e-ISSN: 2148-7456

a peer-reviewed
online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 10 Issue: 2 June 2023

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

IJATE

e-ISSN: 2148-7456

Volume 10 Issue 2 2023

Editor	Dr. Ömer KUTLU
Address	Ankara University, Faculty of Educational Sciences, Cebeci Yerleşkesi, Çankaya, Ankara, Türkiye
E-mail	ijate.editor@gmail.com omerkutlu@ankara.edu.tr
Publisher	Dr. İzzet KARA
Address	Pamukkale University, Education Faculty, Kinikli Campus, 20070 Denizli, Türkiye
Phone	+90 258 296 1036
Fax	+90 258 296 1200
E-mail	ikara@pau.edu.tr ijate.editor@gmail.com
Journal Contact	Dr. Eren Can AYBEK
Address	Department of Educational Sciences, Pamukkale University, Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye
Phone	+90 258 296 31050
Fax	+90 258 296 1200
E-mail	erencanaybek@gmail.com
Frequency	4 issues per year (March, June, September, December)
Online ISSN	2148-7456
Website	https://dergipark.org.tr/en/pub/ijate https://ijate.net/index.php/ijate
Cover Design	Merve SENTURK

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed and academic online journal. The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).

International Journal of Assessment Tools in Education

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

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

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

Indexes and Platforms:

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

Editor

Section Editors

Editorial Board

English Language Editors

Editorial Assistants

Technical Assistant

CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 175–196

https://doi.org/10.21449/ijate.1160120

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

Validation of cognitive models for subtraction of time involving years and centuries

Huan Chin[1], Cheng Meng Chew[2,*]

[1]School of Educational Studies, Universiti Sains Malaysia, Penang, Malaysia
[2]School of Education, Humanities and Social Sciences, Wawasan Open University, Penang, Malaysia

ARTICLE HISTORY

Received: Aug. 10, 2022
Revised: May 01, 2023
Accepted: May 10, 2023

Keywords:
Cognitive models,
Cognitive diagnostic assessment,
Validation,
Attribute hierarchy method,
Learning of time.

Abstract: Years and Centuries are the measurement units used to quantify a longer time duration, while subtraction is the operation required to determine the duration based on two given time points. However, subtraction of time is a difficult skill to be mastered by many elementary students. To identify the root cause of the student's failure in performing subtraction involving the unit of time, we developed and validated the three cognitive models related to this skill by conducting a descriptive study which involved 119 Grade Five students from three Malaysian elementary schools. The cognitive diagnostic assessment developed based on the three cognitive models was used to elicit the participants' responses. Then, Attribute Hierarchy Method and Classical Test Theory were employed to analyse the data. The findings indicated that the hierarchical structures of all cognitive models are supported by the student's responses. The three student-based cognitive models were also highly consistent with the corresponding expert-based cognitive models. The cognitive models developed could guide diagnostic assessment development and diagnostic inference making.

1. INTRODUCTION

Time is one of the key concepts included in the domain of measurement in elementary school mathematics (National Council of Teachers of Mathematics [NCTM], 2000; Van de Walle et al., 2018). Time telling and determining the duration are the two key skills covered under the concept of time in elementary mathematics (Harris, 2008). Students learn about time telling in Grade One and Grade Two, followed by determining the duration of time in Grade Three onwards. Besides the commonly used time unit such as second, minute, hour, day, week, month, and year, students are also introduced to the year-based time unit such as year, decade, and century for quantifying a longer time duration. This numeracy capability will support the students in interpreting the historical timeline (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2017).

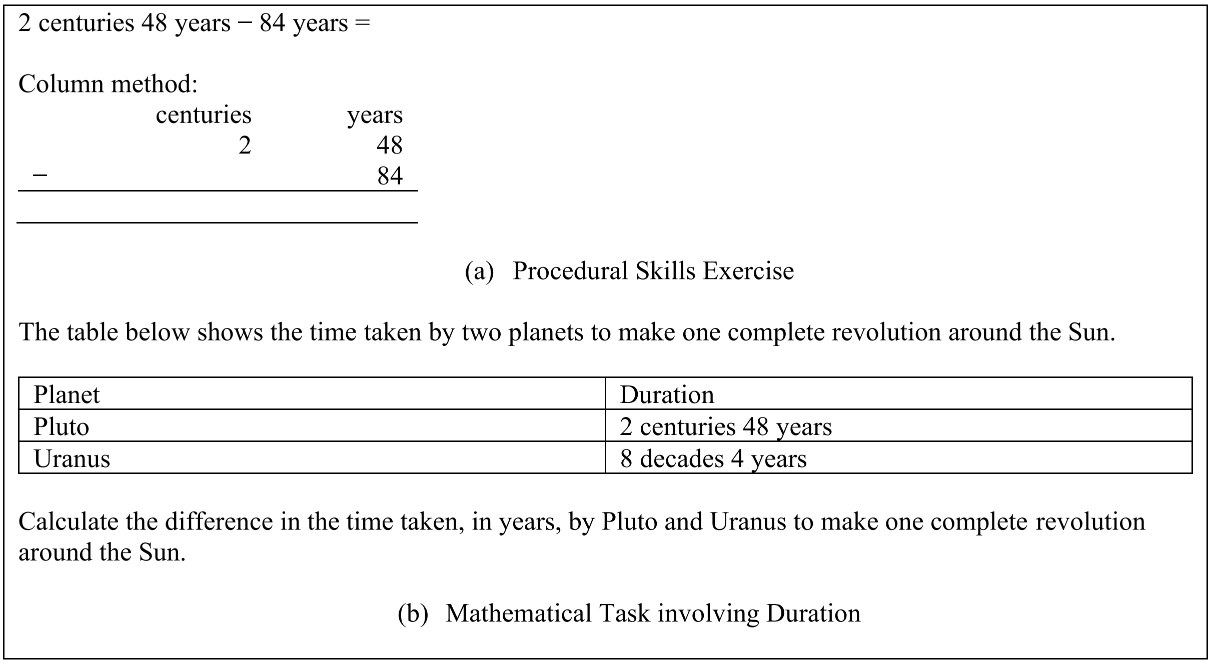
In fact, subtraction of time is the mathematical operation which is needed to determine the duration between two given time points (Sia et al., 2019). Regardless of the time unit, the

*CONTACT: Cheng Meng Chew ✉ cmchew@usm.my ; cmchew@wou.edu.my School of Educational Studies, Universiti Sains Malaysia, Penang, Malaysia & School of Education, Humanities and Social Sciences, Wawasan Open University, Penang, Malaysia

e-ISSN: 2148-7456 /© **IJATE** 2023

students would have to subtract the began time (subtrahend) from the ended time (minuend) to find the duration. Thus, the students were exposed to the procedural skills exercise related to subtraction of time as shown in Figure 1(a) prior proceed with the mathematical task involving duration as shown in Figure 1(b) (Chan et al., 2017). Specifically, the students would perform subtraction involving time by using the column method.

Figure 1. *Mathematical task related to 'Subtraction of Time' which involves Centuries and Years. Adopted from Chan (2017).*



2 centuries 48 years − 84 years =

Column method:

```
            centuries        years
                    2           48
 −                              84
__________________________________

__________________________________
```

(a) Procedural Skills Exercise

The table below shows the time taken by two planets to make one complete revolution around the Sun.

Planet	Duration
Pluto	2 centuries 48 years
Uranus	8 decades 4 years

Calculate the difference in the time taken, in years, by Pluto and Uranus to make one complete revolution around the Sun.

(b) Mathematical Task involving Duration

Even though procedural fluency has been emphasized in the mathematics classroom, students' failure in subtraction of time is frequently highlighted by researchers (i.e., Earnest, 2015; Kamii & Russell, 2012; Sia et al., 2019). Despite subtraction and conversion of time have been introduced to the students in the early grade, the students seem to fail to integrate their knowledge about subtraction and the relationship between the time units for performing subtraction involving the unit time. In this regard, cognitive modelling could be an appealing approach to deepen the educators' understanding of the student's failure in performing subtraction involving the unit of time from the perspective of their cognition (Leighton & Gierl, 2007).

The use of cognitive models in supporting the process of making diagnostic inferences has been illustrated in the studies conducted by Leighton and Gierl (2007), and Sia et al. (2019). While cognitive models were used to guide the test items' construction, students' performance would be tapped into mastery of the subskill which was arranged hierarchically as illustrated in the cognitive models. Thus, the use of cognitive models in test development and result interpretation would eventually highlight the students' cognitive weaknesses which hinder their mastery of the concept and hence explain their performance (Kane & Bejar, 2014). With the rich diagnostic information, the teachers could differentiate the teaching and plan for remedial instruction to support students in learning the subtraction of 'Time'. Thus, this study sought to develop and validate the cognitive models for 'Subtraction of Time' involving *years* and *centuries*.

1.1. Literature Review

1.1.1. *Cognitive models*

In the field of cognitive psychology, the cognitive model refers to the theoretical foundation of the procedures which are assumed to be carried out during complex cognitive activities such as problem-solving and decision-making (Keehner et al., 2017). The integration of cognitive psychology and educational measurement brings about the emergence of two types of cognitive models, namely *cognitive models of learning* and *cognitive models of task*. According to Keehner et al. (2017), the *cognitive models of learning* such as learning progressions describe the stages of knowledge and skill acquisition as well as competence development. Meanwhile, the *cognitive models of task* refer to the descriptions of the attribute used by the students in solving some tasks, for supporting the inferences made on students' performance (Leighton & Gierl, 2007).

Assessment practice can be implemented with the integration of both *cognitive models of task* as well as *cognitive models of learning*. By applying the specific cognitive-psychometric model, students' responses elicited using the test items guided by the *cognitive models of learning* will be mapped onto the corresponding developmental stage as stated in the *cognitive models of learning*. Meanwhile, fitting students' responses prompted using test items guided by *cognitive models of task* onto the cognitive-psychometric model would generate students' attribute mastery. In order to ensure the diagnostic information generated based on the cognitive models is highly precise and specific, the cognitive model of task must possess four significant properties: (i) each fine-grained attribute must be specified in a detailed manner consistently; (ii) each attribute must be able to measure using cognitive task; (iii) each attribute specified must be aligned with the curriculum; and (iv) each attribute must be structured hierarchically in the cognitive model (Gierl et al., 2009b).

The previous relevant studies mainly focused on the development and validation of *cognitive models of learning* for various topics of science, such as ecosystem (Jin et al., 2019), matter (Hadenfeldt et al., 2016), and phase transformation (Schultz et al., 2017). For the subject of mathematics, the cognitive models of learning have been developed to describe the developmental stages of a concept such as subtraction of fractions (Akbay et al., 2018) as well as number sense (Chen et al., 2017). Rather than focusing on the validation of *cognitive models of learning,* Sia et al. (2019) evaluated the consistency of the *cognitive model of task* for the '*duration'* concept which was hypothesized by the expert, and the *cognitive model of task* exhibited on students' actual cognitive processes when solving the tasks.

