexity of writing tasks along the task structure led to a significant difference in the text easibility indices and coherence of EFL learners' narrative writing is displayed in tables and explained in detail under each dependent variable. The analysis results in T.E.R.A. that provide some explanation for these indices explained in the previous section under the subheading of data analysis are also benefited to interpret the results in the tables.

Narrativity

Table 1. Wilcoxon signed-test results for the narrativity of the tests

Narrativity for the simple task - Narrativity for the complex task	
Z	-.57
Asymp. Sig. (2-tailed)	.57

According to the results illustrated in Table 1, the difference between complex and simple tasks is not statistically significant. It is also clear that the narrativity of students' narrative essays is not affected by the increase in the cognitive complexity of the tasks ($Z=-.57$, $p=.57>0.05$). Indeed, the median narrativity rating is higher in the complex task (median=95.0) compared to that in the simple task (median= 96.0).

Accordingly, it is clear that texts high in narrativity are more story-like and probably have more familiar words. Therefore, it is typically easier to understand those texts. In line with the statistical results, it can be concluded that the narrative essays produced based on a picture in a loose task structure (complex task) are more-story like and thus easier to understand

Syntactic Simplicity

Table 2. Wilcoxon signed-test results for the syntactic simplicity

Syntactic Simplicity for the simple task - Syntactic Simplicity for the complex task	
Z	-1.77
Asymp. Sig. (2-tailed)	.08

Table 2 illustrates the results for the difference between two tasks in syntactic complexity of essays produced by the learners. Although the essays produced in the simple task (median=88) had more syntactic

simplicity than those in the complex task (median=83) did, the difference was not found statistically significant ($Z=-1.77, p=.08>0.05$).

Accordingly, students produced more syntactically simple essays in their simple tasks. In other words, narrative essays in the simple task had more simple sentence structures and thus are easier to process.

Word Concreteness

Table 3. Wilcoxon signed-test results for the word concreteness

Word Concreteness for the simple task - Word Concreteness for the complex task	
Z	-3.32
Asymp. Sig. (2-tailed)	.00

A Wilcoxon signed-rank test the result of which is displayed in Table 3 shows that the complexity of the task yielded to a statistically significant difference in word concreteness of learners' narrative essays ($Z=-3.32, p=.00<0.05, r=0.51$). According to the results of this test, students used more concrete words in their essays they produced in the simple task (median=83) than they did in their essays produced in the complex tasks (median=61). The complexity of the task affects word concreteness at a significance level. It can be also concluded that the use of more concrete words in a simple task makes their narrative essays easier to visualize and comprehend.

Referential Cohesion

Table 4. Wilcoxon signed-test results for the referential cohesion

Referential Cohesion for the simple task - Referential Cohesion for the complex task	
Z	-2.22
Asymp. Sig. (2-tailed)	.03

According to Table 4, referential cohesion in the written production of the simple task significantly differs from that in essays of the complex task ($Z=-2.26, p=.03<0.05, r=0.34$). In addition, the simple task essays had stronger referential cohesion (median=60) compared to the essays of the complex task (median=46). The effect of the task complexity on referential cohesion of written production is statistically significant at a medium level.

Based on these results, it is clear that students' narrative essays in the complex task had little overlap in words and ideas and thus required the reader to make inferences. Therefore, these essays are more difficult to comprehend than those produced in a simple task.

Deep Cohesion

Table 5. Wilcoxon signed-test results for the deep cohesion

Deep Cohesion for the simple task - Deep Cohesion for the complex task	
Z	-1.22
Asymp. Sig. (2-tailed)	.22

As shown in Table 5, the difference between the complex and simple tasks in deep cohesion was not statistically significant ($Z=-1.22, p=.22>0.05$). Although students produced narrative essays in their complex tasks which are richer in deep cohesion (median=78) compared to their essays in the simple task (median=70). In this sense, essays in the complex tasks can be said to have more connecting words to clarify the relationships between events, ideas, and information and thus are easier to comprehend due to this added support.

Coherence

Table 5. Wilcoxon signed-test results for the coherence

Coherence for the simple task - Coherence for the complex task	
Z	-.83
Asymp. Sig. (2-tailed)	.41

Although the components of text easibility were analyzed by an automated program, T.E.R.A, essays were assessed by two different raters in terms of coherence using an analytic rubric. Since there was interrater reliability between the scores of the raters, the mean of their scores were taken and statistically analyzed. According to the results displayed in Table 6, there is no statistically significant difference between narrative essays of students in the simple task and those in the complex task in terms of coherence ($Z=-2.36$, $p=.41>0.05$). Indeed, the median coherence rating was 3.5 for both complex and simple tasks. Therefore, the complexity of the tasks does not affect the coherence of the texts produced by the learners.

Discussion and Conclusion

This study was investigated to see whether increasing the complexity of a writing task had a significant effect on the easibility of the texts learners produced as a result of their task performance and also on its coherence. Two writing tasks were designed at two levels as complex and simple according to +/- (tight/loose) task structure, one of the resource-dispersing variables in Robinson's framework. Therefore, the participants of this study were asked to produce a narrative essay for the complex task involving one picture about which they were required to narrate a story and also another one for the simple task involving 16 pictures on which they would ground their narration according to the order of the pictures.

Their essays were analyzed by an automated program, T.E.R.A, for the text easibility including five indices as narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion, and by an analytic rubric for the coherence of the texts. T.E.R.A. suggests that the texts higher in these indices are easy to comprehend.

The results of this study pointed out that students' narrative essays in the simple task had more syntactic simplicity although they were low in narrativity. That is, their essays in the complex task were more story-like and thus easier to comprehend but had more complex structures than their essays in the simple task did. However, although differences in these indices were not statistically significant, the complex task had more syntactically complex texts. Similarly, the studies, based on the Trade-off Hypothesis, found that increasing the complexity of the task led to the appearance of more syntactically complex structures in task performance (Salimi et al., 2011; Tavakoli, 2009). However, Tavakoli and Foster (2008) pointed out that syntactic complexity was high in narrative tasks with background information which was identified as the simple task. Furthermore, some studies obtained similar results that task complexity did have no significant effect on the syntactic complexity of the task outcome (Adams et al., 2015; Frear & Bitchener, 2015).

The current study also found that the texts in the simple task had more concrete words, which enabled the reader to visualize what was said and thus to easily comprehend them. However, their narrative essays produced in the complex task involved more abstract words which are assumed to be more difficult to comprehend. These results seemed to support the results of Frear and Bitchener (2015) showing that the use of complex words was positively related to the complexity of the task. Furthermore, Ong and Zang (2010) revealed that increasing the complexity of the task along the provision of ideas and macro-structure affected lexical complexity of EFL learners' argumentative writing whereas task complexity along draft availability had no impact on lexical complexity. Similarly, some studies provided contradictory results that the lexical complexity of participants' task performance was not influenced by the task complexity (Kuiken & Vedder, 2007).

In addition, students' narrative essays in the simple task had more overlap in words and ideas which helped the readers to make inferences and thus easily understand the text. On the other hand, the essays produced in the complex task were richer in deep cohesion and had more connecting words to clarify the relationships between events, ideas, and information although the difference between the two tasks in deep cohesion was not significant. Similarly, no significant difference was observed in the coherence of the texts produced in the simple and complex tasks. It is not so easy to discuss these results in the light of other studies since no study investigating task complexity in terms of cohesion and coherence was encountered. However, when coherence and cohesion are regarded as components of fluency in writing, the studies investigating task complexity in terms of +/- planning suggested that planning time in task positively affected the fluency of written production (Ellis & Yuan, 2004; Ong & Zang, 2010).

Based on these results, clear conclusions cannot be suggested about the effect of task complexity on text easibility since the results show differences in indices. Whereas the essays in the simple task were higher in some indices suggesting that they were easier to comprehend, the essays produced in the complex task had higher scores in other indices. However, it was seen that students produced easier texts in simple tasks, though not statistically significant in all indices. Therefore, conducting more writing tasks at different levels of task complexity may yield more reliable and concrete results for the text easibility. Furthermore, other measures such as complexity, accuracy, and fluency concurrently in addition to those involved in this study can be employed to have further insight into the effects of task complexity on written task performance..

Acknowledgments or Notes

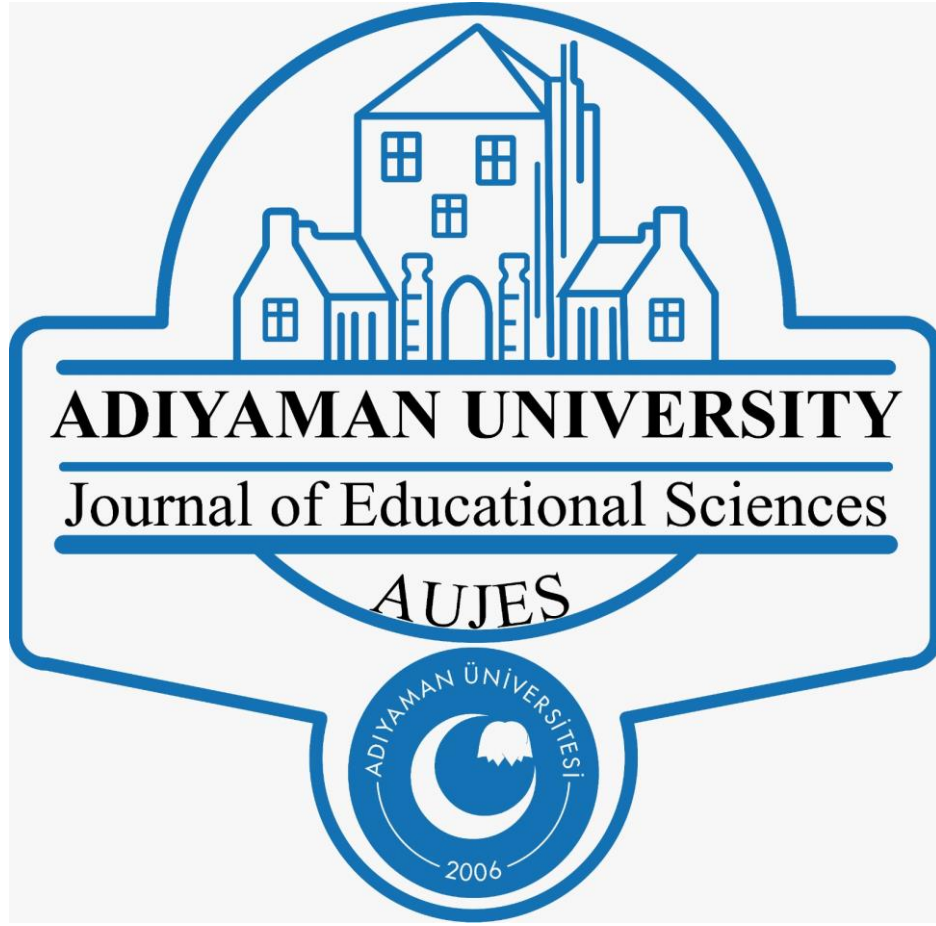
All the data of this study were collected during the PhD study of the first author and partially involved in her doctoral dissertation under the supervision of the second author. We would also like to thank TÜBİTAK (The Scientific and Technological Council of Turkey) for financial support during the process of the first author's PhD study.