In this study, we focused on the development and validation of the cognitive models of task which can be used to support the diagnostic inference made regarding students' cognitive strengths and weaknesses which contribute to the mastery or non-mastery of the skills in 'Subtraction of Time'. We validated the *cognitive models of task* by using various sources of evidence to validate the data, rather than just evaluating the consistency as demonstrated by Sia et al. (2019). The term '*cognitive model*' will be used to indicate the '*cognitive model of task*' in this study.

1.1.2. *Development of cognitive models*

Ideally, the cognitive models should be constructed based on a substantive theory of cognition and learning (Nichols, 1994). Since a suitable substantive theory can rarely be found in the literature, Gierl et al. (2009a) introduced two approaches for developing the cognitive model, namely the top-down approach and the bottom-up approach. The top-down approach involves conducting a task analysis within the domain of interest for developing the cognitive models, whereas the bottom-up approach involves analysing the protocol data collected using the think-aloud method for developing the cognitive models.

The use of the top-down approach has been demonstrated in the study conducted by Gierl et al. (2008), as well as Sia and Lim (2018). They began with specifying the attributes which include the mathematical concepts, skills, and processes used to solve each task, followed by arranging those attributes hierarchically based on their complexity to form the cognitive models. Instead of conducting the task analysis, Chen et al. (2017) identified the attributes by reviewing the skills included in the textbooks.

Meanwhile, the use of the bottom-up approach has been demonstrated in the study conducted by Gierl et al. (2009a). They started with transcribing the recordings, coding the attribute, and presenting the process to solve the task by using the flowchart which represents the cognitive models from the student's perspective. Notably, the top-down approach has been used in majority of the previous studies due to the advantages of this approach in ensuring the cognitive models' credibility. With sufficient teaching experience, the experts would have a strong understanding of students' thinking, learning and instruction. This relevant expertise would eventually support them in identifying and arranging the instructional relevant attributes in hierarchical order to form the cognitive models (Gierl et al., 2009b).

1.1.3. *Validity of cognitive models*

As asserted by Leighton and Gierl (2007), the substantive theory of learning and cognition integrated into the assessment is in high demand. Thus, the cognitive models constructed are regarded as a new theory of learning and cognition (Nichols et al., 2017) which has never been demonstrated to be valid (Nichols, 1994). In other words, the attributes associated with students' cognitive processes in solving the related tasks are only hypothesized by the experts (Graf et al., 2019). Thus, empirical evidence needs to be accumulated to verify the validity of the cognitive models constructed (Graf et al., 2019).

The empirical evidence can be collected from various sources based on the nature of constructs and tasks. Most of the studies (i.e., Akbay et al., 2018; Chen et al., 2017; Graf et al., 2019; Langenfeld et al., 2020; Sia et al., 2019) employed cognitive-psychometric modelling to validate the cognitive models because the fit of cognitive models to the data can be computed by using the mathematics formula which take into consideration of the constraints imposed in the cognitive models (Keehner et al., 2017). In order to triangulate the data, Graf et al. (2019), Langenfeld et al. (2020) and Sia et al. (2019) further analysed the protocols collected by using the think-aloud method and conducting cognitive labs, respectively.

Unlike the study conducted by Langenfeld et al. (2019) and Sia et al. (2019), we began the empirical evaluation of cognitive models by conducting Item Difficulty Modelling (IDM) as suggested by Keehner et al. (2017). Following this, we could explore the cognitive models' hierarchical structure by taking into account the item difficulty (Gorin, 2006) and attribute complexity (Keehner et al., 2017). After conducting IDM, we evaluated the consistency between the student-based cognitive models (S-CM) and expert-based cognitive models (E-CM) by applying the cognitive-psychometric modelling.

1.1.4. *The learning of subtraction concept*

The concept of subtraction is commonly introduced to children informally (Clement et al., 2020) using the 'taking away', 'part-part-whole' and 'comparing two sets of items' problem situations (Carpenter et al., 1999). After conceptualising subtraction operation, the subtraction learning will begin with single-digit numbers with a unitary conceptual structure (Fuson, 1990). The students will engage in solving simple subtraction problems involving single-digit numbers (Clement et al., 2020) using Count-All-and Taking-Away (Murata & Kattubadi, 2012). After that, the students will be guided to make use of the numerical information at the subtrahend to find the difference using the Counting-Up or Counting Down strategy (Murata & Kattubadi, 2012). Once the students understand the relationship between minuend and subtrahend, they will be introduced to find the difference using the subtraction algorithm (Murata & Kattubadi, 2012).

The learning of subtraction is then extended to multi-digit numbers. The learning of multi-digit subtraction is more complex because the multi-digit number is conceptualized as 'multiunit quantities associated with multiunit names and position (Fusion, 1990, p. 350)'. For example, the two-digit number '23' is conceptualized as a combination of two bundles of 10 sticks and three single sticks. Thus, the understanding of the base-ten place value system preceded multi-digit subtraction (Fuson, 1990; Nuerk et al., 2015). Following this, Nuerk et al. (2015) suggested that multi-digit subtraction involves three processes: (i) place identification, (ii) place-value activation, and (iii) place-value computation. It begins with assigning each digit to the correct base-10 place value stack position (*place identification*), followed by conceptualising the digit located at the respective base-10 place value stack position as corresponding numerical magnitude (*place-value activation*), and regrouping the numbers and performing the subtraction across place-value stacks (*place-value computation*) (Nuerk et al., 2015).

The measurement of time involving a composite unit can be expressed as a multi-unit conceptual structure because it involves the pairing of a number associated with a higher measurement unit and a number associated with a lower measurement unit (Fusion, 1990b). Even though the higher measurement unit quantifies a collection of the lower measurement unit, the value of the number associated with the higher measurement units might not always equal 10 times the number associated with the lower measurement units. Thus, subtraction involving measurement might be slightly different from multi-digit subtraction which follows the base-10 place value system.

In fact, the learning of subtraction involving the measurement of time can be extended from multi-digit subtraction using different number bases considering the relationship between the units of time. In this study, learning of subtraction involving century and year could be extended from multi-digit subtraction with a number base of 100 because a century is quantified as 100 years.

1.1.5. *Past related studies on time concept*

While time telling and duration of time are two fundamental concepts in the topic of 'Time', several studies have been conducted on assessing students' performance on time telling (Brace et al., 2019; Lambert et al., 2020), as well as identifying students' common errors (Tan et al., 2019) and knowledge state (Tan et al., 2017) in finding duration of calendar time. Besides that, past studies (e.g., Chin et al., 2021b, 2022) also focused on developing assessments to measure students' attribute mastery level for addition and multiplication of time involving *hours, days, weeks, months, and years.* Meanwhile, several interventions have been introduced by the researchers to support students' learning of time-telling (Earnest, 2021; Pelton et al., 2018; Wang et al., 2016) as well as the learning of time concepts involving *years, decades, and centuries* (Chin et al., 2021a).

For the aspect of subtraction of time, the past studies mainly focused on error analysis (i.e., Earnest, 2015; Kamii & Russell, 2012; Ojose, 2015). The findings constantly reported that most of the students tend to make regrouping errors when performing subtraction of time. For instance, they tend to subtract the measurement of time without performing necessary regrouping (Kamii & Russell, 2012). Even though some of the students understand the concept of regrouping, Earnest (2015) and Ojose (2015) found that the students tend to make mistakes in regrouping the time notation. Besides confusing the time notation with the base ten number system (Earnest, 2015), the students also regrouped the time notation using the wrong time relationship (Ojose, 2015). For example, the students might regroup 1 day into 12 hours rather than 24 hours.

1.1.6. *The present study*

'Subtraction of Time' is regarded as a difficult skill to be mastered by students. Despite the importance of subtraction of time involving years and centuries in determining a longer duration between two given time points, the previous relevant studies mainly concentrated on the frequently used time unit, such as *hours* and *minutes* (e.g., Earnest, 2015; Kamii & Russell, 2012; Sia et al., 2019). The study focused on the subtraction of time involving *years*, and *centuries*, which are rarely found in the literature (Chin et al., 2021a).

To determine the persistence made by the students in subtraction involving time, several studies (i.e., Earnest, 2015; Kamii & Russell, 2012; Ojose, 2015) have been conducted by performing error analysis. However, this approach fails to pinpoint the underlying cognitive attribute deficit which leads to the errors made (Ketterlin-Geller & Yovanoff, 2009). Consequently, the students' procedural errors were usually corrected without considering the conceptual understanding (Russell & Masters, 2009) which provides strong support for the development of procedural knowledge and permits the extension of the mathematical idea (Rittle-Johnson & Schneider, 2015).

In this regard, the use of cognitive models in test development and result interpretation could provide informative diagnostic data for supporting the teachers in planning the remedies to support students' acquisition of 'subtraction of time'. Yet, the available cognitive models which can be used to support the highly specific diagnostic inference made are rarely found in the literature (Gierlet al., 2009a; Sia et al., 2019). Besides that, the cognitive process of performing subtraction of time involving years and centuries is left unexplored in the past.

To fill the research gap, this study sought to develop and validate the cognitive models for 'Subtraction of Time' involving years and centuries. In this paper, we present the process of developing the cognitive models. To ensure the validity of the diagnostic claims made, we validate the cognitive models developed by addressing the following research questions:

(1) To what extent are hierarchical arrangement of attribute in expert-based cognitive models supported by students' responses?
(2) To what extent are the attribute dependency in the expert-based cognitive models supported by students' responses?
(3) To what extent are the expert-based cognitive models consistent with the student-based cognitive models for 'Subtraction of Time'?

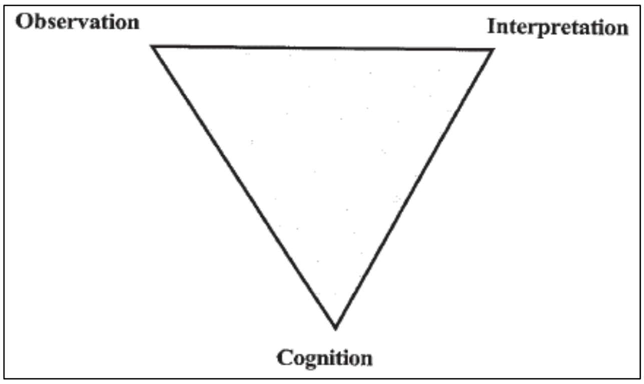
1.2. Theoretical Framework

1.2.1. *Assessment triangle*

The development and validation of cognitive models were grounded in the framework called Assessment Triangle (Pellegrino et al., 2001). This framework explains the mechanism of linking educational measurement with human cognition by using a triangle as illustrated in Figure 2. The three basic elements of assessment, namely (i) a cognitive model which illustrates the students' skills or knowledge acquisition in the tested domain, (ii) the task which triggers students' response to manifest their skills or knowledge and (iii) the interpretation method used to make diagnostic inferences (Pellegrino & Chudowsky, 2003) are pivoted on each vertex of the Assessment Triangle namely, cognition, observation, and interpretation, respectively. Following this, each element is linked to the other two elements and works in synchrony. Hence, students' cognition can be used to explain their strengths and weaknesses (Pellegrino et al. 2001).

The cognitive models which are embedded in the cognition vertex of the Assessment Triangle were developed by using the top-down approach in this study. Since the attributes which form the cognitive models are considered latent traits that are non-observable (Keehner et al., 2017), the assessment tasks were developed to elicit students' responses which could demonstrate their mastery of attributes. These assessment tasks are pivoted to the observation vertex of the Assessment Triangle. While the cognitive models represent the hierarchically ordered attributes, fitting the cognitive-psychometric model in the interpretation vertex, such as Attribute Hierarchy Method (AHM) onto students' responses collected in the observation vertex can be used to validate the cognitive models. (Leighton et al., 2004).

Figure 2. *Assessment triangle (Pellegrino & Chudowsky, 2013, p. 112).*



2. METHOD

A descriptive research design was adopted for the empirical validation of the cognitive models developed, which is predominantly descriptive. In this section, we discuss the process of developing the cognitive models, followed by describing the participants, research instruments, and the research procedure of the empirical validation of the cognitive models constructed.

2.1. Participants

Since the development of the cognitive model is an iterative process, the participants of the study were selected by employing convenience sampling, which is commonly used for piloting an under-developed instrument (Salkind, 2010) because of its main advantage in terms of cost efficiency. A total of 119 Grade Five students from one National School (NS), one National-Type Chinese School (NTCS) and one National-Type Tamil School (NTTS) in Penang, Malaysia with the Malay, Mandarin, and Tamil language as the medium of mathematics instruction, respectively were chosen to participate in the study. Since class streaming is no longer practised in the schools, each class consisted of students with mixed abilities. In order to ensure the representativeness of the data, the intact class of the students were chosen. With the sample size of 119, which surpassed the minimum sample size required (n=100) for employing the psychometric model named AHM, which is rooted in latent class analysis (Wrupts & Geiser, 2014), the findings of the study could be reliable.

2.2. Development of Expert-Based Cognitive Models

The development of expert-based cognitive models started with the identification of attributes through task analysis and expert reviews following the suggestions were given by Gierl et al. (2010). A workshop which involved two invited mathematics education experts was conducted to specify the fine-grained attributes through task analysis and expert reviews in a workshop. The background of the two experts is presented in Table 1.

Table 1. *Background of mathematics education experts.*

Expert	Academic Qualification	Specialization	Position	Affiliation
Expert 1	Doctor of Philosophy	Mathematics Education	Associate Professor	Public University in Malaysia
Expert 2	Doctor of Philosophy	Mathematics Education	Associate Professor	Public University in Malaysia

During the workshop, the two experts reviewed Year Four Mathematics Textbook as well as the Curriculum and Assessment Standard Document to deepen their understanding of the learning standards related to the intended construct, that is 'Subtraction of Time'. Then, they listed the main skills about 'Subtraction of Time' as tabulated in Table 2 based on the learning standards.

Table 2. *Main skills related to 'Subtraction of Time'.*

Main Skill	Description
Main Skill 1	Subtraction of time involving century and year without regrouping
Main Skill 2	Subtraction of time involving century and year with single regrouping
Main Skill 3	Subtraction of time involving century and year with double regrouping

After that, task analysis was performed by the two experts on the task selected from the textbook based on each main skill as demonstrated in Figure 3. During the task analysis, each step involved in solving the tasks is depicted as a detailed description. Based on this description, the experts outlined the attributes which can be measured using the test items (Alves, 2012). For example, the attribute *'Convert 1 century to 100 years, add the 100 years into the number of years in the first minuend and subtract the number of centuries and years in the subtrahend from the first minuend'* is used to summarize the description of steps: (i) '*Borrow 1 century from the century column in the first minuend*'; (ii) '*Convert 1 century to 100 years and add the 100 years into the number of years in the first minuend';* and (iii) '*Subtract the number of centuries and years in the first subtrahend from the first minuend*'. Notably, this attribute could barely be measured using an item because it is less precise. Thus, it was rephrased into a clearer version such as '*Subtract one unit of time from one unit of time involving century and year with regrouping*'. The modified attributes are shown in the square brackets in Figure 3.

To ensure the instructional relevance of the attributes, a panel of subject matter experts (SMEs) were invited to validate the attributes. The background of the SMEs is shown in Table 3. All attributes were rated as '5' by each pair of experts on the 5-point Likert-scale validation form. With the simple agreement of 100 percent, the relevancy of the attributes with respect to the content standards and learning standards was very high.

Table 3. *Background of the subject matter experts.*

Expert	Academic Qualification	Specialization	Position	Affiliation
Expert 1	Master in Education	Mathematics Education	Experienced Teacher	NS in Malaysia
Expert 2	Master in Education	Mathematics Education	Experienced Teacher	NS in Malaysia
Expert 3	Master in Education	Mathematics Education	Experienced Teacher	NTCS in Malaysia
Expert 4	Master in Education	Mathematics Education	Experienced Teacher	NTCS in Malaysia
Expert 5	Master in Education	Mathematics Education	Experienced Teacher	NTTS in Malaysia
Expert 6	Doctor of Philosophy	Mathematics Education	Experienced Teacher	NTTS in Malaysia

After the validation process, the attributes were ordered hierarchically by the two mathematics experts, based on the complexity of each attribute to derive the attribute hierarchy. For example, the attribute *'Subtract one unit of time from one unit of time involving century and year with regrouping'* was positioned at the lowest level of the hierarchy since it is less complex compared to the attribute *'Subtract two units of time from one unit of time involving century and year with single regrouping'*. A total of six attribute hierarchies related to 'Subtraction of Time' as shown in Table 4 were specified by the experts in the field of mathematics education. These attribute hierarchies are considered as Expert-based Cognitive Models (E-CM).

Table 4. *Three attribute hierarchies related to 'Subtraction of Time'.*

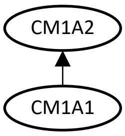
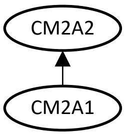
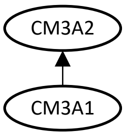
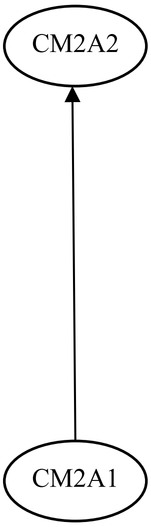
Cognitive Model	Attribute Hierarchy	Attributes
Cognitive Model 1		CM1A2: Subtract two units of time from one unit of time involving century and year without regrouping. CM1A1: Subtract one unit of time from one unit of time involving century and year without regrouping.
Cognitive Model 2		CM2A2: Subtract two units of time from one unit of time involving century and year with single regrouping. CM2A1: Subtract one unit of time from one unit of time involving century and year with regrouping.
Cognitive Model 3		CM3A2: Subtract two units of time from one unit of time involving decade and year with double regrouping. CM3A1: Subtract one unit of time from one unit of time involving century and year with regrouping.

Figure 3. *Task analysis for Main Skill 2.*

Working	Description of the steps	Attributes	Attribute Hierarchy
$\begin{array}{clll} & \text{Century} & \text{Year} & \\ \text{①} & {}^{7}\cancel{8} & \begin{matrix}100+6\\=106\end{matrix} & \text{②} \\ & & \cancel{6} & \\ - & 2 & 7 & \text{③} \\ \hline & 5 & 99 & \\ - & & 9 & \text{④} \\ \hline & 5 & 90 & \end{array}$	① Borrow 1 century from the decade column in the first minuend. ② Convert 1 century to 100 years and add the 100 years into the number of years in the first minuend. ③ Subtract the number of cenrturies and years in the first subtrahend from the first minuend. ④ Subtract the number of centuries and years in the second subtrahend from the first minuend.	**CM2A1**: Convert 1 century to 100 years, add the 100 years into the number of years in the first minuend and subtract the number of centuries and years in the subtrahend from the first minuend. [Subtract one unit of time from one unit of time involving century and year with regrouping] **CM2A2**: Subtract the number of centuries and years in second subtrahend from the second minuend. [Subtract two units of time from one unit of time involving century and year with single regrouping]	

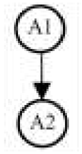
To ensure the appropriateness of the sequence of the attributes being ordered, these attribute hierarchies underwent validation that involved six subject matter experts. All attribute hierarchies were rated as '5' by each pair of experts on the 5-point Likert-scale validation form. With the simple agreement of 100 percent, the arrangement of the attribute hierarchies was considered very appropriate. These validated attribute hierarchies were regarded as expert-based cognitive models (Sia et al., 2019).

In order to validate the cognitive models, a matrix-formed test specification, named reduced Q matrix (**Q$_r$** matrix), was derived from the expert-based cognitive models. The hypothesized attributes that need to be mastered in order to answer each test item correctly are depicted in the reduced Q-matrix (Li & Suen, 2013). The **Q$_r$** matrix derivation process is illustrated in Figure 4. The derivation of **Q$_r$** matrix began with using the second order binary square matrix named *adjacent matrix* (**A** matrix) to specify the direct attributes' relationship in the hierarchy. In the A matrix, the presence or absence of the direct attributes' relationships was represented using '1' or '0' respectively. Then, the direct and indirect attributes relationships in the hierarchy were represented using the second order binary square matrix, named *reachability matrix* (**R** matrix) derived by applying Boolean arithmetic following the formula $\mathbf{R} = (\mathbf{A} + \mathbf{I})^n$, where **I** refers to the *Identity matrix* and *n* refers to the smallest integer needed to obtain a constant **R** matrix. After that, the number of potential items (*i*) was determined by employing the formula $i = 2^k - 1,$ where *k* indicates the number of attributes.

Then, the incidence matrix (**Q** matrix) was derived to portray the attribute combinations which might be involved in solving each potential item correctly. Then a further derivation was conducted to reduce the **Q** matrix of order 2 × 3 into a binary second order squared matrix named reduced incidence matrix (**Q$_r$** matrix) by establishing the direct and indirect attribute relationships following its specification shown in the **R** matrix. For instance, the removal of the second column of **Q** matrix was made due to the fact that the items that involve attribute CM2A2, would also involve attribute CM2A1 indirectly. In other words, none of the items could be used to measure solely the attribute CM2A2, without measuring attribute CM2A1 indirectly. After deriving the **Q$_r$** matrix, the cognitive diagnostic assessment (CDA) can be constructed to collect the empirical data for validating the expert-based cognitive models developed.