References

- Adams, R., Nik Mohd Alwi, N. A., & Newton, J. (2015). Task complexity effects on the complexity and accuracy of writing via text chat. *Journal of Second Language Writing*, 29, 64-81. <https://doi.org/10.1016/j.jslw.2015.06.002>
- Ahmadian, M. J., Abdolrezapour, P., & Ketabi, S. (2012). Task difficulty and self-repair behaviour in second language oral production. *International Journal of Applied Linguistics*, 22(3), 310-330. <https://doi.org/10.1111/j.1473-4192.2012.00313.x>
- Arslanyilmaz, A. (2013). Computer-assisted foreign language instruction: task based vs. form focused. *Journal of Computer Assisted Learning*, 29(4), 303-318. <https://doi.org/10.1111/jcal.12003>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1-16. <https://doi.org/10.1016/j.jslw.2016.01.003>
- Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, 30(4), 474-509. <https://doi.org/10.1093/applin/amp042>
- Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26, 59-84. <https://doi.org/10.1017/S0272263104261034>
- Frankel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. Mc Graw Hill.
- Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing*, 30, 45-57. <https://doi.org/10.1016/j.jslw.2015.08.009>
- Genç, Z. S. (2012). Effects of strategic planning on the accuracy of oral and written tasks in the performance of Turkish EFL learners. In A. Shehadeh & C. A. Coombe (Eds.), *Task-Based Language Teaching in Foreign Language Contexts* (pp. 67-88). John Benjamins Publishing.
- Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. Longman.
- Jackson, D. O., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning*, 63(2), 330-367. <https://doi.org/10.1111/lang.12008>
- Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and task performance in as second language* (pp. 143-164). Amsterdam: John Benjamins.
- Kim, N. (2020a). The effects of different task sequences on novice L2 learners' oral performance in the classroom. *Language Teaching Research*, <https://doi.org/10.1177/1362168820937548>
- Kim, N. (2020b). Reconsidering task complexity through different planning and proficiency in l2 written tasks. *응용언어학*, 36(1), 65-95. <https://doi.org/10.17154/kjal.2020.3.36.1.65>
- Kırkgöz, Y. (2014). Task-based language teaching. In S. Çelik (Ed.), *Approaches and principles in English as a foreign language (EFL) education*. Egiten Kitap.
- Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, 12(2), 108-128. <https://doi.org/10.1016/j.asw.2007.07.002>
- Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, 20(2), 148-161. <https://doi.org/10.1016/j.jslw.2011.02.001>
- Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen, M. d. P. Garcia Mayo, & J. Cenoz (Eds.), *Eurosla Yearbook* (Vol. 5, pp. 195-222). John Benjamins.

- Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *IRAL - International Review of Applied Linguistics in Language Teaching*, 45(3). <https://doi.org/10.1515/iral.2007.012>
- Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, 17(1), 48-60. <https://doi.org/10.1016/j.jslw.2007.08.003>
- Larsen-Freeman, D., & Anderson, M. (2011). *Techniques and principles in language teaching* (Third Edition ed.). Oxford University Press.
- Lee, I. (2002). Teaching coherence to ESL students-a classroom inquiry. *Journal of Second Language Writing*, 11, 135-159. [https://doi.org/160-3743/02/\\$](https://doi.org/160-3743/02/$)
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2009). Linguistic Features of Writing Quality. *Written Communication*, 27(1), 57-86. <https://doi.org/10.1177/0741088309351547>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
- McNamara, D. S., Louwerson, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4), 292-330. <https://doi.org/10.1080/01638530902959943>
- Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences-A classroom-based study. *The Modern Language Journal*, 95(Supplementary Issue), 162-181. <https://doi.org/10.1111/j.1540-4781.2011.01241>
- Richards, J. C., & Rodgers, T. (2001). *Approaches and methods in language teaching*. Cambridge University Press.
- Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27-57. <https://doi.org/10.1093/applin/22.1.27>
- Robinson, P. (2003). The Cognition Hypothesis, task design and task-based language learning. *Second Language Studies*, 21(2), 45-107.
- Robinson, P. (2005). Cognitive complexity and task sequencing: A review of studies in a Componential Framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 43(1), 1-33. <https://doi.org/10.1515/iral.2005.43.1.1>
- Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects of L2 speech production, interaction, uptake and perceptions of task difficulty. *Iral-International Review of Applied Linguistics in Language Teaching*, 45, 161-176. <https://doi.org/10.1515/iral.2007.009>
- Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *Iral-International Review of Applied Linguistics in Language Teaching*, 45(3), 161-176. <https://doi.org/10.1515/iral.2007.007>
- Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing*, 28, 1-19. <https://doi.org/10.1016/j.jslw.2015.02.001>
- Salimi, A., & Dadaspour, S. (2012). Task complexity and SL development: Does task complexity matter? *Procedia - Social and Behavioral Sciences*, 46, 726-735. <https://doi.org/10.1016/j.sbspro.2012.05.189>
- Salimi, A., Dadaspour, S., & Asadollahfam, H. (2011). The effect of task complexity on EFL learners' written performance. *Procedia - Social and Behavioral Sciences*, 29, 1390-1399. <https://doi.org/10.1016/j.sbspro.2011.11.378>
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1), 1-14. <https://doi.org/10.1017/s026144480200188x>
- Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Eds.) *Processing perspectives on task performance* (pp. 211-260). Johns Benjamin Publishing Company.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49, 93-120. <https://doi.org/10.1111/1467-9922.00071>

- Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). Cambridge University Press.
- Tavakoli, P. (2009). Assessing L2 task performance: Understanding effects of task design. *System*, 37(3), 482-495. <https://doi.org/10.1016/j.system.2009.02.013>
- Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 58(2), 439-473. <https://doi.org/10.1111/j.1467-9922.2011.00642.x>
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language*. John Benjamins.
- Todd, R. W., Thienpermpool, P., & Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. *Assessing Writing*, 9(2), 85-104. <https://doi.org/10.1016/j.asw.2004.06.002>
- Watson Todd, R., Khongput, S., & Darasawang, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing Writing*, 12(1), 10-25. <https://doi.org/10.1016/j.asw.2007.02.002>
- Willis, J. (1996). *A framework for task-based learning*. Longman.
- Yang, W. (2014). *Mapping the relationships among the cognitive complexity of independent writing tasks, L2 writing quality, and complexity, accuracy and fluency of L2 writing*. (Unpublished Dissertation Thesis). Georgia State University, Retrieved from http://scholarworks.gsu.edu/alesl_diss/29/
- Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53-67. <https://doi.org/10.1016/j.jslw.2015.02.002>
- Yıldız, M., & Yeşilyurt, S. (2017). Effects of task planning and rhetorical mode of writing on lexical complexity, syntactic complexity, and overall writing quality of EFL writers' task performance. *Journal of Language and Linguistic Studies*, 13(2), 440-464.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24(1), 1-27.



Article History

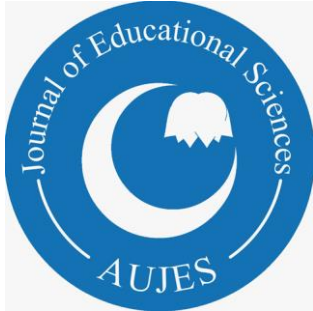
Received: 23.05.2021

Received in revised form: 16.06.2021

Accepted: 23.06.2021

Available online: 31.12.2020

Article Type: Research Article




ADIYAMAN UNIVERSITY
Journal of Educational Sciences
(AUJES)

<https://dergipark.org.tr/tr/pub/adyuebd>

Sample Size Determination and Optimal Design of Randomized/Non-equivalent Pretest-posttest Control-group Designs

Metin Buluş¹

¹Adiyaman University, Department of Educational Sciences, Adiyaman, Turkey 

To cite this article:

Bulus, M. (2021). Sample size determination and optimal design of randomized/non-equivalent pretest-posttest control-group designs. *Adiyaman University Journal of Educational Sciences*, 11(1), Page 48-69.

Sample Size Determination and Optimal Design of Randomized/Non-equivalent Pretest-posttest Control-group Designs

Metin Bulus^{1*}

¹Adiyaman University, Department of Educational Sciences, Adiyaman, Turkey

Abstract

A recent systematic review of experimental studies conducted in Turkey between 2010 and 2020 reported that small sample sizes had been a significant drawback (Bulus & Koyuncu, 2021). A small chunk of the studies in the review were randomized pretest-posttest control-group designs. In contrast, the overwhelming majority of them were non-equivalent pretest-posttest control-group designs (no randomization). They had an average sample size below 70 for different domains and outcomes. Designing experimental studies with such small sample sizes implies a strong (and perhaps an erroneous) assumption about the minimum relevant effect size (MRES) of an intervention; that is, a standardized treatment effect of Cohen's $d < 0.50$ is not relevant to education policy or practice. Thus, an introduction to sample size determination for randomized/non-equivalent pretest-posttest control group designs is warranted. This study describes nuts and bolts of sample size determination (or power analysis). It also derives expressions for optimal design under differential cost per treatment and control units, and implements these expressions in an Excel workbook. Finally, this study provides convenient tables to guide sample size decisions for MRES values between $0.20 \leq \text{Cohen's } d \leq 0.50$.

Keywords: pretest-posttest, experimental design, random assignment, non-equivalent control-group design, sample size, power analysis, optimal design

Introduction

One crucial question in education policy and practice is whether a program, product, or service produces favorable outcomes. The first step to answering such a research question is to solicit funding from stakeholders in a grant proposal to cover research expenses. The description of the research design in the grant proposal should convince stakeholders (and peers in the publication process) that the study employs rigorous methodological procedures and that the sample is not fundamentally flawed to produce biased or inconclusive results.

In education policy research, experiments are indispensable research designs that can establish a cause-effect relationship between an independent variable (e.g., receiving a program, product, or service) and an outcome variable (e.g., academic achievement) (Campbell & Stanley, 1963; Cook et al., 2002; Mostseller & Boruch, 2004). An experiment's main characteristic is that researchers can manipulate the independent variable to isolate its effect from unobserved confounders. In the simplest form, this is achieved via randomly assigning subjects in the sample into the treatment and control groups. Randomization assures that effects of unobserved confounders on the outcome – a significant threat to the internal validity of experiments – are canceled out on average (Campbell & Stanley, 1963; Cook et al., 2002; Mostseller & Boruch, 2004). In this case, treatment and control groups do not systematically differ (especially in large samples). This type of design is referred to as a true experiment.