Figure 4. *The derivation of the Qr matrix (Chin et al., 2021b, p. 300).*

Attribute Hierarchy



- The attributes A1, and A2 are arranged in a linear hierarchical order based on their complexity to form the attribute hierarchy.

A matrix

$$\begin{array}{c|cc} & A1 & A2 \\ \hline A1 & 0 & 1 \\ A2 & 0 & 0 \end{array}$$

- ***A*** matrix is used to represent the direct relationship of the attributes in cognitive diagnostic assessment (Tatsuoka 1986).
- The presence and absence of the direct relationship between attributes are represented as ‘1’ and ‘0’ respectively in the ***A*** matrix.

R matrix

$$\begin{array}{c|cc} & A1 & A2 \\ \hline A1 & 1 & 1 \\ A2 & 0 & 1 \end{array}$$

- ***R*** matrix is used to represent the direct and indirect prerequisite relationship among the attributes in cognitive diagnostic assessment (Tatsuoka 1986).
- The presence or absence of the direct and indirect relationship between attributes is represented as ‘1’ and ‘0’ respectively in the ***R*** matrix.
- ***R*** matrix can be derived from ***A*** matrix by performing Boolean addition and multiplication using formula R = $(\boldsymbol{A}+\boldsymbol{I})^n$, where ***I*** refers to the identity matrix and n refers to the integer required to reach invariance.

Q matrix

$$\begin{array}{c|ccc} & Q1 & Q2 & Q3 \\ \hline A1 & 1 & 0 & 1 \\ A2 & 0 & 1 & 1 \end{array}$$

- The incidence matrix (**Q** matrix) is the potential pool of items which represent all combination of attributes when the attributes are assumed to be not related to each other.
- The number of potential items can be determined using the formula
  $i = 2^k$-1, where k is the number of attributes

Q

$\boldsymbol{Q_r}$ matrix

$$\begin{array}{c|cc} & Q1 & Q3 \\ \hline A1 & 1 & 1 \\ A2 & 0 & 1 \end{array}$$

- According to Gierl, Leighton, and Hunka.(2000), the reduced incidence matrix ($\boldsymbol{Q_r}$) represents the items from the potential pool that satisfy the attribute hierarchy as shown in the cognitive model.
- ***Q*** matrix can be reduced to form the $\boldsymbol{Q_r}$ matrix by imposing the prerequisite relation of the attributes as specified in the ***R*** matrix (Gierl et al. 2000; Tatsuoka 1991, 2009).

2.3. Research Instrument

The validity evidence of the cognitive models was collected by using the CDA for ‘Subtraction of Time’ which was developed based on the $\mathbf{Q_r}$ matrix. Each combination of attributes depicted at each column of the $\mathbf{Q_r}$ matrix was probed using three parallel open-ended items as recommended by Sia and Lim (2018) to enhance the reliability of the CDA (Gierl et al., 2009). Following this, the CDA for ‘Subtraction of Time’ consisted of 18 items to elicit student responses for the two attribute combinations as shown in each $\mathbf{Q_r}$ matrix of the six cognitive models (2 cognitive models × 2 attribute combinations × 3 parallel items = 18 items). Corresponding to the six cognitive models, the CDA consisted of 3 sections as listed in Table 5. These English-written items were then translated into Malay, Mandarin and Tamil languages to comply with the instruction medium of mathematics lessons in NS, NTCS and NTTS, respectively.

Table 5. *Content of CDA.*

Section	Cognitive Model	Skill	Number of Items
Section A	Cognitive Model 1	Subtraction of time involving century and year with no regrouping	6
Section B	Cognitive Model 2	Subtraction of time involving century and year with single regrouping	6
Section C	Cognitive Model 3	Subtraction of time involving century and year with double regrouping	6
		Total	18

To ensure the content validity of the CDA, the two subject matter experts with at least 10 years of teaching experience each from NS, NTCS and NTTS were invited to validate the instrument after the translation process. All items were rated as ‘5’ by the six experts on the 5-point Likert-scale validation form. With the content validity index of 1.00 at the scale level, all items in the CDA were highly relevant with respect to the corresponding attribute combination measured (Polit & Beck, 2006). After the validation process, the CDA was piloted using 32 Year Five NS pupils, 35 Year Five NTCS pupils and 15 Year Five NTTS pupils selected through convenience sampling, for determining the reliability of the instrument. Although the CDA comprised a set of open-ended items, it was scored dichotomously in accordance with the use of AHM as the psychometric model (Wang & Gierl, 2011). With the reliability coefficient of .90 which was calculated using Kuder Richardson 20, the dichotomously scored open-ended CDA was reliable (Multon & Coleman, 2010).

2.4. Research Procedures

The CDA was administered to the 119 Grade Five students in one NS, one NTCS, and one NTTS located in Penang, Malaysia. No time limit was imposed upon the test because the CDA was not served for students' performance comparison. After the test administration, the answer scripts were scored dichotomously. For each item, one mark was awarded to the correct response, while no mark was awarded to the incorrect response.

The pupils’ responses were then further analysed by applying AHM. Specifically, Artificial Neural Network (ANN) pattern recognition analysis (PRA) was performed using Statistical Package for Social Sciences (SPSS) version 24 to estimate the pupil’s attribute probability that corresponded with their response patterns in CDA. The ANN PRA is a two-stage data analysis process. During the first stage, the training of ANN was conducted so that the expected response pattern (ERP) could be associated with the corresponding expected attribute pattern (EAP) as shown in Table 6. To prevent the arisen of the issue regarding model-underfit, the ANN training data set consisted of the data made up of 100 replications of each ERP Vector and the corresponding EAP Vector pairs (Briggs & Kizil, 2017) as shown in Table 6. Following this, the ANN was trained with 300 samples (3 ERP-ERP pairs × 100 times of replication = 300 samples) by using the gradient transient backpropagation algorithm (Cui et al., 2016). With the architecture of 6 input nodes, 2 hidden nodes, and 2 output nodes, the error of the ANN converged at nearly zero (Root Mean Squared Error = .0088), which is acceptable (Cui et al., 2016). This indicates the relationship between the ERP Vectors and the corresponding EAP Vectors has been established. Following this, the pupils' attribute probabilities were estimated using the trained ANN at the second stage of ANN PRA.

Table 6. *Expected response pattern and expected attribute pattern.*

Expected Response Pattern	Expected Attribute Pattern
[0 0 0 0 0 0]	[0 0]
[1 1 1 0 0 0]	[1 0]
[1 1 1 1 1 1]	[1 1]

To determine the hierarchical arrangement of the attributes in the student-based cognitive models, item difficulty modelling was performed for each section of the assessment. Since the sample size of the study failed to meet the minimum requirement for performing Rasch Analysis (n=250) suggested by Linacre (1994), the item difficulty of each item was computed following the Classical Test Theory. Then, the position of each attribute in the hierarchy was determined based on the mean difficulty of the item measuring each attribute. Meanwhile, the mean attribute probabilities were used to confirm the hierarchical arrangement of the attributes in the S-CM. Then, the dependency of the attributes in the cognitive models was determined by referring to the correlation between the attribute probabilities. After portraying the S-CM, the Hierarchical Consistency Index (HCI) for each cognitive model based on the formulae as shown below, by using Microsoft Excel 2016 to determine the extent to which the S-CM are consistent with the E-CM.

$$HCI_i = 1 - \frac{2\sum_{j\in S_{correct_i}}\sum_{g\epsilon S_j} X_{i_j}(1 - X_{i_g})}{N_{c_i}}$$

Where,

$S_{correct_i}$ is a set which consists of the items that are correctly answered by the student *i*

X_{i_j} is the score (1 or 0) of student *i* for the item *j,* where item *j* is an element in the set $S_{correct_i}$

S_j is a set which consists of the items which required subset of attributes measured by item *j,* where item $j \notin S_j$

X_{i_g} is the score (1 or 0) of student *i* for the item *g,* where item *g* is an element in the set S_j

N_{c_i} is the total number of comparisons for all the items that are correctly answered by the student *i*

(Cui & Leighton, 2009, p. 436)

3. FINDINGS

3.1. Hierarchical Arrangement of Cognitive Models

Item difficulty modelling was used to verify the position of attributes in the cognitive models based on student responses. The item difficulty index, attribute-level mean item difficulty, and attribute-level z-score was computed for dispersion of the data. As shown in Table 7, the attribute level z-scores ranged from -1.11 to 1.15. This indicates that the difficulty of each item only differs by nearly one standard deviation of the respective attribute-level mean item difficulty. In other words, the three parallel items measuring the same attribute were calibrated at almost the same level of difficulty. As shown in Table 7, the p-value range of the first three items did not overlap with the p-value range of the last three items in each section. In other words, the items which measured the same attribute were grouped into a cluster calibrated with different complexity. This suggests the existence of a clear distinction in terms of the complexity between each attribute pair (i.e., CM1A1-CM1A2, CM2A1-CM2A2, and CM3A1-CM3A2) in the cognitive models. Thus, the two attributes in each cognitive model were not at the same level of hierarchy.