However, randomization is not always feasible. For example, in education research, it is common to assign entire clusters to treatment and control groups (e.g., classrooms) without randomization. In this case, the treatment effect may be contaminated with unobserved confounders. In other words, treatment and control groups may systematically differ. This type of design is a non-equivalent design (see Campbell & Stanley, 1963; Oakes & Feldman, 2001) and categorized as one of the weak experiments in the literature. Nonetheless, weak experiments can be manipulated to mimic true experiments via matching subjects on the pretest or covariates (Fraenkel et al., 2011; see also Campbell & Stanley, 1963). This type of design is referred to as a quasi-experiment.

* Corresponding Author: *Metin Bulus*, bulusmetin@gmail.com

Recent reviews of experiments in Turkey indicated that they had inadequate sample sizes (e.g., Bulus & Koyuncu, 2021; Yildirim et al., 2019). Overwhelming majority of the reviewed experiments in Bulus and Koyuncu (2021) and Yildirim et al. (2019) were small-scale weak or quasi-experiments. Most of them were based on convenience sampling where intact classrooms received the treatment or control protocols (often, one classroom in each). Average sample size was 70 for experiments reviewed in Bulus and Koyuncu (2021) and was 54 for those reviewed in Yildirim et al. (2019). Such small sample sizes imply a strong (and perhaps an erroneous) assumption about an intervention's minimum relevant effect size (MRES) before an experiment is undertaken. In other words, a standardized treatment effect of Cohen's $d < 0.50$ is not relevant to education policy or practice. MRES is related to the "What is the minimum treatment effect that is meaningful and relevant to education policy and practice?" question, and its value should carefully be justified.

The result of a small-scale experiment is sometimes "too good to be true." There are several potential sources of bias inherent to small-scale experiments. For example, the treatment effect in a small-scale experiment could be overestimated due to publication bias (Hedges, 1992; Vevea & Hedges, 1995), small study effect (Sterne et al., 2000), overfitting problem where the model picks up noise (Yarkoni, 2017), teaching treatment group to perform superior on the researcher developed test, shorter pretest-posttest interval (Slavin, 2008), baseline incomparability, classroom or school confounding, researcher bias such as choosing the more able subjects for the treatment group, or a combination of them.

Bulus and Koyuncu (2021) reported large treatment effects for 106 experiments targeting cognitive outcomes (Cohen's $d = 1.02$, on average) and for 81 experiments targeting affective outcomes (Cohen's $d = 1.01$, on average). The authors did not adjust effect size estimates for the pretest. Yildirim et al. (2019) also reported large treatment effects of learning strategies on academic achievement based on a random-effect meta-analysis of 28 experiments (Cohen's $d = 1.21$, on average). The authors did not explicitly state whether they adjusted effect size estimates for the pretest. We do not know whether the effects reported in Bulus and Koyuncu (2021) and Yildirim et al. (2019) were artifacts (due to several potential sources of bias mentioned earlier) or actual effects. Effects sizes of this magnitude, if considered artifacts, cannot be explained by failure to adjust for the pretest alone. If these are actual effects, it begs why these programs are not scaled-up.

One effective way to decipher this ambiguity and ameliorate potential sources of bias mentioned earlier is to conduct an experiment with sufficient sample size. A sufficient sample size would allow the experiment to detect a minimum effect relevant to policy and practice with sufficient statistical power (probability to detect an effect when there is an effect in the underlying population). This study mainly describes formulas and software to determine sample size for randomized pretest-posttest control-group design (true experiment) and non-equivalent pretest-posttest control-group design (weak experiment). It derives expressions for the optimal design of true experiments under differential cost per treatment and control units, and provides a convenient Excel workbook for this purpose (Optimal Design: <https://osf.io/uerbw/download>). Moreover, it provides convenient tables to guide sample size decisions for MRES values between $0.20 \leq \text{Cohen's } d \leq 0.50$ (Appendix and Supplement: <https://osf.io/t2as3/download>).

In what follows, first, the approximate standard error of the treatment effect for several types of experimental designs will be described. Approximate standard errors are required for power analysis routines. Suppose approximate standard errors are formulated in terms of known design parameters such as MRES, treatment group allocation rate, and explanatory power or covariates. Then, one can conveniently find the minimum required sample size (MRSS) for true and weak experiments given design parameters. Second, illustrative examples are provided to find MRSS depending on common design characteristics. Finally, key points are discussed and summarized

Approximate Standard Error Formulas for Power Analysis

To answer the crucial question of "At least how many participants are needed in treatment and control groups to detect an effect that is relevant to policy and practice?" one will need to have a guestimate for the standard error of the treatment effect. Fortunately, there are many important studies in this line of work. Several scholars derived expressions for approximate standard errors, which is a function of the known design parameters such as total sample size, treatment group allocation rate, and explanatory power of covariates (e.g., Bloom, 2006, Dong & Maynard, 2013; Oakes & Feldman, 2001). Expressions for approximate standard errors considering true and weak experiments will be described momentarily.

Approximate standard error expressions presented in this study apply to several experimental designs described in Campbell and Stanley (1963) and Fraenkel et al. (2011) when Analysis of Variance (ANOVA) or Analysis of Covariance (ANCOVA) model is the method of choice. Randomized posttest-only control-group and randomized pretest-posttest control-group designs are categorized as true experiments (Campbell & Stanley, 1963; Fraenkel et al., 2011). Static-group comparison design (SCD; Campbell & Stanley, 1963) and static-group

pretest-posttest design (SPPD; Fraenkel et al., 2011) are categorized as weak experiments. SCD and SPPD designs are also known as non-equivalent designs. There is no guarantee that treatment and control groups are comparable at the baseline in non-equivalent designs (see Campbell & Stanley, 1963; Oakes & Feldman, 2001). This study adopts the latter naming convention; non-equivalent posttest-only control-group design for SCD and non-equivalent pretest-posttest control-group design for SPPD.

True Experiments

In a simple true experiment, subjects are randomly assigned into the treatment and control groups. While treatment group subjects benefit from a program, product, or service, no procedures are undertaken for the control group except for the administration of questionnaires. Information is collected at the baseline (e.g., pretest) to control bias resulting from baseline differences (mostly in small-scale weak or quasi-experiments) and improve the estimate's precision. In the end, outcomes between the two groups are compared to gauge the effectiveness of an intervention.

Randomized Pretest-posttest Control-group Design

The diagram of the randomized pretest-posttest control-group design is described below. R refers to the randomization process, X refers to the implementation of the treatment protocol, and O refers to the observation of the pretest before X or posttest after X.

Treatment group	R	O	X	O
Control group	R	O		O

The following procedures are followed in this type of design; (i) subjects are randomized into the treatment and control groups, (ii) a pretest questionnaire is administered before subjects receive treatment and control protocols, (iii) treatment and control group protocols are administered, and (iv) a posttest questionnaire is administered after subjects receive treatment and control protocols. Control group subjects could receive the business-as-usual approach or another intervention different from the treatment group. Data collected from this type of design can be analyzed via an ANCOVA model. The approximate standard error for the treatment effect takes the form of

$$SE(\widehat{ES}) = \sqrt{\frac{1 - R^2}{p(1 - p)n}} \quad 1$$

with $v = n - g - 2$ degrees of freedom (Bloom, 2006, p. 12; Dong & Maynard, 2013, p. 45). R^2 is the proportion of variance in the posttest explained by the pretest. p is the treatment group allocation rate (proportion of subjects in the treatment group). n is the total sample size in the treatment and control groups. g indicates the number of covariates ($g = 1$ when pretest is the only covariate). To determine MRSS for this type of design, one can use PowerUpR (Bulus et al., 2021) R package or PowerUp! (Dong & Maynard, 2021) Excel workbook for this purpose. These freeware will be described in the software illustration section momentarily.

Randomized Posttest-only Control-group Design

The diagram of the randomized posttest-only control-group design is described below.

Treatment group	R	X	O
Control group	R		O

The following procedures are followed in this type of design; (i) subjects are randomized into the treatment and control groups, (ii) treatment and control group protocols are administered, and (iii) a posttest questionnaire is administered after subjects receive treatment and control protocols. Similarly, control group subjects could receive the business-as-usual approach or another intervention different from the treatment group. Data collected from this type of design can be analyzed via an ANOVA model. Per G*Power 3.1 guide (p. 49), the approximate standard error for the treatment effect takes the form of

$$SE(\widehat{ES}) = \sqrt{\frac{1}{p(1 - p)n}} \quad 2$$

with $v = n - 2$ degrees of freedom. The remaining parameters are defined earlier. The relevant specification in G*Power is "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)." Note that when pretest information is not available in Equation 1 ($R^2 = 0$ & $g = 0$), it converges to Equation 2. Alternatively, one can use PowerUpR (Bulus et al., 2021) R package or PowerUp! (Dong &

Maynard, 2021) Excel workbook for this purpose. Note that in this case $R^2 = 0$ and $g = 0$ in PowerUpR and PowerUp!

Optimal Design of True Experiments

Conducting an experiment can be costly. Naturally, costs for the treatment group could be higher than costs for the control group. When the cost per subject in treatment and control groups is differential, it is desirable to sample less from the group with higher costs. Higher costs associated with the treatment group may emerge from new materials, new approaches to learning, hiring experts, and other overhead costs needed to develop and implement an intervention. Overhead costs for treatment and control groups can be divided by the number of subjects in each group and added to the subject-unique costs. In this case, each subject in the treatment and the control groups will be associated with differential costs. Therefore, it is reasonable to sample fewer subjects from the treatment group and more subjects from the control group. In what follows, analytic expressions are derived to find optimal p and n given total cost or budget.

Let C_{TRT} and C_{CTRL} be the cost per subject in treatment and control groups, respectively. Let also C_{TOT} be the total cost or budget. Total cost is the sum of the costs for treatment and control groups. Costs for the treatment and control groups can be expressed as the subject-level cost in each group multiplied by the number of subjects in each group. There are pn subjects in the treatment and $(1 - p)n$ subjects in the control group.

Then, the following equation can be defined as

$$C_{TOT} = pnC_{TRT} + (1 - p)nC_{CTRL} \tag{3}$$

Re-arranging Equation 3, n can be expressed as

$$n = \frac{C_{TOT}}{pC_{TRT} + (1 - p)C_{CTRL}} \tag{4}$$

Plugging Equation 4 for n in Equation 1, the squared standard error can be expressed as

$$SE(\widehat{ES})^2 = \frac{1 - R^2}{C_{TOT}} \left(\frac{pC_{TRT} + (1 - p)C_{CTRL}}{p(1 - p)} \right) \tag{5}$$

In order to find optimal p that minimizes the squared standard error in Equation 5, one needs to take the derivative of $SE(\widehat{ES})^2$ with respect to p as

$$\frac{\partial SE(\widehat{ES})^2}{\partial p} = \frac{1 - R^2}{C_{TOT}} \left(\frac{p^2C_{TRT} - (1 - p)^2C_{CTRL}}{p^2(1 - p)^2} \right) \tag{6}$$

Setting Equation 6 to zero and solving for p produces the optimal p as

$$p = \frac{\sqrt{C_{CTRL}}}{\sqrt{C_{TRT}} + \sqrt{C_{CTRL}}} \tag{7}$$

Equation 7 can be further simplified. Define cost ratio as $CR = C_{TRT}/C_{CTRL}$, then

$$p = \frac{1}{1 + \sqrt{CR}} \tag{8}$$

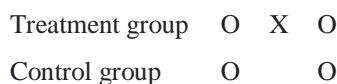
Equations 4 and 8 can be used to devise a randomized pretest-posttest control-group design optimally. First, one would need to have information on the cost ratio. Once the cost ratio is known, optimal p can be obtained using Equation 8. In the second step, optimal p can be plugged in Equation 4 to get an estimate for n .

Weak Experiments

Although weak experiments are presented here, they are not the first choice to produce knowledge for evidence-based practices. They should be preferred when randomization is not feasible. They are described below for interested readers.

Non-equivalent Pretest-posttest Control-group Design

The diagram of the non-equivalent pretest-posttest control-group design is described below.



The following procedures are followed in this type of design; (i) a pretest questionnaire is administered to subjects in two naturally occurring groups (e.g., classroom) before they receive treatment and control protocols, (iii) treatment and control group protocols are administered to these two groups, and (iv) a posttest questionnaire is administered after these two groups receive treatment and control protocols, respectively. Note that there is no randomization. Data collected from this type of design can also be analyzed via an ANCOVA model. The approximate standard error for the treatment effect is adapted from Oakes and Feldman (2001, p. 15) as

$$SE(\widehat{ES}) = \sqrt{\frac{1 - R^2}{p(1 - p)n(1 - R_{TX}^2)}} \quad 9$$

with $v = n - g - 2$ degrees of freedom. Unlike earlier designs, R_{TX}^2 is the squared point-biserial correlation between the pretest variable and the treatment indicator. It represents the proportion of variance in the pretest explained by the treatment indicator.

Non-equivalent Posttest-only Control-group Design

The diagram of the non-equivalent posttest-only control-group design is described below.

Treatment group	X	O
Control group		O

The following procedures are followed in this type of design; (i) treatment and control group protocols are administered to two naturally occurring groups, and (ii) a posttest questionnaire is administered after these two groups receive treatment and control protocols, respectively. There is no randomization. Data collected from this type of design can also be analyzed via an ANOVA model. The approximate standard error for the treatment effect can be obtained via re-expressing Equation 9 as

$$SE(\widehat{ES}) = \sqrt{\frac{1}{p(1 - p)n(1 - R_{TX}^2)}} \quad 10$$

with $v = n - 2$ degrees of freedom. One could rightfully argue that R_{TX}^2 does not apply to this formulation because pretest information is not collected. Although pretest information is not collected, differences between treatment and control groups at the baseline would affect standard error of the treatment effect. Thus, it would be a good practice to have a guesstimate for R_{TX}^2 and determine sample size accordingly. Other parameters are defined earlier.

Sample Size Determination in True Experiments

In this section, the nuts and bolts of sample size determination in randomized pretest-posttest control-group design will be described. First, in the software illustrations section, PowerUpR and PowerUp! will be used to determine the sample size for a hypothetical intervention. Second, in the optimal design section, a step-by-step guide will be provided to optimally design a hypothetical intervention, along with the description of the Optimal Design Excel workbook accompanying this article. Finally, in the table illustration section, the relevant table in the Appendix will be used to determine sample size without using any software packages.

Software Illustrations

There are a few points to consider when determining the minimum required sample size (MRSS):

- Type I error rate can be defined as the probability of finding a treatment effect in the sample when there is no effect in the underlying population. It is usually specified as 05%, the default value in PowerUpR (`alpha = .05`).
- Power rate can be defined as the probability of finding a treatment effect in the sample when there is an effect in the underlying population. It is usually defined as 80% in social science, which is the default value in PowerUpR (`power = .80`).
- Whether the hypothesis test is one-tailed or two-tailed. Generally, a two-tailed hypothesis test is performed assuming that the intervention could either be beneficial or detrimental, the default value in PowerUpR (`two.tailed = TRUE`).
- The minimum relevant effect size (MRES), standardized according to Cohen's d . MRES is usually defined as 0.20 or 0.25 in education research, the default value in PowerUpR (`es = 0.25`). An MRES of 0.25 means that a minimum meaningful treatment effect bumps an average student's score by ten percentile points.

- Treatment group allocation rate (p) is defined as the proportion of subjects in the treatment group. Allocating half of the sample into the treatment group produces the smallest variance (or maximum power rate), which is the default value in PowerUpR ($p = .50$).
- The proportion of variance in the posttest explained by the pretest and other covariates (R^2). There is not much research in Turkey that provides R^2 values for planning experimental designs beyond Bulus and Koyuncu (2021). Brunner et al. (2018) analyzed PISA data for 81 countries, including Turkey, and provide design parameters for planning cluster-randomized trials. Their results apply to 15 years old students. If the interest is the explanatory power of socio-demographic variables for high school students, R^2 values reported for student-level can possibly be used. Socio-demographic variables explain a small amount of variance in academic achievement (Median $R^2 = .05$), affect and motivation (Median $R^2 = .01$), and learning strategies (Median $R^2 = .01$) at the student level. R^2 should rely on earlier literature or some existing data targeting the same outcome. The correlation between the pretest and the posttest tends to be higher with affective outcomes because, in comparison to cognitive outcomes, they tend to persist over time. This tendency for a stronger relationship manifests itself as higher R^2 values. In fact, for true experiments, Bulus and Koyuncu (2021, p. 32) reported that average values for affective and cognitive outcomes are $R^2 = .38$ and $R^2 = .22$, respectively ($r2 = .38$ or $r2 = .22$).

MRSS computations can be performed considering the information presented above. For this purpose, PowerUpR R package and PowerUp! Excel workbook will be used. These two freeware have the same naming conventions and employ the same algorithms to determine MRSS. Although these statistical packages are mainly designed for multilevel randomized experiments, they also include a function for randomized pretest-posttest control-group design under the "Individual Random Assignment" function or module.

First, we need to install the PowerUpR package in the R environment and load it into the current session using the following code (or any other package installment routine). GitHub code repository has the most recent version of the package. Once available, the package can also be downloaded from the CRAN repository.

```
require(devtools)
install_github("metinbulus/PowerUpR")
library(PowerUpR)
```

The function that allows MRSS computation in PowerUpR is `mrss.ira()`. Earlier versions of the PowerUpR package available on CRAN uses `mrss.ira1r1()` name. Considering R^2 from Bulus and Koyuncu (2021), MRSS for an intervention targeting to improve an affective outcome (e.g. affect and motivation) or a cognitive outcome (e.g. achievement) can be computed as:

```
# MRSS for an affective outcome
mrss.ira(alpha = .05, power = .80, two.tailed = TRUE,
          es = .25, g = 1, r2 = .38, p = .50)
# n = 313

# MRSS for a cognitive outcome
mrss.ira(alpha = .05, power = .80, two.tailed = TRUE,
          es = .25, g = 1, r2 = .22, p = .50)
# n = 394
```

If one opts for PowerUp! Microsoft Excel workbook, it should be downloaded from <https://www.causalevaluation.org/uploads/7/3/3/6/73366257/powerup.xlsm>. MRSS can be computed for each type of outcome using PowerUp! Module IRA with identical specifications (see Figures 1 and 2).

Model 1.0: Sample Size Calculator for Individual Random Assignment Designs (IRA)—Completely Randomized Controlled Trials		
Assumptions	Comments	
MRES = MDES	0.25	Minimum Relevant Effect Size = Minimum Detectable Effect Size
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1- β)	0.80	Statistical power (1-probability of a Type II error)
P	0.50	Proportion of the sample randomized to treatment: $n_T / (n_T + n_C)$
R ²	0.38	Percent of variance in the outcome explained by covariates
k*	1	The number of covariates used
M (Multiplier)	2.81	Automatically computed
N (Sample Size)	314	The number of individuals needed for the given MDES.

Figure 1. MRSS for an intervention targeting an affective outcome.

Model 1.0: Sample Size Calculator for Individual Random Assignment Designs (IRA)—Completely Randomized Controlled Trials		
Assumptions	Comments	
MRES = MDES	0.25	Minimum Relevant Effect Size = Minimum Detectable Effect Size
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1- β)	0.80	Statistical power (1-probability of a Type II error)
P	0.50	Proportion of the sample randomized to treatment: $n_T / (n_T + n_C)$
R ²	0.22	Percent of variance in the outcome explained by covariates
k*	1	The number of covariates used
M (Multiplier)	2.81	Automatically computed
N (Sample Size)	394	The number of individuals needed for the given MDES.

Figure 2. MRSS for an intervention targeting a cognitive outcome.

Considering MRSS result for an intervention targeting a cognitive outcome only, for example, one can report the power analysis procedure in a paragraph as follows:

For this randomized pretest-posttest control-group design, we assume that the pretest explains 22% of the posttest variance (Bulus and Koyuncu, 2021). We further assume that the hypothesis test is two-tailed, the Type I error rate is 5%, and the power rate is 80%. Under these conditions, based on PowerUpR (Bulus et al., 2021) or PowerUp! (Dong & Maynard, 2013), a sample of 394 subjects equally allocated to treatment and control groups is needed to detect an effect size as small as 0.25.

Readers are referred to Dong and Maynard (2013) for more complicated randomized experiments. In multisite randomized experiments, subjects are randomly assigned into the treatment and control groups within sites or blocks (Bloom, 2006; Dong & Maynard, 2013; Hedges & Rhoads, 2010; Raudenbush & Liu, 2000; Konstantopoulos, 2008a). In cluster-randomized experiments, entire clusters are randomly assigned into the treatment and control groups (Dong & Maynard, 2013; Hedges & Rhoads, 2010; Konstantopoulos, 2008b). Finally, in multisite cluster-randomized experiments, entire clusters are randomly assigned into the treatment and control groups within sites or blocks (Dong & Maynard, 2013; Hedges & Rhoads, 2010; Konstantopoulos, 2008a; Schochet, 2008; Spybrook, 2007). To estimate sample size in such complex experiments, researchers can use PowerUpR (also available through <https://powerupr.shinyapps.io/index/>) or PowerUp!.

Optimal Design under Differential Costs

The task of undertaking an experiment can be costly. Expenses can either be covered by the researcher or can be solicited from funding agencies. In either case, one can optimally allocate subjects into treatment and control groups if costs associated with treatment and control units are available. Optimal Design Excel workbook accompanying this article implements optimal design formulas presented in this study. The step-by-step approach to optimal design of randomized pretest-posttest control-group design is presented in Figures 3 to

6. The Optimal Design Excel workbook can also be used to optimally devise a randomized posttest-only control group design.

Assume that the reserved budget is 2000£, which cannot be increased (fixed budget). Further, assume that costs associated with each treatment and control unit are 20£ and 5£, respectively. Defining these values in the Optimal Design Excel workbook (yellow highlighted cells) produces a sample size of 200 with an allocation rate of $p = 0.33$ (see Step 1 in Figure 3).