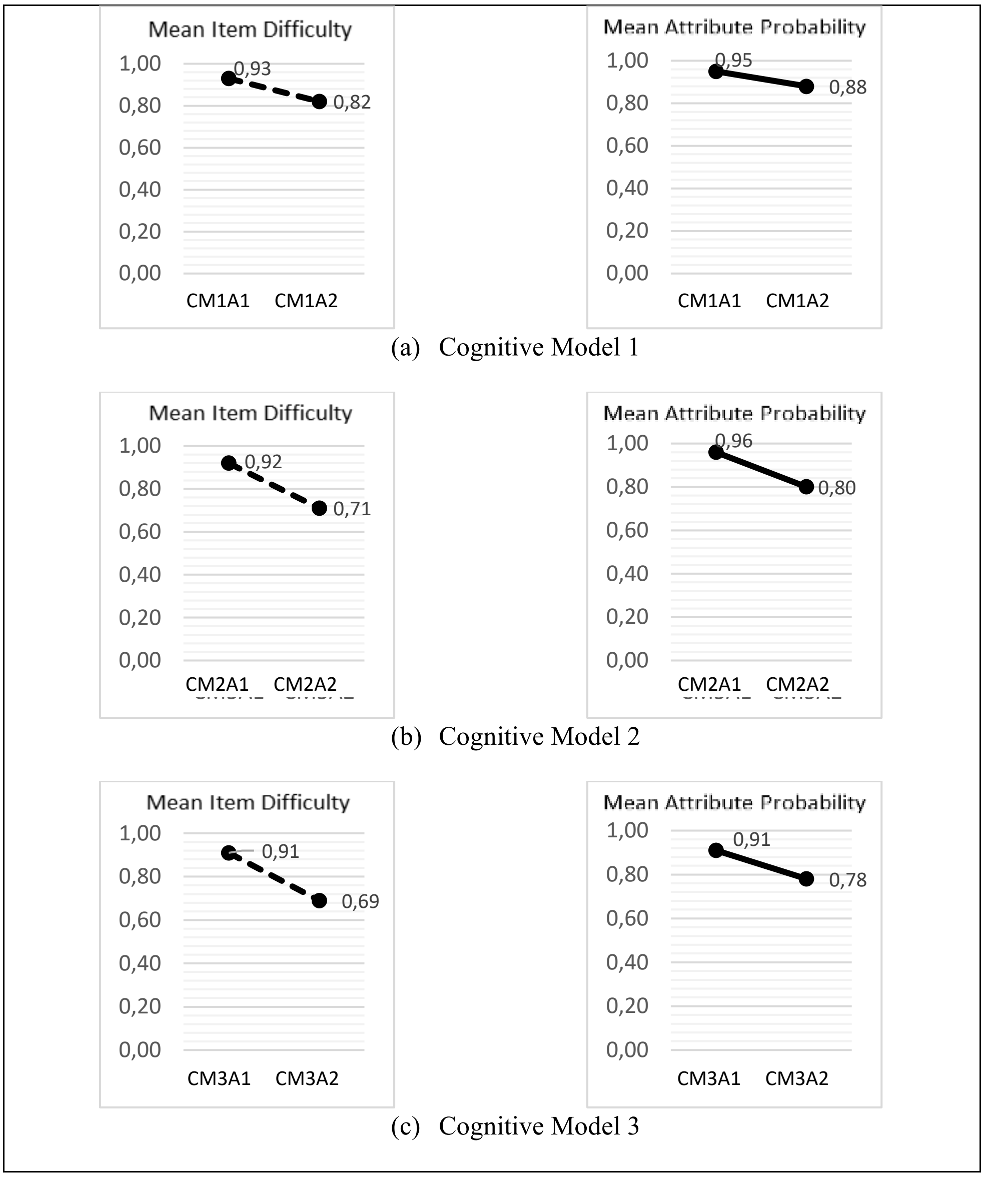
Table 7. *Item difficulty index, attribute-level mean item difficulty, and attribute-level z-score.*

Section [Cognitive Model]	Item	Item Difficulty Index (p-value)	Attribute Level Mean Item Difficulty Index	Attribute Level Z-score
Section A [Cognitive Model 1]	CM1A1		0.93	
	Item 1	0.94		1.15
	Item 2	0.92		-0.58
	Item 3	0.92		-0.58
	CM1A2		0.82	
	Item 4	0.84		1.06
	Item 5	0.84		1.06
	Item 6	0.79		1.00
Section B [Cognitive Model 2]	CM2A1		0.92	
	Item 1	0.89		-1.11
	Item 2	0.93		0.28
	Item 3	0.95		0.83
	CM2A2		0.71	
	Item 4	0.77		1.32
	Item 5	0.77		1.32
	Item 6	0.58		1.00
Section C [Cognitive Model 3]	CM3A1		0.91	
	Item 1	0.92		1.00
	Item 2	0.91		0.00
	Item 3	0.89		-1.00
	CM3A2		0.70	
	Item 4	0.74		1.06
	Item 5	0.66		0.94
	Item 6	0.70		1.00

Then, the attribute level mean item difficulty was compared for each section in order to determine the hierarchical position of the attributes in each cognitive model. The attribute level mean item difficulty was tabulated in Table 7. For each cognitive mode, the first attribute (i.e., CM1A1, CM2A1, and CM3A1) has a higher mean item difficulty compared to the second attribute (i.e., CM1A2, CM2A2, and CM3A2). This implies more students answered the items that probe the first attribute in each cognitive model correctly on average. In other words, the first attribute of each cognitive model is more basic than the second attribute of each cognitive model. Thus, the first attribute of each cognitive model was placed at a lower position in the hierarchy structure, compared to the second attribute in each cognitive model.

Then, the comparison of the mean item difficulty and mean attribute probability trend was made to further confirm the cognitive models' hierarchical structure. The results of the comparison were illustrated in Figure 5.

Figure 5. *Comparison of mean item difficulty and mean attribute probability.*



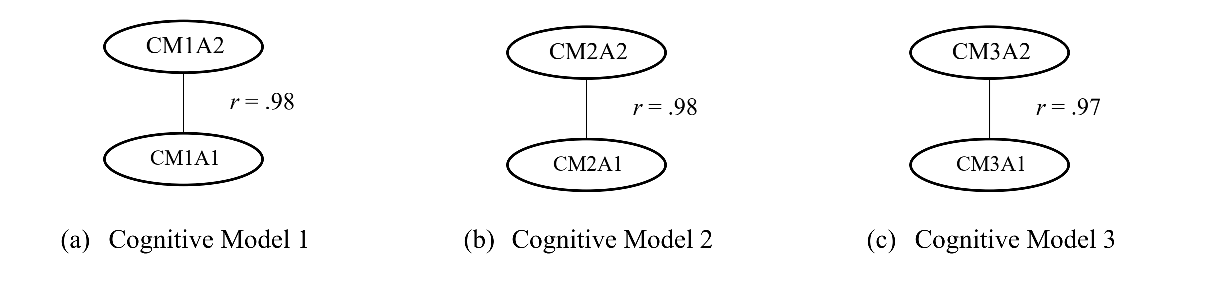
(a) Cognitive Model 1

(b) Cognitive Model 2

(c) Cognitive Model 3

As shown in Figure 5, both the mean item difficulty and mean attribute probability of each cognitive model showed a similar trend. The mean item difficulty of the second attribute [CM1A2: 0.82; CM2A2: 0.71; CM3A2: 0.69] of each cognitive model was lower than that of the first attribute [CM1A1: 0.93; CM2A2: 0.92; CM3A2: 0.91] in each cognitive model. Likewise, the mean attribute probability of the second attribute [CM1A2: 0.88; CM2A2: 0.80; CM3A2: 0.78] of each cognitive model was lower than that of the first attribute [CM1A1: 0.95; CM2A2: 0.96; CM3A2: 0.91] in each cognitive model. This indicates that the mean item difficulty and mean attribute probability decrease as the attribute gets more complex. Besides, this also reflects the linear-shaped cognitive models' hierarchical structure (Alves, 2012) as portrayed in Figure 6.

Figure 6. *Linear hierarchical structure of cognitive model based on the students' responses.*



(a) Cognitive Model 1

(b) Cognitive Model 2

(c) Cognitive Model 3

As shown in Figure 6, the attributes with lower complexity (i.e., CM1A1, CM2A1, and CM3A1) are positioned at the lower hierarchy in the corresponding cognitive models. Meanwhile, the attributes with higher complexity (i.e., CM1A2, CM2A2, dan CM3A2) are positioned at the higher hierarchy in the corresponding cognitive models. Since the two attributes in each cognitive models are related to each other, they are linked together using a straight line and form a linear hierarchical structure. These linear hierarchical structures are also considered as student-based cognitive models (S-CM).

3.2. Dependency among the Attributes in the Cognitive Models

To verify the attribute dependency in each cognitive model, a correlation analysis of the attribute probability for each attribute pair in the cognitive model was conducted. Since the attributes CM1A1, CM2A1, and CM3A1 were negatively skewed ($\text{Skewness}_{\text{CM1A1}}$ = -2.65; $\text{Skewness}_{\text{CM2A1}}$ = -4.78; $\text{Skewness}_{\text{CM3A1}}$ = -2.89), the correlational relationship between each pair of attributes in the six cognitive models were analysed by using Spearman Rank Correlation Coefficient. As illustrated in Figure 6, the correlation coefficients ranged from .97 to .98 indicating that the attribute pair in each cognitive model exhibited a strong positive correlational relationship at the significant level of .05 (Pallant, 2016). This implies that there exists a dependency relationship between the attribute pair in each cognitive model. Hence, the two attributes in each of the cognitive models were positioned next to each other in the attribute hierarchy.

3.3. Consistency between Student-Based and Expert-Based Cognitive Models

The overall consistency between S-CM and E-CM was evaluated based on the HCI computed. Instead of the mean HCI as suggested by Alves (2012), the median HCI was used to represent the heavily left-skewed HCI distributions of the cognitive models constructed in the study with the skewness coefficient ranging from -1.98 to -2.31. The median of HCI for each cognitive model was reported in Table 8. With the median of HCI surpassing the cut score of .80, the six cognitive models derived in this study exhibited excellent fit (Cui et al., 2016). This indicates that the S-CM were highly consistent with the E-CM.

Table 8. *Overall consistency between student-based and expert-based cognitive models.*

Cognitive Model	Skewness	*Md*	Interquartile Range	Interpretation
Cognitive Model 1	-2.08	1.00	(0.64, 1.00)	Excellent
Cognitive Model 2	-1.98	1.00	(0.64, 1.00)	Excellent
Cognitive Model 3	-2.31	1.00	(0.75, 1.00)	Excellent

4. DISCUSSION and CONCLUSION

4.1. To What Extent are Hierarchical Arrangement of Attributes in Expert-Based Cognitive Models Supported by Students' Responses?

The findings indicate that the cognitive models' hierarchical structure developed from the expert perspective was affirmed based on the decreasing trend of mean item difficulty and mean attribute probability which was computed based on the students' responses. This is expected because the attributes were arranged by the experts in increasing order of complexity following the claim made by Iuculano et al. (2018) whereby the mathematics skills are acquired following hierarchical sequences. Since the more basic pre-requisite skill serves as the foundation for mastering a new skill (Iuculano et al., 2018). The new skill is relatively complex. With the increasing attribute complexity, attribute acquisition becomes tougher, and the items become more difficult for the students (Morrison & Embretson, 2014). Thus, the probability of mastering the attributes would be decreased and fewer students would be able to answer the related items correctly. The decreases in mean item difficulty and mean attribute probabilities with the increase of attribute complexity further confirmed the structure of the linearly ordered attribute hierarchy as specified in the E-CMs illustrated in Table 4.

4.2. To What Extent are the Attribute Dependency in The Expert-Based Cognitive Models Supported by Students' Responses?

Meanwhile, the correlational analysis of the attribute probabilities provided convergence evidence to support the hierarchical structure of each cognitive model. This finding is expected because the pre-requisite relationships between the two attributes were exhibited between the two attributes (Sia, 2017). For each cognitive model, the students who must master the second attribute that is more complex are more likely to master the first attribute which is more basic. Since the attribute pair in all cognitive models exhibited a strong positive correlation, we verified the cognitive models' hierarchical structure with the two attributes being positioned adjacent to each other (Sia, 2017).

4.3. To What Extent are the Expert-Based Cognitive Models Consistent with The Student-Based Cognitive Models For 'Subtraction of Time'?