Optimal Design of Randomized Pretest-Posttest Control-group Design under Differential Cost	
Parameters	Values
Total cost or budget	2,000£
Cost per treatment unit	20£
Cost per control unit	5£
Treatment group sampling rate (p)	0.33
Total sample size (n)	200

Figure 3. Step 1 in Optimal Design Excel workbook.

We know this is the best allocation that produces minimum variance (or maximum power) compared to alternative allocations under identical budget constraints. However, we still do not know what power rate this allocation will produce. The question is: What is the power rate for the optimal allocation rate ($p = .33$) and the sample size ($n = 200$)? Using PowerUpR, the power rate is computed as 47% (see Step 2 in Figure 4). If the total cost or budget is fixed at 2000£, this the best we can do.

<p>Step 2: Check the power rate in PowerUpR or PowerUp! given optimal p and n produced in Step 1. Specify other design parameters according to your study field. If the total cost or budget is fixed stop here.</p>	<pre>power.ira(alpha = .05, two.tailed = TRUE, es = .25, g = 1, r2 = .22, p = .33, n = 200) # Statistical power: # ----- # 0.465 # ----- # Degrees of freedom: 197 # Standardized standard error: 0.095 # Type I error rate: 0.05 # Type II error rate: 0.535 # Two-tailed test: TRUE</pre>
---	---

Figure 4. Step 2 in Optimal Design Excel workbook.

Suppose the total cost or budget is flexible. In that case, we can demonstrate that we opted for a cost-efficient allocation via exploring alternatives. The allocation rate does not change because it depends on per unit costs in treatment and control groups. The question is: What is the sample size and the total cost for a power rate of 80% given the optimal allocation rate ($p = .33$)? PowerUpR produces a sample size of 445, which will cost 4450£ (see Step 3 in Figure 5).

<p>Step 3: For the desired power rate (80%), find the required sample size given optimal p produced in Step 1. Then, re-estimate the total cost or budget.</p>	<pre>mrss.ira(alpha = .05, power = .80, two.tailed = TRUE, es = .25, g = 1, r2 = .22, p = .33) # n = 445</pre>						
	<table border="1"> <tr> <td>Total sample size (n)</td> <td>445</td> </tr> <tr> <td>Treatment group sampling rate (p)</td> <td>0.33</td> </tr> <tr> <td>Total cost or budget</td> <td>4,450£</td> </tr> </table>	Total sample size (n)	445	Treatment group sampling rate (p)	0.33	Total cost or budget	4,450£
Total sample size (n)	445						
Treatment group sampling rate (p)	0.33						
Total cost or budget	4,450£						

Figure 5. Step 3 in Optimal Design Excel workbook.

The next question is: What the sample size would have been for a power rate of 80% had we used a balanced allocation ($p = .50$) and how much would that cost? Had we used a $p = .50$ allocation rate instead of $p = .33$, we would have needed 394 subjects which would have cost 4925£ (see Step 4 in Figure 6).

<p>Step 4: For the desired power rate (80%), find the required sample size</p>	<pre>mrss.ira(alpha = .05, power = .80, two.tailed = TRUE, es = .25, g = 1, r2 = .22,</pre>
---	---

(n) with the balanced allocation rate (p = .50). Then, re-estimate the total cost or budget.	p = .50)	
	# n = 394	
	Total sample size (n)	394
	Treatment group sampling rate (p)	0.50
	Total cost or budget	4,925₺
Save	475₺	

Figure 6. Step 4 in Optimal Design Excel workbook.

Using an optimal allocation rate of $p = .33$, we save 475₺ while preserving a power rate of 80%. Researchers can decide whether they should spend the extra 475₺ and go with the more balanced sample. Sometimes, severely unbalanced samples produce unstable estimates in the analysis of variance. Readers are referred to Bulus & Dong (2021a) for the optimal design of more complicated experimental designs. Researchers can use the cosa R package (also available through <https://cosa.shinyapps.io/index/>; Bulus & Dong, 2021b) for this purpose.

Table Illustration

Tables 1A – 7A in the Appendix tabulate the main factors affecting MRSS. MRSS depends on whether the hypothesis test is two-tailed, the Type I error rate (α), the treatment group allocation rate (p), the explanatory power of the pretest (R^2), and the minimum relevant effect size (MRES). Tables are reproduced considering MRES values ranging from 0.20 to 0.50. There are two rationales for these specifications; an MRSS capable of detecting the MRES = 0.20 is an acceptable standard in education research. It is considered the minimum meaningful effect according to Cohen's d when there is no theory that guides MRES specification. Besides, Bulus and Koyuncu (2021) found that the average sample size for experiments conducted in Turkey between 2010 and 2020 is insufficient to detect MRES values of 0.50 and below. Type I error rate (α) specifications are based on common reporting guidelines in scholarly work (* $p < .05$, ** $p < .01$, and *** $p < .001$). The treatment group allocation rate (p) ranges from .35 to .50 because differential costs may impel researchers to draw more subjects from the control group. After all, it is less costly. $p = .50$ produces the smallest MRSS (minimum variance or maximum power) under no cost considerations. R^2 can be as high as .70, according to values reported in Hedges and Hedberg (2013). Thus, the explanatory power of the pretest (R^2) ranges from 0 to .70.

Table 2A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Experimental Design when MRES = 0.25

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power				PowerUPR						
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	1094	1092	765	711	657	602	548	494	439	385	331
	0.001	1.50	0.40	1036	1035	726	674	623	571	520	468	417	365	314
	0.001	1.22	0.45	1006	1004	704	654	604	554	504	454	404	354	304
	0.001	1.00	0.50	996	994	697	647	598	549	499	450	400	351	301
	0.01	1.86	0.35	710	708	497	461	426	391	355	320	285	250	214
	0.01	1.50	0.40	672	672	471	437	404	371	337	304	270	237	203
	0.01	1.22	0.45	652	651	457	424	392	359	327	295	262	230	197
	0.01	1.00	0.50	646	645	452	420	388	356	324	292	260	227	195
	0.05	1.86	0.35	438	436	306	284	262	241	219	197	175	154	132
	0.05	1.50	0.40	414	414	290	269	249	228	208	187	166	146	125
Two-tailed	0.001	1.86	0.35	1208	1206	846	785	725	665	605	545	485	425	365
	0.001	1.50	0.40	1144	1143	802	745	688	631	574	517	460	403	346
	0.001	1.22	0.45	1110	1109	778	722	667	612	557	502	446	391	336
	0.001	1.00	0.50	1100	1098	770	715	661	606	551	497	442	387	333
	0.01	1.86	0.35	826	824	578	537	496	455	414	373	332	291	249
	0.01	1.50	0.40	782	782	548	509	470	431	392	353	315	276	237
	0.01	1.22	0.45	760	758	532	494	456	418	381	343	305	267	230
	0.01	1.00	0.50	752	751	526	489	452	414	377	339	302	265	227
	0.05	1.86	0.35	554	554	388	361	333	306	278	250	223	195	168
	0.05	1.50	0.40	526	525	368	342	316	290	264	237	211	185	159
	0.05	1.22	0.45	510	509	357	332	306	281	256	230	205	180	154
	0.05	1.00	0.50	506	504	354	328	303	278	253	228	203	178	153

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. Allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the post-test explained by the pre-test variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t tests" and "Statistical test: Means: Difference between two independent means (two groups)".

Figure 7. Finding MRSS from tables in the Appendix (or Supplemental Excel workbook) based on MRES and R^2 specifications.

Let us find the MRSS for an experiment targeting an affective outcome. The default option for linear regression or t -test in SPSS and R produces p -values for a two-tailed hypothesis testing. Thus, we look at the rows in the "Two-tailed" section (see Figure 7). One could argue that the MRES value of 0.25 is the minimum meaningful improvement in education policy and practice. An MRES = 0.25 means that an intervention could bump up an average student's score from the 50th percentile to the 60th percentile. Thus, Table 2A in the Appendix is chosen. Bulus and Koyuncu (2021) reported that the explanatory power of the pretest for affective outcomes is .38 on average, a value between $R^2 = .35$ and $R^2 = .40$ (see Figure 7). It is common to deem a program effective if the p -value for the treatment effect is below .05. Thus, the row with $\alpha = .05$ is chosen (see Figure 7). Without any cost considerations, it is ideal to choose a balanced sample ($p = .50$).

For $R^2 = .35$ we need 328 subjects whereas for $R^2 = .40$ we need 303 subjects. A difference of .05 in R^2 corresponds to a difference of 25 subjects in MRSS. $R^2 = .38$ is .02 (2/5 of the difference) units away from the $R^2 = .40$, so approximately the sample size will be 2/5 of 25 (10 subjects) more. As a result $303 + 10 = 313$ subjects are needed in total. Note that this number is the same as the MRSS found in the software illustration section. An MRSS of 313 is the minimum required number. Surely more subjects can be recruited. Finally, one could randomly allocate 157 subjects into the treatment group and the remaining 157 subjects into the control group.

One can report the power analysis procedure in a paragraph as follows:

For this randomized pretest-posttest control-group design, we assume that the pretest explains 38% of the posttest variance (Bulus and Koyuncu, 2021). We further assume that the hypothesis test is two-tailed, the Type I error rate is 5%, and the power rate is 80%. Under these conditions, based on Table 2A in Bulus (2021), we decided on a sample of 314 subjects equally allocated to treatment and control groups to detect an effect size as small as 0.25.

Sample Size Determination in Weak Experiments

Table Illustration

There is no known software to determine MRSS for a non-equivalent pretest-posttest control-group design ($R^2 > 0$) and non-equivalent posttest-only control-group designs ($R^2 = 0$) yet. Researchers can use Tables S1–S28 in the Supplement for this purpose. Using the same specifications in Figure 7, except that now treatment and control groups are not equivalent on the pretest score, we can find the MRSS for a non-equivalent pretest-posttest control-group design. Assume that the point-biserial correlation between the pretest and treatment indicator is 0.243, translating into a standardized pretest difference of 0.50 between treatment and control groups. From the INDEX worksheet in Figure 8, one can choose Table S8 for this purpose.

Appendix - Randomized Pretest-posttest Control-group Design (True Experiment)

R	O ₁	X	O ₂
R	O ₃		O ₄

R: Random assignment. O: Observed measurement. X: Exposure to treatment.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design (True Experiment)

Reference Table	Minimum Relevant Effect Size (MRES)	Pretest Difference (PREDIFF) (as $n \rightarrow \infty$)	Point-biserial Correlation (r_{TX}) (as $n \rightarrow \infty$)
Table A1	0.20	0.00	0.00
Table A2	0.25	0.00	0.00
Table A3	0.30	0.00	0.00
Table A4	0.35	0.00	0.00
Table A5	0.40	0.00	0.00
Table A6	0.45	0.00	0.00
Table A7	0.50	0.00	0.00

Note. PREDIFF: Standardized pretest difference between treatment and control groups. r_{TX} : Point-biserial correlation between pretest and treatment indicator.

Supplement - Non-equivalent Pretest-posttest Control-group Design (Weak-experiment)

	O ₁	X	O ₂
	O ₃		O ₄

O: Observed measurement. X: Exposure to treatment.

Minimum Required Sample Size for Non-equivalent Pretest-posttest Control-group Design (Weak experiment)

Reference Table	Minimum Relevant Effect Size (MRES)	Pretest Difference (PREDIFF)	Point-biserial Correlation (r_{TX})
Table S1	0.20	0.20	0.100
Table S2	0.20	0.30	0.148
Table S3	0.20	0.40	0.195
Table S4	0.20	0.50	0.243
Table S5	0.25	0.20	0.100
Table S6	0.25	0.30	0.148
Table S7	0.25	0.40	0.195
Table S8	0.25	0.50	0.243
Table S9	0.30	0.20	0.100
Table S10	0.30	0.30	0.148

Figure 8. Finding the relevant table from the Supplemental Excel workbook based on MRES and pretest difference specifications.

For $R^2 = .35$ we need 349 subjects whereas for $R^2 = .40$ we need 322 subjects (see Figure 9). A difference of .05 in R^2 corresponds to a difference of 27 subjects in MRSS. $R^2 = .38$ is .02 (2/5 of the difference) units away from the $R^2 = .40$, so approximately the sample size will be 2/5 of 27 (~11 subjects) more. As a result, $322 + 11 = 333$ subjects are needed in total. Twenty more subjects are needed compared to the earlier example with randomized pretest-posttest control-group design due to the pretest differences between treatment and control groups.

Table S8. Minimum Required Sample Size for Non-equivalent Pretest-posttest Control-group Experimental Design MRES = 0.25 &

PREDIFF = .50

	Alpha	p	R2=0	R2=.30	R2=.35	R2=.40	R2=.45	R2=.50	R2=.55	R2=.60	R2=.65	R2=.70
One-Tailed	0.001	0.35	1160	813	755	698	640	582	524	467	409	351
	0.001	0.40	1100	771	716	662	607	552	497	442	388	333
	0.001	0.45	1066	748	695	642	589	535	482	429	376	323
	0.001	0.50	1056	740	688	635	583	530	478	425	372	320
	0.01	0.35	753	528	490	453	415	378	340	303	265	228
	0.01	0.40	714	500	465	429	394	358	323	287	251	216
	0.01	0.45	692	485	451	416	382	347	313	278	244	209
	0.01	0.50	685	480	446	412	378	344	310	276	241	207
	0.05	0.35	464	325	302	279	256	233	209	186	163	140
	0.05	0.40	440	308	286	264	242	221	199	177	155	133
	0.05	0.45	426	299	278	256	235	214	193	171	150	129
	0.05	0.50	422	296	275	254	233	212	191	170	149	128
Two-Tailed	0.001	0.35	1281	898	834	771	707	643	579	515	452	388
	0.001	0.40	1215	852	791	731	670	610	549	489	428	368
	0.001	0.45	1178	826	767	709	650	591	533	474	415	357
	0.001	0.50	1166	818	760	702	644	586	528	469	411	353
	0.01	0.35	876	614	570	527	483	440	396	352	309	265
	0.01	0.40	830	582	541	500	458	417	375	334	293	251
	0.01	0.45	805	565	525	484	444	404	364	324	284	244
	0.01	0.50	797	559	519	480	440	400	361	321	281	241
	0.05	0.35	589	413	383	354	325	295	266	237	207	178
	0.05	0.40	558	391	363	336	308	280	252	224	197	169
	0.05	0.45	541	379	352	325	298	272	245	218	191	164
	0.05	0.50	536	376	349	322	296	269	242	215	189	162

GO TO INDEX

Note. Statistical power is fixed at 80% for all designs. Alpha is the Type I error rate. p is the treatment group allocation rate (proportion of subjects in the treatment group). Values in table (n) refers to the total sample size. There will be pn subjects in the treatment and (1-p)n subjects in the control condition. MRES: Minimum relevant effect size. PREDIFF: Standardized pretest difference between treatment and control groups. R2: Proportion of variance in the posttest explained by the pretest variable. R2 = 0 applies to non-equivalent posttest only control-group experimental designs.

MRES = .25 & PREDIFF = .40 | MRES = .25 & PREDIFF = .50 | MRES = .30 & PREDIFF = .20 | MRES = .30 & PREDIFF = .30

Figure 9. Finding MRSS from the Supplemental Excel workbook based on MRES, R^2 , and pretest difference specifications.

One can report the power analysis procedure in a paragraph as follows:

This non-equivalent pretest-posttest control-group design assumes that the pretest explains 38% of the posttest variance (Bulus and Koyuncu, 2021). We further assume a point-biserial correlation of .243 between the pretest and treatment indicator, translating into a standardized pretest difference of 0.50 between treatment and control groups. We further assume that the hypothesis test is two-tailed, the Type I error rate is 5%, and the power rate is 80%. Under these conditions, based on Table 8S in Bulus (2021), we decided on a sample of 334 subjects (167 of them in the treatment and 167 of them in the control group) to detect an effect size as small as 0.25.

Discussion

Researchers can use G*Power for randomized posttest-only control-group designs. They can also use PowerUpR or PowerUp! via setting $R^2 = 0$ and $g = 0$ for this purpose. Collecting pretest information and other covariates means that $R^2 > 0$. This reduces the required sample size for an experiment. As for the randomized pretest-posttest control-group designs, researchers can use PowerUpR or PowerUp! via setting $R^2 > 0$ and $g > 0$ depending on the explanatory power of the pretest and covariates. G*Power and PowerUpR results are comparable when the explanatory power pretest or covariates is zero ($R^2 = 0$). PowerUpR allows $R^2 > 0$, whereas there is no convenient option in G*Power for pretest adjustment. Results differ by one or two units in some cases, possibly due to internal rounding differences used during intermediate computations. It is possible to convert G*Power results for $R^2 = 0$ to other scenarios with $R^2 > 0$. If one multiplies G*Power results for $R^2 = 0$ by the term $(1 - R^2)$, they will obtain sample sizes comparable to PowerUpR. For example, to detect MRES = 0.20 using a two-tailed test with $\alpha = .05$, $p = .50$, and $R^2 = .50$, PowerUpR produces an MRSS = 394 (see Table

1A in the Appendix). G*Power produces an MRSS = 788 with the same specifications. If we multiply the result from G*Power by $(1 - R^2)$, we get 394, which is the same as the result produced by PowerUpR.

Alternatively, one can use Tables 1A through 7A in the Appendix for randomized posttest-only control group design ($R^2 = 0$ & $g = 0$) and randomized pretest-posttest control-group designs ($R^2 > 0$ & $g > 0$). There are some evident trends in MRSS values reported in Tables 1A–7A in the Appendix. Two-tailed hypothesis tests require larger sample sizes compared to one-tailed hypothesis tests. The smaller the Type I error rate (α), the larger the sample size requirement. A balanced sample ($p = .50$) requires a smaller sample size than an unbalanced sample (though one may favor unbalanced samples under differential costs). The bigger the value of R^2 , the smaller the sample size requirement. Finally, to detect smaller MRES, larger sample sizes are required.

There is no known software to find MRSS for non-equivalent posttest-only control-group design ($R^2 = 0$) and non-equivalent pretest-posttest control group design ($R^2 > 0$). One can use Tables 1S through 28S in the Supplemental Excel workbook for this purpose. Trends observed in Tables 1A–7A for true experiments apply to Tables 1S–28S for weak experiments. For a small point-biserial correlation between pretest and treatment indicator ($r_{TX} \cong .10$), in other words, for a small standardized difference on the pretest between treatment and control groups, MRSS values hardly differ between tables in the Appendix and tables in the Supplement. For a moderate to large correlation ($r_{TX} \cong .30$ and above), in other words, a moderate standardized difference on the pretest between treatment and control groups, differences between Tables in the Appendix, and those in the Supplement become noticeable. Weak experiments typically require larger sample sizes.

Weak experiments could be manipulated before an intervention so that treatment and control groups are comparable on the pretest. One such procedure is known as matching. Subjects not only can be matched on the pretest but they can also be matched on other relevant covariates. These designs are referred to as quasi-experimental designs (Fraenkel et al., 2011). The corresponding quasi-experimental designs would be the matching-only pretest-posttest control-group and matching-only posttest-only control-group designs (Fraenkel et al., 2011). Reserving only matched pairs and discarding remaining subjects will reduce the sample size and result in a loss of power. Assuming that the pretest difference between treatment and control groups is negligible after matching, one can use Tables 1A–7A to determine MRSS values and plan their sample size accordingly. There are other methods to ensure that treatment and control groups are comparable; propensity score matching (Rosenbaum & Rubin, 1983), prognostic scores (Hansen, 2006, 2008; Wyss et al., 2015), prognostic propensity scores (Leacy & Stuart, 2013), coarsened exact matching (Iacus et al., 2012), inverse probability of treatment weighting (Huber, 2014). The description of these methods is beyond the scope of this study. Readers are referred to the references.

Formulas described in this study, software illustrations, and MRSS values in Tables 1A–7A and 1S–28S assume that observations are independent of each other. This assumption is often violated in practice because students are nested within classrooms (or teachers), and classrooms are nested within schools. Students in the same classroom or school tend to perform similarly. In other words, their scores are correlated due to contextual effects. Design and analysis experiments with nested structure require specialized statistical tools. An emerging bulk of studies consider this nested structure in the design of experiments (e.g., Bloom, 2006; Dong & Maynard, 2013; Hedges & Rhoads, 2010; Raudenbush & Liu, 2000; Konstantopoulos, 2008a; Konstantopoulos, 2008b; Schochet, 2008; Spybrook, 2007 and many others). To find MRSS for such complex experimental designs, researchers can use the PowerUpR or PowerUp!

Conclusion

This study elaborated on the nuts and bolts of sample size determination (or power analysis) in true experiments (randomized pretest-posttest control groups design and randomized posttest-only control-group design) and weak experiments (non-equivalent pretest-posttest control-group design and non-equivalent posttest-only control group design). In addition, illustrations provided step-by-step guidance on using G*Power, PowerUpR, and PowerUp! freeware to determine MRSS for true experiments. Furthermore, the optimal design of true experiments is illustrated using the companion Optimal Design Excel workbook. Finally, this study provided MRSS values for common scenarios in Tables 1A–7A for true experiments and Tables 1S–28S for weak experiments.