The findings also reveal a high consistency between the S-CMs and E-CMs. This could be due to the use of the curriculum standards and the textbook as the main resources for guiding the attribute specification (Sia et al., 2019). The process of performing the subtraction involving decades and years as well as centuries and years captured in the task analysis is almost similar to the steps shown in the textbook which serve as the main reference in mathematics teaching and learning in the classroom. This could explain the reason underlying the excellent fit among students' responses and the experts' predictions.

4.4. Implications of Study

A total of three cognitive models for 'Subtraction of Time' has been developed in this study. While the attributes' pre-requisite relationship is illustrated in cognitive models, the cognitive models constructed in this study would suggest the instructional sequence for 'Subtraction of time' involving centuries and years' which could foster both conceptual understanding and procedural fluency.

Based on the cognitive models developed, the students should be exposed to the subtraction without regrouping involving measurement of time with the composite unit (i.e., *centuries* and *years*) which serve as an extension from the base-ten subtraction learned in the early grade. Then, the teachers should help the students to recall the relationship between the units of time. Moreover, the underlying reasoning behind regrouping should be explained explicitly when introducing the subtraction of time involving regrouping. With a stronger conceptual understanding, students would have a better procedural fluency in subtraction of time involving *centuries* and *years.*

Besides proposing the instructional sequence, the cognitive models constructed also could be valid to guide the assessment development and make diagnostic inferences related to students' performance on 'Subtraction of Time' involving *centuries* and *years*. This informative diagnostic data would eventually support teachers in the remedial intervention planning for helping the students in overcoming their cognitive weaknesses and thereby foster their mastery of performing subtractions involving these time units.

4.5. Conclusion

To illustrate the cognitive process in performing subtraction of time involving *centuries* and *years,* a total of three cognitive models have been developed in this study. The findings of this study warrant the quality of the cognitive models developed with highly convincing validity evidence which are. Perhaps the findings would encourage the use of cognitive models in guiding the instructional sequence, assessment development and students' result interpretation to support the mathematics teaching and learning for 'Subtraction of Time' involving *centuries* and *years*.

4.6. Limitations and Recommendations

This study is subject to some limitations. Because of the practical constraint, the sample was selected using convenience sampling. Thus, the generalisability of the findings could be reduced. Besides that, the small sample size restricted the choice of the psychometric model used to measure the item difficulty. This eventually reduces the robustness of the findings. To address this limitation, the probabilistic sampling technique is recommended to be used for selecting the larger samples in future studies so that a more robust psychometric model can be applied to calibrate the item difficulty, and the findings would be more generalisable.

Acknowledgments

This study was made possible with funding from the Research University Grant (RUI) Scheme 1001/PGURU/8011027 of Universiti Sains Malaysia, Penang, Malaysia. The authors would like to thank all the teachers and students who voluntarily participated in this study.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Human Research Ethics Committee of Universiti Sains Malaysia, USM/JePEM/18030175.

Authorship Contribution Statement

Huan Chin: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing - original draft. **Cheng Meng Chew**: Conceptualization, Methodology, Formal analysis, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Orcid

Huan Chin https://orcid.org/0000-0003-0991-7299
Cheng Meng Chew https://orcid.org/0000-0001-6533-8406

REFERENCES

Akbay, L., Terzi, R., Kaplan, M., & Karaaslan, K.G. (2018). Expert-based attribute identification and validation: An application of cognitively diagnostic assessment. *Journal on Mathematics Education*, *9*, 103-120.

Alves, C.B. (2012). *Making diagnostic inferences about student performance on the Alberta education diagnostic mathematics project: An application of the Attribute Hierarchy Method.* (Publication No. 919011661) [Doctoral Thesis, University of Alberta, Ann Arbor, Canada]. ProQuest Dissertations and Theses database.

Australian Curriculum Assessment and Reporting Authority [ACARA]. (2017). *Numeracy learning progression and history*. https://www.australiancurriculum.edu.au/media/3666/numeracy-history.pdf

Brace, N., Doran, C., Pembery, J., Fitzpatrick, E., & Herman, R. (2019). Assessing time knowledge in children aged 10 to 11 years. *International Journal of Assessment Tools in Education*, *6*(4), 580-591.

Briggs, D.C., & Kizil, R.C. (2017). Challenges to the use of artificial neural networks for diagnostic classifications with student test data. *International Journal of Testing*, *17*(4), 302-321.

Carpenter, T.P., Fennema, E., Franke, M.L., Levi, L., & Empson, S.B. (1999). *Children's mathematics: Cognitively guided instruction.* Heinemann.

Chan, Y.L. (2017). *Super skills modul aktiviti integrasi: Mathematics Year 5 KSSR.* Sasbadi.

Chan, Y.L., Maun, R., & Krishnan, G. (2017). *Dual language programme mathematics Year 5 textbook.* Dewan Bahasa dan Pustaka.

Chen, F., Yan, Y., & Xin, T. (2017). Developing a learning progression for number sense based on the rule space model in China. *Educational Psychology, 37*(2), 128-144.

Chin, H., Chew, C.M., & Lim, H.L. (2021). Development and validation of online cognitive diagnostic assessment with ordered multiple-choice items for 'Multiplication of Time'. *Journal of Computers in Education*, *8*(2), 289-316.

Chin, H., Chew, C.M., & Lim, H.L. (2021b). Development and validation of online cognitive diagnostic assessment with ordered multiple-choice items for 'Multiplication of Time'. *Journal of Computers in Education*, *8*(2), 289-316.

Chin, H., Chew, C.M., Lim, H.L., & Thien, L.M. (2022). Development and validation of a cognitive diagnostic assessment with ordered multiple-choice items for addition of time. *International Journal of Science and Mathematics Education*, *20*(4)*,* 817-837.

Clements, D.H., Sarama, J., Baroody, A.J., & Joswick, C. (2020). Efficacy of a learning trajectory approach compared to a teach-to-target approach for addition and subtraction. *ZDM Mathematics Education*, *52*, 637–648.

Cui, Y., & Leighton, J.P. (2009). The Hierarchy Consistency Index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*(4), 429-449.

Cui, Y., Gierl, M., & Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology*, *36*(6), 1065-1082.

Earnest, D. (2015). When "half an hour" is not "thirty minutes": Elementary students solving elapsed time problem. In T.G. Bartell, K.N. Bieda, R.T. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the 37th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 285-291). Michigan State University.

Earnest, D. (2021). About time: Syntactically-guided reasoning with analog and digital clocks. *Mathematical Thinking and Learning*. [Advance Online Publication].

Fuson, K.C. (1990). Conceptual structures for multiunit numbers: Implications for learning and teaching multidigit addition, subtraction, and place value. *Cognition and Instruction*, *7*(4), 343-403.

Gierl, M.J., Alves, C., & Taylor-Majeau, R. (2010). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*(4), 318-341.

Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2000). An NCME instructional module on exploring the logic of Tatsuoka's Rule-Space Model for test development and analysis. *Educational Measurement: Issues and Practice*, *19*(3), 34-44.

Gierl, M.J., Leighton, J.P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2009a). *Validating cognitive models of task performance in algebra on the SAT* (College Board Research 2009-3). The College Board.

Gierl, M.J., Roberts, M.P.R., Alves, C., & Gotzmann, A. (April, 2009b). *Using judgments from content specialists to develop cognitive models for diagnostic assessments*. Paper presented at the Annual Meeting of National Council on Measurement in Education, San Diego, CA.

Gierl, M.J., Wang, C., & Zhou, J. (2008). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6*(6), 1-49.

Gorin, J.S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25(4), 21-35.

Graf, E.A., Peters, S., Fife, J.H., Van Rijn, P.W., Arieli-Attali, M., & Marquez, E. (2019). *A Preliminary Validity Evaluation of a Learning Progression for the Concept of Function* (Report No: ETS RR–19-21). Wiley.

Hadenfeldt, J.C., Neumann, K., Bernholt, S., Liu, X., & Parchmann, I. (2016). Students' progression in understanding the matter concept. *Journal of Research in Science Teaching*, *53*(5), 683-708.

Harris, S. (2008). It's about time: Difficulties in developing time concepts. *Australian Primary Mathematics Classroom*, *13*(1), 28-31.

Iuculano, T., Padmanabhan, A., & Menon, V. (2018). Systems neuroscience of mathematical cognition and learning: Basic organization and neural sources of heterogeneity in typical and atypical development. In A. Henik & W. Fias (Eds.), *Heterogeneity of function in numerical cognition* (pp. 287-336). Academic Press.

Jin, H., Shin, H.J., Hokayem, H., Qureshi, F., & Jenkins, T. (2019). Secondary students' understanding of ecosystems: A learning progression approach. *International Journal of Science and Mathematics Education*, *17*(2), 217-235.

Kamii, C., & Russell, K.A. (2012). Elapsed time: Why is it so difficult to teach? *Journal for Research in Mathematics Education*, *43*(3), 296-315.

Kane, M.T., & Bejar, I. I. (2014). Cognitive frameworks for assessment, teaching, and learning: A validity perspective. *Psicología Educativa*, *20*(2), 117-123.

Keehner, M., Gorin, J.S., Feng, G., & Katz, I.R. (2017). Developing and validating cognitive models in assessment. In A. Rupp & J.P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (1st ed., pp. 75-101). Wiley Blackwell.

Ketterlin-Geller, L.R., & Yovanoff, P. (2009). Diagnostic assessments in mathematics to support instructional decision making. *Practical Assessment, Research & Evaluation, 14*(16), 1-11.

Lambert, K., Wortha, S.M., & Moeller, K. (2020). Time reading in middle and secondary school students: The influence of basic-numerical abilities. *The Journal of Genetic Psychology*, *181*(4), 255-277.

Langenfeld, T., Thomas, J., Zhu, R., & Morris, C.A. (2020). Integrating Multiple Sources of Validity Evidence for an Assessment-Based Cognitive Model. *Journal of Educational Measurement*, *57*(2), 159-184.

Leighton, J.P., & Gierl, M.J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3-16.

Leighton, J.P., Cui, Y., & Cor, M.K. (2009). Testing expert-based and student-based cognitive models: An application of the Attribute Hierarchy Method and Hierarchy Consistency Index. *Applied Measurement in Education*, *22*(3), 229-254.

Leighton, J.P., Gierl, M.J., & Hunka, S.M. (2004). The Attribute Hierarchy Method for cognitive assessment: A variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement*, *41*(3), 205-237.

Levin, I. (1989). *Principles underlying time measurement: The development of children's constraints on counting time*. In I. Levin and D. Zakay (Eds.), Advances in psychology (Vol. 59, pp. 145-183). Elsevier.

Li, H., & Suen, H.K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1-25.

Linacre, J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Morrison, K.M., & Embretson, S.E. (2014). Using cognitive complexity to measure the psychometric properties of mathematics assessment items. *Multivariate Behavioral Research*, *49*(3), 292-293.

Multon, K.D., & Coleman, J.S.M. (2010). *Coefficient alpha*. In N. Salkind (Ed), *Encyclopedia of research design* (pp. 159–162). Sage Publication.

Murata, A., & Kattubadi, S. (2012). Grade 3 students' mathematization through modeling: Situation models and solution models with mutli-digit subtraction problem solving. *The Journal of Mathematical Behavior*, *31*(1), 15-28.

National Council of Teachers of Mathematics [NCTM] (2000*). Principles and standards for school mathematics.* NCTM.

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, 64(4), 575-603.

Nichols, P.D., Kobrin, J.L., Lai, E., & Koepfler, J.D. (2017). *The role of theories of learning and cognition in assessment design and development*. In A.A. Rupp & J.P. Leighton (Eds.), The handbook of cognition and assessment: Frameworks, methodologies, and applications (1st ed., pp. 41–74). Wiley Blackwell.

Nuerk, H.C., Moeller, K., & Willmes, K. (2015). *Multi-digit number processing: Overview, conceptual clarifications, and language influences*. Oxford University Press.

Ojose, B. (2015). *Common misconceptions in mathematics: Strategies to correct them*. University Press of America.

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using SPSS program* (6th ed.). McGraw-Hill Education.

Pellegrino, J.W., & Chudowsky, N. (2003). Focus article: The foundations of assessment. *Measurement: Interdisciplinary Research and Perspectives*, *1*(2), 103-148.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* National Academy Press.

Pelton, T., Milford, T., & Pelton, L.F. (2018). Developing Mastery of Time Concepts by Integrating Lessons and Apps. In N. Calder, K. Larkin & N. Sinclair (Eds.), *Using Mobile Technologies in the Teaching and Learning of Mathematics* (pp. 153-166). Springer.

Polit, D.F., & Beck, C.T. (2006). The Content Validity Index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing and Health*, *29*(5), 489-497.

Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge in mathematics. In R. Cohen Kadosh & A. Dowker (Eds.), *Oxford handbook of numerical* *cognition* (pp. 1102-1118). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.014

Russell, M., & Masters, J. (2009, April 13-17). *Formative Diagnostic Assessment in Algebra and Geometry.* Paper presented at the Annual Meeting of the American Education Research Association, San Diego, CA.

Salkind, N. (2010) *Convenience sampling*. In N. Salkind (Ed.), Encyclopedia of research design (p. 254). Sage publications.

Schultz, M., Lawrie, G.A., Bailey, C.H., Bedford, S.B., Dargaville, T.R., O'Brien, G., ... & Wright, A.H. (2017). Evaluation of diagnostic tools that tertiary teachers can apply to profile their students' conceptions. *International Journal of Science Education*, *39*(5), 565-586.

Sia, C.J.L. (2017). *Development and validation of Cognitive Diagnostic Assessment (CDA) for primary mathematics learning of time* [Unpublished master's thesis]. Universiti Sains.

Sia, C.J.L., & Lim, C.S. (2018). *Cognitive diagnostic assessment: An alternative mode of assessment for learning*. In D.R. Thompson, M. Burton, A. Cusi, & D. Wright (Eds.), Classroom assessment in mathematics (pp. 123-137). Springer.

Sia, C.J.L., Lim, C.S., Chew, C.M., & Kor, L.K. (2019). Expert-based cognitive model and student-based cognitive model in the learning of "Time": Match or mismatch? *International Journal of Science and Mathematics Education*, *17*(6), 1–19.

Tan, P.L., Kor, L.K. & Lim, C.S. (2019). Abstracting common errors in the learning of time intervals via cognitive diagnostic assessment. *Creative Practices in Language Learning and Teaching (CPLT) Special Issue: Generating New Knowledge through Best Practices in Computing and Mathematical Sciences*, *7*(1), 3-10.

Tan, P.L., Lim, C.S., & Kor, L.K. (2017). Diagnosing primary pupils' learning of the concept of" after" in the topic" time" through knowledge states by using cognitive diagnostic assessment. *Malaysian Journal of Learning and Instruction*, *14*(2), 145-175.

Tatsuoka, K.K. (1986). Toward an integration of Item-Response Theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Lawrence Erlbaum Associates.

Tatsuoka, K.K. (1991). *Boolean algebra applied to determination of universal set of knowledge states* (Research Report No: RR-91-44-0NR). Educational Testing Service.

Tatsuoka, K.K. (2009). *Cognitive assessment: An introduction to the Rule Space Method.* Routledge.

Van de Walle, J.A., Karp, K.S., & Bay-Williams, J.M. (2012). *Elementary and secondary school mathematics: Teaching with developmental approach.* Pearson.

Van Steenbrugge, H., Valcke, M., & Desoete, A. (2010). Mathematics learning difficulties in primary education: Teachers' professional knowledge and the use of commercially available learning packages. *Educational Studies*, *36*(1), 59-71.

Wang, C., & Gierl, M.J. (2011). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, *48*(2), 165-187.

Wurpts, I.C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Frontiers in Psychology*, *5*(920), 1-15.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 197–217

https://doi.org/10.21449/ijate.1222410

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

The effects of learning oriented assessment on academic writing

Haticetul Kubra Er[1,*], Hossein Farhady[2]

[1]Erzurum Technical University, School of Foreign Languages, Erzurum, Türkiye
[2]Yeditepe University, Faculty of Education, Department of English Language Teaching, Istanbul, Türkiye

ARTICLE HISTORY

Received: Dec. 21, 2022
Revised: Feb. 22, 2023
Accepted: May 03, 2023

Keywords:
Learning Oriented Assessment,
Academic Writing Ability,
Argumentative Essay.

Abstract: The purpose of this study is to investigate the impact of learning-oriented assessment (LOA) on the academic writing ability of EFL students (N:40) during a 12-week in the semester of 2019-2020 academic year in the context of a higher education. Within a pretest-posttest intact group design, the experimental group received instruction following the principles of LOA, and the comparison group received routine procedures for academic writing. This is a quantitative experimental design. The test of normality, Kolmogorov-Smirnov, Mann-Whitney Tests and Wilcoxon Signed Ranks Test were administered in order to see the significance of the intervention the data for this study included scores of a pretest, several assessments, and a posttest at the beginning, during, and at the end of instruction, respectively. The prompt for both pre and post-tests required participants to write argumentative essays. To rate the writing tasks, we followed the rubrics developed by the testing office of the institution. The findings revealed that the experimental group outscored the comparison group indicating the effectiveness of LOA procedures in student learning. Further, the findings indicated that implementing LOA could have significant implications and applications for EFL writing education.

1. INTRODUCTION

According to Hyland (2014), academic writing provides opportunities for learners to create social negotiations and understand the process of constructing knowledge with the help of reasoning skills and critical thinking. In other words, academic writing serves as a communication booster that helps writers convey a message on a specific topic. The issues such as 'opportunities for learners to create social negotiations and understand the process of constructing knowledge' are also some of the principles of LOA.

As Carless (2014) states, LOA promotes higher-order thinking because the learners are active participants in generating, applying, and engaging with instructional criteria. These principles include students' and instructors' active engagement along with a focus on procedures that integrate assessment, learning, and teaching (William & Thompson, 2007; Stiggins, 2005). For instance, LOA fosters learners' self-directed learning skills in the context where active collaboration and cooperation take place along with using the feedback/feedforward process

*CONTACT: Haticetul Kubra Er ✉ kubra.er@gmail.com Erzurum Technical University, School of Foreign Languages, Erzurum, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

(Mok, 2013). Moreover, the LOA context requires that learning activities occur while learners are actively involved in the assessment process via thinking about their strategies for achieving learning objectives (Zeng, Huang, Yu & Chen, 2018).

The significance of this study lies in combining these two important aspects of ELT by investigating the effects of LOA on academic writing. To put the issue in an appropriate context, a brief description of LOA along with the framework used for this study seems necessary.

1.1. Statement of the Problem

Recently, several studies have attempted to deal with the concept of LOA and recently LOA has been a subject of various research studies due to several reasons. LOA promotes higher-order thinking and various approaches to learning since learners are active participants in generating, applying, and engaging with criteria (Carless, 2014). In language testing, formative assessment and LOA has gained popularity thanks to late and ongoing advancements (Carless, 2007). Mentioned advancements include students' and instructors' cognitive involvement along with a focus on procedures to promote assessment for learning (William & Thompson, 2007).

To illustrate, Hamp-Lyons (2017) examined the factors affecting learning orientation in assessment. According to research, LOA is as closely related to beliefs and principles of teaching as it is with principles in testing and assessment. Hamp-Lyons (2017) aimed to explore the possible ways that might encourage instructors and test developers to provide greater chances of learning for large-scale tests such as Cambridge Speaking Tests of CEFR B2 level. According to Hamp-Lyons (2017), LOA opportunities might be extremely useful in speaking tests for teacher trainers. Hence, Hamp-Lyons (2017) explored the effects of LOA on speaking assessment by showing the LOA processes. Furthermore, a similar study has been carried out by Green (2017) exploring the impact of using learning-oriented language test preparation materials for the speaking part of a General English proficiency test (Cambridge English). Besides, few studies have also explored the effects of both assessment and LOA in different ways. In his study, Ibrahim (2013) explored the support the idea of using LOA in an EFL setting and how to implement it along with challenges. However, the lack of a comprehensive view of the implementation of LOA in different contexts with different skills still exists. Carless (2014) also explored the LOA processes by observing classes. The research did not have a goal to explore the students' success, but it explored the process that learners and instructors were engaged in. As it can be understood from above, there is not much literature on ‘effects of learning-oriented assessment'. Also, earlier research on LOA has generally focused on the detailed description of the LOA process.

Writing skills and assessment have also been investigated in much of the previous studies. There are many studies about the Cognitive Process of Theory of Writing (Flower & Hayes, 1981), reading writing relations and its theoretical perspectives (Grabe, 2016), the genre in second language writing (SLW from now on) (Bawarshi &Reiff, 2010; Swales, 1990), fluency in writing (Hayes & Chenoweth, 2001), paraphrasing texts in SLW (Shi, 2012), contrastive rhetoric: cross-cultural aspects of SLW (Grabe & Kaplan, 1996), writing models and their effects on writing performances (Nicolas et al. ,2014), writing assessment (Grabe & Kaplan, 1996), written corrective feedback in writing accuracy (Han & Hyland, 2015), error correction in SLW (Beuningen, Jong & Kuiken, 2012). Lastly, regarding academic writing skills in a university context and formative assessment, research by Horstmanshof and Brownie (2013) investigated the effect of using a scaffold approach for formative assessment in academic writing skills. The researchers addressed the academic challenges of writing in the formative assessment such as timely feedback, and different abilities to improve academic writing skills in higher education. The authors also focused on student satisfaction, assessment, the role of feedback, and teaching/learning online.