G*Power and PowerUpR produced the same results for randomized posttest-only control-group designs. G*Power results can be converted to PowerUpR via multiplying them by $(1 - R^2)$. PowerUpR and PowerUp! cover a broader range of experimental designs. Either of them can be used to design a randomized pretest-posttest control-group design. The software illustration section defined relevant design parameters and discussed reasonable values for them. One crucial design parameter is the minimum relevant effect size (MRES). Effects below the benchmark MRES would not be an interest to education policy and practice. When no data or literature is available for benchmark MRES value, 0.20 or 0.25 can be used. The second crucial

parameter is R^2 value defined as the proportion of variance in the posttest explained by the pretest. R^2 values should rely on earlier studies of a similar kind. When no information is available, researchers can use $R^2 = .22$ for cognitive outcomes and $R^2 = .38$ for affective outcomes. These values are based on 155 experimental studies reviewed in Bulus and Koyuncu (2021).

This study also provided optimal design formulas for randomized pretest-posttest control-group designs under differential cost assumption. When treatment units are more expensive than control units, and the total cost or budget is fixed, researchers can find optimal p and n . Optimal p depends on the cost ratio (cost per treatment unit/cost per control unit), and n depends on total cost or budget given p . Suppose the total cost or budget is flexible. In this case, the researcher can explore several options described in the illustration. They can then compare the total cost with $p = .50$ and decide whether it is worth pursuing an unbalanced design. Suppose the additional cost induced by the balanced design is not that much. In that case, it is probably better to use a balanced design. Optimal design formulas are implemented in the Optimal Design Excel workbook accompanying this article.

Finally, MRSS values in Tables 1A–7A allow researchers unfamiliar with R programming and Excel workbook to decide on an MRSS for randomized pretest-posttest control groups design and randomized posttest-only control-group design. There is no known software for finding MRSS in non-equivalent pretest-posttest control-group design and non-equivalent posttest-only control group design. Tables 1S–28S in the Supplement Excel workbook are helpful in this aspect.

References

- Bloom, H. S. (2006). *The core analytics of randomized experiments for social research* (MDRC Working Papers on Research Methodology). <http://www.mdrc.org/publications/437/full.pdf>
- Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-school variation in students' achievement, motivation, affect, and learning strategies: Results from 81 countries for planning group-randomized trials in education. *Journal of Research on Educational Effectiveness*, *11*(3), 452-478. <https://doi.org/10.1080/19345747.2017.1375584>
- Bulus, M., & Dong, N. (2021a). Bound constrained optimization of sample sizes subject to monetary restrictions in planning of multilevel randomized trials and regression discontinuity studies. *The Journal of Experimental Education*, *89*(2), 379–401. <https://doi.org/10.1080/00220973.2019.1636197>
- Bulus, M., & Dong, N. (2021b). cosa: Bound constrained optimal sample size allocation. R package version 2.1.0. <https://CRAN.R-project.org/package=cosa>
- Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2021). PowerUpR: Power analysis tools for multilevel randomized experiments. R package version 1.1.0. <https://CRAN.R-project.org/package=PowerUpR>
- Bulus, M., & Koyuncu, I. (2021). Statistical power and precision of experimental studies originated in the Republic of Turkey from 2010 to 2020: Current practices and some recommendations. *Journal of Participatory Education Research*, *8*(4), 24-43. <https://doi.org/10.17275/per.21.77.8.4>
- Bulus, M., & Sahin, S. G. (2019). Estimation and standardization of variance parameters for planning cluster-randomized trials: A short guide for researchers. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(2), 179-201. <https://doi.org/10.21031/epod.530642>
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67. <https://doi.org/10.1080/19345747.2012.673143>
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers*, *28*(1), 1-11. <https://doi.org/10.3758/BF03203630>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. <https://doi.org/10.3758/BRM.41.4.1149>

- Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2011). *How to design and evaluate research in education* (10th Ed.). McGraw-Hill.
- Hansen, B. B. (2006). *Bias reduction in observational studies via prognosis scores*. Technical report #441, University of Michigan Statistics Department. <http://dept.stat.lsa.umich.edu/~bbh/rspaper2006-06.pdf>
- Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2), 481-488. <https://doi.org/10.1093/biomet/asn004>
- Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2), 246-255. <https://www.jstor.org/stable/2246311>
- Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, 37(6), 445-489. <https://doi.org/10.1177/0193841X14529126>
- Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. <https://files.eric.ed.gov/fulltext/ED509387.pdf>
- Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6), 920-943. <https://doi.org/10.1002/jae.2341>
- Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1), 1-24. <https://www.jstor.org/stable/41403736>
- Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, 1, 265-288. <https://doi.org/10.1080/19345740802328216>
- Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level cluster-randomized designs. *Journal of Research on Educational Effectiveness*, 1, 66-88. <https://doi.org/10.1080/19345740701692522>
- Leacy, F. P., & Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine*, 33(20), 3488-3508. <https://doi.org/10.1002/sim.6030>
- Mosteller, F. F., & Boruch, R. F. (Eds.). (2004). *Evidence matters: Randomized trials in education research*. Brookings Institution Press.
- Oakes, J. M., & Feldman, H. A. (2001). Statistical power for non-equivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review*, 25(1), 3-28. <https://doi.org/10.1177%2F0193841X0102500101>
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5, 199-213. <https://doi.org/10.1037/1082-989X.5.2.199>
- Rickles, J., Zeiser, K., & West, B. (2018). Accounting for student attrition in power calculations: Benchmarks and guidance. *Journal of Research on Educational Effectiveness*, 11(4), 622-644. <https://doi.org/10.1080/19345747.2018.1502384>
- Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. <https://doi.org/10.1093/biomet/70.1.41>
- Slavin, R. E. (2008). Perspectives on evidence-based research in education: What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14. <https://doi.org/10.3102%2F0013189X08314117>
- Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53(11), 1119-1129. [https://doi.org/10.1016/s0895-4356\(00\)00242-0](https://doi.org/10.1016/s0895-4356(00)00242-0)
- Wyss, R., Ellis, A. R., Brookhart, M. A., Jonsson Funk, M., Girman, C. J., Simpson, R. J., & Stürmer, T. (2015). Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and Drug Safety*, 24(9), 951-961. <https://doi.org/10.1002/pds.3810>
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419-435. <https://doi.org/10.1007/BF02294384>

- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122. <https://doi.org/10.1177%2F1745691617693393>
- Yildirim, I., Cirak-Kurt, S., & Sen, S. (2019). The effect of teaching "Learning strategies" on academic achievement: A meta-analysis study. *Eurasian Journal of Educational Research*, 79, 87-114. <https://doi.org/10.14689/ejer.2019.79.5>

Appendix

Table 1A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when $MRES = 0.20$

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power				PowerUpR						
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	1704	1703	1194	1109	1024	939	854	769	684	599	514
	0.001	1.50	0.40	1616	1615	1132	1051	971	890	810	729	649	568	487
	0.001	1.22	0.45	1568	1566	1097	1019	941	863	785	707	629	551	473
	0.001	1.00	0.50	1552	1550	1087	1009	932	855	777	700	623	545	468
	0.01	1.86	0.35	1106	1105	775	719	664	609	554	499	444	389	333
	0.01	1.50	0.40	1050	1048	734	682	630	578	525	473	421	368	316
	0.01	1.22	0.45	1018	1016	712	661	611	560	509	459	408	357	307
	0.01	1.00	0.50	1008	1006	705	655	605	555	504	454	404	354	304
	0.05	1.86	0.35	682	681	477	443	409	375	341	307	273	239	205
	0.05	1.50	0.40	646	646	452	420	388	356	324	291	259	227	195
Two-tailed	0.001	1.86	0.35	1882	1881	1318	1225	1131	1037	943	849	755	662	568
	0.001	1.50	0.40	1786	1784	1250	1161	1072	983	894	805	716	627	539
	0.001	1.22	0.45	1732	1730	1212	1126	1040	954	867	781	695	609	522
	0.001	1.00	0.50	1714	1712	1200	1115	1029	944	859	773	688	603	517
	0.01	1.86	0.35	1288	1286	901	837	773	709	645	581	516	452	388
	0.01	1.50	0.40	1220	1220	855	794	733	672	611	551	490	429	368
	0.01	1.22	0.45	1184	1183	829	770	711	652	593	534	475	416	357
	0.01	1.00	0.50	1172	1171	821	762	704	645	587	529	470	412	353
	0.05	1.86	0.35	866	864	606	563	519	476	433	390	347	304	261
	0.05	1.50	0.40	820	820	574	533	492	452	411	370	329	288	247
0.05	1.22	0.45	796	795	557	517	478	438	398	359	319	279	240	
0.05	1.00	0.50	788	787	551	512	473	434	394	355	316	277	237	

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only the pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 2A.

Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.25

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power				PowerUpR						
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	1094	1092	765	711	657	602	548	494	439	385	331
	0.001	1.50	0.40	1036	1035	726	674	623	571	520	468	417	365	314
	0.001	1.22	0.45	1006	1004	704	654	604	554	504	454	404	354	304
	0.001	1.00	0.50	996	994	697	647	598	549	499	450	400	351	301
	0.01	1.86	0.35	710	708	497	461	426	391	355	320	285	250	214
	0.01	1.50	0.40	672	672	471	437	404	371	337	304	270	237	203
	0.01	1.22	0.45	652	651	457	424	392	359	327	295	262	230	197
	0.01	1.00	0.50	646	645	452	420	388	356	324	292	260	227	195
	0.05	1.86	0.35	438	436	306	284	262	241	219	197	175	154	132
	0.05	1.50	0.40	414	414	290	269	249	228	208	187	166	146	125
	0.05	1.22	0.45	402	401	281	261	241	221	201	181	161	141	121
	0.05	1.00	0.50	398	397	278	259	239	219	199	180	160	140	120
Two-tailed	0.001	1.86	0.35	1208	1206	846	785	725	665	605	545	485	425	365
	0.001	1.50	0.40	1144	1143	802	745	688	631	574	517	460	403	346
	0.001	1.22	0.45	1110	1109	778	722	667	612	557	502	446	391	336
	0.001	1.00	0.50	1100	1098	770	715	661	606	551	497	442	387	333
	0.01	1.86	0.35	826	824	578	537	496	455	414	373	332	291	249
	0.01	1.50	0.40	782	782	548	509	470	431	392	353	315	276	237
	0.01	1.22	0.45	760	758	532	494	456	418	381	343	305	267	230
	0.01	1.00	0.50	752	751	526	489	452	414	377	339	302	265	227
	0.05	1.86	0.35	554	554	388	361	333	306	278	250	223	195	168
	0.05	1.50	0.40	526	525	368	342	316	290	264	237	211	185	159
	0.05	1.22	0.45	510	509	357	332	306	281	256	230	205	180	154
	0.05	1.00	0.50	506	504	354	328	303	278	253	228	203	178	153

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 3A.
 Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.30

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power				PowerUpR						
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	760	759	533	495	457	420	382	344	306	269	231
	0.001	1.50	0.40	722	720	505	470	434	398	362	326	291	255	219
	0.001	1.22	0.45	700	698	490	456	421	386	351	317	282	247	213
	0.001	1.00	0.50	692	691	485	451	417	382	348	314	279	245	211
	0.01	1.86	0.35	494	493	346	321	297	272	248	223	199	174	150
	0.01	1.50	0.40	468	467	328	305	281	258	235	212	188	165	142
	0.01	1.22	0.45	454	453	318	295	273	250	228	205	183	160	138
	0.01	1.00	0.50	450	449	315	293	270	248	226	203	181	159	136
	0.05	1.86	0.35	304	303	213	198	183	168	152	137	122	107	92
	0.05	1.50	0.40	288	288	202	188	173	159	145	130	116	102	87
	0.05	1.22	0.45	280	279	196	182	168	154	140	126	113	99	85
	0.05	1.00	0.50	278	276	194	180	166	153	139	125	111	98	84
Two-tailed	0.001	1.86	0.35	840	839	589	547	505	464	422	380	339	297	255
	0.001	1.50	0.40	796	795	558	519	479	440	400	361	321	282	242
	0.001	1.22	0.45	772	771	542	503	465	427	388	350	312	273	235
	0.001	1.00	0.50	766	764	536	498	460	422	384	346	309	271	233
	0.01	1.86	0.35	574	573	402	374	345	317	288	260	231	203	174
	0.01	1.50	0.40	546	544	382	355	327	300	273	246	219	192	165
	0.01	1.22	0.45	528	527	370	344	318	291	265	239	213	187	160
	0.01	1.00	0.50	524	522	366	340	315	289	263	237	211	185	159
	0.05	1.86	0.35	386	385	270	251	232	213	194	174	155	136	117
	0.05	1.50	0.40	366	365	256	238	220	202	184	165	147	129	111
	0.05	1.22	0.45	356	354	249	231	213	196	178	161	143	125	108
	0.05	1.00	0.50	352	351	246	229	211	194	176	159	141	124	107

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 4A.
 Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.35

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power		PowerUpR								
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	560	559	393	365	337	309	282	254	226	199	171
	0.001	1.50	0.40	532	530	372	346	320	294	267	241	215	188	162
	0.001	1.22	0.45	516	514	361	336	310	285	259	234	208	183	157
	0.001	1.00	0.50	510	509	358	333	307	282	257	232	206	181	156
	0.01	1.86	0.35	364	363	255	237	219	201	183	165	147	129	111
	0.01	1.50	0.40	346	344	242	224	207	190	173	156	139	122	105
	0.01	1.22	0.45	334	334	234	218	201	185	168	152	135	118	102
	0.01	1.00	0.50	332	330	232	216	199	183	166	150	134	117	101
	0.05	1.86	0.35	224	223	157	146	135	124	112	101	90	79	68
	0.05	1.50	0.40	212	212	149	138	128	117	107	96	86	75	65
	0.05	1.22	0.45	206	205	144	134	124	114	103	93	83	73	63
0.05	1.00	0.50	204	203	143	133	123	113	102	92	82	72	62	
Two-tailed	0.001	1.86	0.35	620	618	434	403	373	342	311	281	250	219	189
	0.001	1.50	0.40	588	586	411	382	353	324	295	266	237	208	179
	0.001	1.22	0.45	570	568	399	371	343	315	287	258	230	202	174
	0.001	1.00	0.50	564	562	395	367	339	312	284	256	228	200	172
	0.01	1.86	0.35	424	422	296	275	255	234	213	192	171	150	129
	0.01	1.50	0.40	402	400	281	261	241	222	202	182	162	142	122
	0.01	1.22	0.45	390	388	273	253	234	215	196	176	157	138	119
	0.01	1.00	0.50	386	384	270	251	232	213	194	175	156	137	118
	0.05	1.86	0.35	284	284	199	185	171	157	143	129	115	101	86
	0.05	1.50	0.40	270	269	189	175	162	149	135	122	109	95	82
	0.05	1.22	0.45	262	261	183	170	157	144	131	118	106	93	80
0.05	1.00	0.50	260	258	181	169	156	143	130	117	104	92	79	

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 5A.
 Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.40

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power				PowerUpR						
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	430	429	302	280	259	238	217	196	174	153	132
	0.001	1.50	0.40	408	407	286	266	246	226	206	186	165	145	125
	0.001	1.22	0.45	396	395	278	258	239	219	200	180	161	141	122
	0.001	1.00	0.50	392	391	275	256	236	217	198	178	159	140	120
	0.01	1.86	0.35	280	278	196	182	168	154	140	127	113	99	85
	0.01	1.50	0.40	266	264	186	172	159	146	133	120	107	94	81
	0.01	1.22	0.45	258	256	180	167	155	142	129	117	104	91	79
	0.01	1.00	0.50	254	253	178	166	153	141	128	116	103	90	78
	0.05	1.86	0.35	172	171	120	112	103	95	86	78	69	61	53
	0.05	1.50	0.40	164	163	114	106	98	90	82	74	66	58	50
Two-tailed	0.001	1.86	0.35	476	474	333	310	286	263	239	216	193	169	146
	0.001	1.50	0.40	452	450	316	294	272	250	227	205	183	161	138
	0.001	1.22	0.45	438	436	307	285	264	242	221	199	177	156	134
	0.001	1.00	0.50	434	432	304	282	261	240	218	197	176	154	133
	0.01	1.86	0.35	326	324	228	212	196	180	164	147	131	115	99
	0.01	1.50	0.40	308	307	216	201	186	170	155	140	125	110	94
	0.01	1.22	0.45	300	298	210	195	180	165	151	136	121	106	92
	0.01	1.00	0.50	296	295	207	193	178	164	149	134	120	105	91
	0.05	1.86	0.35	218	218	153	142	131	121	110	99	88	77	67
	0.05	1.50	0.40	208	206	145	135	125	114	104	94	84	74	63
	0.05	1.22	0.45	202	200	141	131	121	111	101	91	81	71	61
	0.05	1.00	0.50	200	198	139	130	120	110	100	90	80	71	61

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 6A.
 Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.45

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power		PowerUpR								
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	342	340	239	223	206	189	172	155	139	122	105
	0.001	1.50	0.40	324	322	227	211	195	179	163	148	132	116	100
	0.001	1.22	0.45	314	313	220	205	189	174	159	143	128	112	97
	0.001	1.00	0.50	312	310	218	203	188	172	157	142	127	111	96
	0.01	1.86	0.35	222	220	155	144	133	122	112	101	90	79	68
	0.01	1.50	0.40	210	209	147	137	126	116	106	96	85	75	65
	0.01	1.22	0.45	204	203	143	133	123	113	103	93	83	73	63
	0.01	1.00	0.50	202	201	141	131	122	112	102	92	82	72	62
	0.05	1.86	0.35	136	136	95	89	82	75	69	62	55	49	42
	0.05	1.50	0.40	130	129	91	84	78	72	65	59	52	46	40
0.05	1.22	0.45	126	125	88	82	76	69	63	57	51	45	39	
0.05	1.00	0.50	124	124	87	81	75	69	63	57	50	44	38	
Two-tailed	0.001	1.86	0.35	376	376	264	246	227	209	190	172	153	135	116
	0.001	1.50	0.40	358	356	251	233	216	198	181	163	146	128	110
	0.001	1.22	0.45	348	346	243	226	209	192	175	158	141	124	107
	0.001	1.00	0.50	344	342	241	224	207	190	174	157	140	123	106
	0.01	1.86	0.35	258	257	181	168	155	143	130	117	105	92	79
	0.01	1.50	0.40	244	243	171	159	147	135	123	111	99	87	75
	0.01	1.22	0.45	238	236	166	155	143	131	120	108	96	85	73
	0.01	1.00	0.50	236	234	165	153	142	130	118	107	95	84	72
	0.05	1.86	0.35	174	172	121	113	104	96	87	79	70	62	53
	0.05	1.50	0.40	164	163	115	107	99	91	83	75	67	59	50
	0.05	1.22	0.45	160	159	112	104	96	88	80	72	65	57	49
	0.05	1.00	0.50	158	157	110	103	95	87	80	72	64	56	49

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."

Table 7A.
 Minimum Required Sample Size for Randomized Pretest-posttest Control-group Design when MRES = 0.50

Hypothesis Test	α	Allocation Ratio	p	Minimum Required Sample Size (n)										
				G*Power		PowerUpR								
				$R^2=0$	$R^2=0$	$R^2=.30$	$R^2=.35$	$R^2=.40$	$R^2=.45$	$R^2=.50$	$R^2=.55$	$R^2=.60$	$R^2=.65$	$R^2=.70$
One-tailed	0.001	1.86	0.35	278	276	195	181	167	154	140	127	113	100	86
	0.001	1.50	0.40	264	262	185	172	159	146	133	120	108	95	82
	0.001	1.22	0.45	256	254	179	167	154	142	129	117	104	92	79
	0.001	1.00	0.50	254	252	178	165	153	140	128	116	103	91	79
	0.01	1.86	0.35	180	179	126	117	108	100	91	82	73	64	56
	0.01	1.50	0.40	170	170	120	111	103	95	86	78	70	61	53
	0.01	1.22	0.45	166	165	116	108	100	92	84	76	68	59	51
	0.01	1.00	0.50	164	163	115	107	99	91	83	75	67	59	51
	0.05	1.86	0.35	112	110	78	72	67	61	56	50	45	40	34
	0.05	1.50	0.40	106	105	74	69	63	58	53	48	43	38	33
	0.05	1.22	0.45	102	101	71	66	62	57	52	47	42	37	32
0.05	1.00	0.50	102	100	71	66	61	56	51	46	41	36	31	
Two-tailed	0.001	1.86	0.35	306	305	215	200	185	170	155	140	125	110	95
	0.001	1.50	0.40	292	290	204	190	176	161	147	133	119	105	90
	0.001	1.22	0.45	282	281	198	184	171	157	143	129	115	102	88
	0.001	1.00	0.50	280	278	196	183	169	155	142	128	114	101	87
	0.01	1.86	0.35	210	208	147	137	126	116	106	96	85	75	65
	0.01	1.50	0.40	198	198	139	130	120	110	100	91	81	71	62
	0.01	1.22	0.45	194	192	135	126	116	107	97	88	79	69	60
	0.01	1.00	0.50	192	190	134	125	115	106	97	87	78	69	59
	0.05	1.86	0.35	140	140	99	92	85	78	71	64	57	50	43
	0.05	1.50	0.40	134	133	94	87	80	74	67	61	54	48	41
	0.05	1.22	0.45	130	129	91	84	78	72	65	59	53	46	40
0.05	1.00	0.50	128	128	90	84	77	71	65	59	52	46	40	

Note. MRES: Minimum relevant effect size. Statistical power is fixed at 80% for all designs. α is the Type I error rate. The allocation ratio is $(1-p) / p$ and is the required input for G*Power. n refers to the total sample size. R^2 is the proportion of variance in the posttest explained by the pretest variable (and other covariates, if available). If only pretest is included in the model, R^2 can be interpreted as the squared correlation between the pretest and posttest. There will be $p \times n$ subjects in the treatment group and $(1-p) \times n$ subjects in the control group. G*Power specifications: "Test family: t-tests" and "Statistical test: Means: Difference between two independent means (two groups)."