Horstmanshof and Brownie's (2013) study fail to address significant components of assessment which are embedded in LOA since LOA assessment is a dynamic process as well as including the combination of not only formative assessment but summative as well. In addition to these LOA captures the centrality of learning within assessment whether summative or formative, the main goal of LOA is to promote active student learning (Barker, 2013). It should be also noted that LOA assessment has its root from both the features of both summative assessment and formative assessment. In other words, summative assessment evaluates what has happened before; that is to say, judgment and backward- looking, on the other hand, formative assessment guides what will happen next that is to say development and forward looking. Therefore, above mentioned features of both summative and formative assessment are within LOA that support learning. As can be seen above there are plenty of studies on academic writing. However, there is not any study conducted regarding the administration of LOA in higher education specifically for academic writing.

In brief, what is known about LOA is that it is largely based on studies that investigate the process of learning-oriented assessment rather than its effect on a specific skill. Previously published studies mostly describe the principles and process of LOA or LOA and technology relevance. To illustrate, Keppell, Au, Ma, and Chan (2007) investigated themes of group work, group projects, collaborative learning, and peer learning in LOA for technology-enhanced environments. As mentioned previously, similarly, Hamp-Lyons (2014) explored the effects of LOA on speaking assessment by showing the LOA processes. Furthermore, a similar study has been carried out by Green (2017) exploring the impact of using learning-oriented language test preparation materials for the speaking part of a General English proficiency test (Cambridge English)

1.2. Significance of the Study

Although there is a growing body of studies on academic writing, assessment in higher education, specifically LOA on academic writing, has received less attention. A number of authors have considered LOA in large-scale testing thus, the administration of LOA in an academic writing context in higher education is investigated to fill this gap in the literature.

To provide another example of why the current study is significant is that the implementation of LOA in different contexts has been investigated but they make no attempt to engage with higher education within academic writing specifically. To illustrate, the studies reported by Ashton and Salamoura (2012) illustrate the implementation of LOA in the primary and early secondary educational context. In addition to this Keppell (2006) asserts the significance of distance learners and distance learning with regard to flexible curriculum and learning at Hong Kong University for LOA implementation. Also, details of how teachers can use those strategies in their classrooms are shown as well. However, mentioned studies above did not consider the academic setting, especially for academic writing skills. Thus, it can be concluded that previous studies on LOA have dealt with large-scale testing and curriculum-based LOA. Therefore, the current research may contribute to the field with the implementation of LOA in higher education specifically for academic writing skills in the School of Foreign Languages.

Besides, regarding negative aspects of the traditional type of assessments Hamp-Lyons (2017) make a comparison between the former type of assessments and claims that learner-oriented assessment is against the traditional type of assessment which is about assessments that consist of judgment-focused tasks, learner excluded assessment and judgment-focused feedback. Thus, traditional assessment practices may have some weaknesses and limitations such as underestimating learners' capacities to evaluate their own work (Boud & Falchinov, 2006). It can then be argued that according to the studies mentioned above there are numerous challenges of assessment in higher education. Therefore, it would be useful, beneficial, and effective for describing principles and stages, which are linked to patterns of LOA and connect these specifically to academic writing.

In this manner, it could be concluded that LOA is of paramount importance and should be definitely used to support and promote effective learning specifically in higher education. As previously mentioned, the existing literature on writing skills and LOA is detailed but failed to address both academic writing skills and implementation of LOA in higher education. A more comprehensible study would then include several unresolved issues. Thus, the present study would hopefully be valuable and significant for a more efficient assessment of academic writing in higher education. In brief, the present research would make several contributions to the field of applied linguistics to fill the gap in terms of 'detailed description of LOA implementation process, it is being conducted in higher education, focusing specifically on academic writing skill'.

1.3. Definition and History of LOA

The roots of LOA reside in the sociocultural theory that (Westbury et al., 2000, p. 47) emphasizes the associations between the theory and practice within the framework of the philosophy of didactics. Under this philosophy, teachers need to focus on 'learners' learning and learning activities'. Besides, the Didactic paradigm focuses on the reflective processes for the assessment as the core element of teaching and learning processes (Vallberg, & Roth, 2014). This paradigm offers a framework for the reflective processes where the teacher needs to consider what, why, and how to assess questions in the context of instruction (Westbury et al., 2000, p.33).

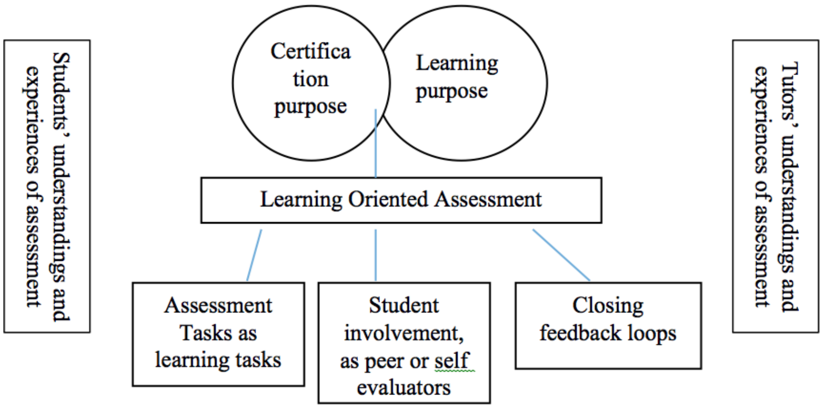
Many studies have investigated the interrelationship of teaching, learning, and assessment under different topics such as assessment of learning (AoL) (Hume & Coll 2009), assessment for learning (AfL) (Martinez & Lipson, 1989), and assessment during learning (Gibbons & Kankkonen, 2011) (AdL). Some scholars classify them all under the term formative assessment to contrast with summative assessment (Hume & Coll 2009; Stiggings, 2005). According to Mok (2013), the LOA framework seems to be comprehensive enough to function as an umbrella term and accommodate the multiple concepts developed in combination with the word "assessment". LOA comprises a blending of various assessments such as performance, alternative, authentic, and dynamic assessments. Within this conceptualization, Huang, Yu, and Chen (2018) expanded Carless' (2006, 2007) framework and offered a model of LOA that is more apt for a productive way of students' learning and an effective way for teachers' teaching.

Some principles emerged from the models that provided a roadmap for teachers in implementing LOA in a real context of the classroom. First, assessment tasks ought to promote learning among students. That is, the teacher serves as a curriculum designer to arrange a desirable assessment task that would promote learning. Second, students need to actively engage, with understanding, in the application of the criteria for self and peer assessment. It implies that the teacher as a test developer integrates AoL, AaL, and AfL and tries to help learners involve in the processes of teaching, learning, and assessment. Third, teachers need to provide timely feedback that is prompt and forward-looking for future learning (Zeng, et al., 2018). The teachers should receive training on providing feedback and feed-forward to support future learning.

In the context of the current study, the dependent variable (DV) is defined as the score of students' writing tasks prepared from the testing department of Erzurum Technical University School of Foreign Languages. Regarding operational definitions for instruction following LOA, there are two widely known frameworks: Learning-Oriented Assessment Framework (LOAF) proposed by Carless, (2007); Carless, Joughin, Liu and Associates, (2006) and 'Framework of

LOA' proposed by Turner and Purpura (2014). The LOAF has two main goals, which are evaluating learners' performance and the learning component. According to Carless (2009), the goal of LOA is to focus on the learning component of assessment in order to achieve it via both summative and formative assessment. Figure 1 is a graphic representation of LOA components.

Figure 1. *Framework for learning-oriented assessment (Carless, 2009).*


Students' understandings and experiences of assessment
Certifica tion purpose
Learning purpose
Learning Oriented Assessment
Assessment Tasks as learning tasks
Student involvement, as peer or self evaluators
Closing feedback loops
Tutors' understandings and experiences of assessment

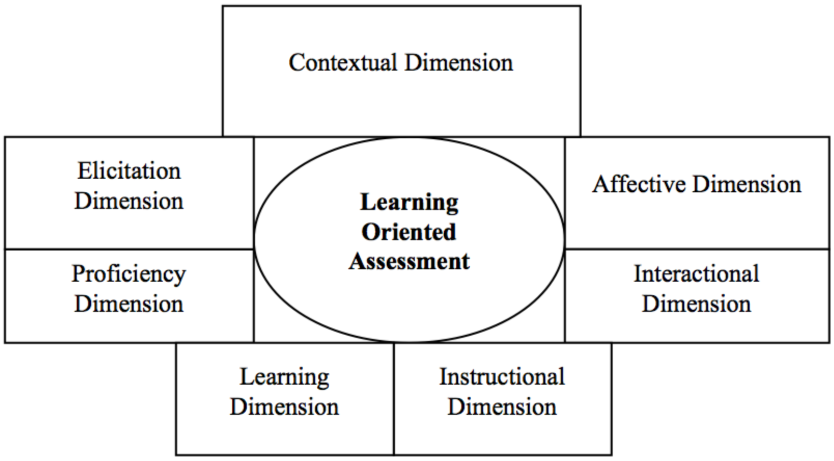
As Carless proposes, three strands of LOA are viewed as unified rather than composed of discrete elements that can be clearly seen from the above Figure 1; 1) Assessment tasks as learning tasks, 2) Student Involvement (Peer and Self-Assessment), 3) Feedback Loops or Feedback and Feed-forward.

As for the second framework called 'Framework of LOA' proposed by Turner and Purpura (2014), LOA can be described as an embedded assessment, focusing on the learner through seven interrelated dimensions. This framework also contributes to instructors with the goal of helping to facilitate the determination of best practices for teaching (Turner & Purpura, 2014). 'Framework for LOA' is adapted to serve the purpose of the study. Turner and Purpura's (2014) 'Framework of LOA' is administered for the current research due to its detailed descriptions of various dimensions. The LOA framework consists of seven dimensions that are the contextual, the elicitation, the proficiency, the learning, the instructional, the interactional, and the affective.

The Contextual Dimension of LOA has two phases, which are macro level and micro level. In the former one, curriculum, instruction, and assessment are affected by several factors such as socio-cultural norms and socio-political forces as well as classroom expectations. In the latter one, curriculum, instruction, and assessment are driven by personal attributes of teachers, teacher's choices, the creation of classroom culture. Thus, it can be concluded that the Contextual Dimension indicates teachers' characteristics (assessment literacy) that has an effect on learning and assessment in a class context. The Elicitation Dimension of LOA involves the situations in which language is elicited in various methods. In the form of a feedback for potential intervention action, students' performance is noticed, argued, commented on, and responded to. The Proficiency Dimension of LOA is utilized to identify 'what to assess? How to follow the performance? and what to focus on regarding feedback? The Learning Dimension of LOA consists of a perception of how students deal with knowledge and finally learn. Furthermore, it is crucial to know how instruction and assessment are conceptualized and administered. The role of feedback and self-regulation (responsible for their own learning) are also considered as critical features of the learning dimension of LOA.

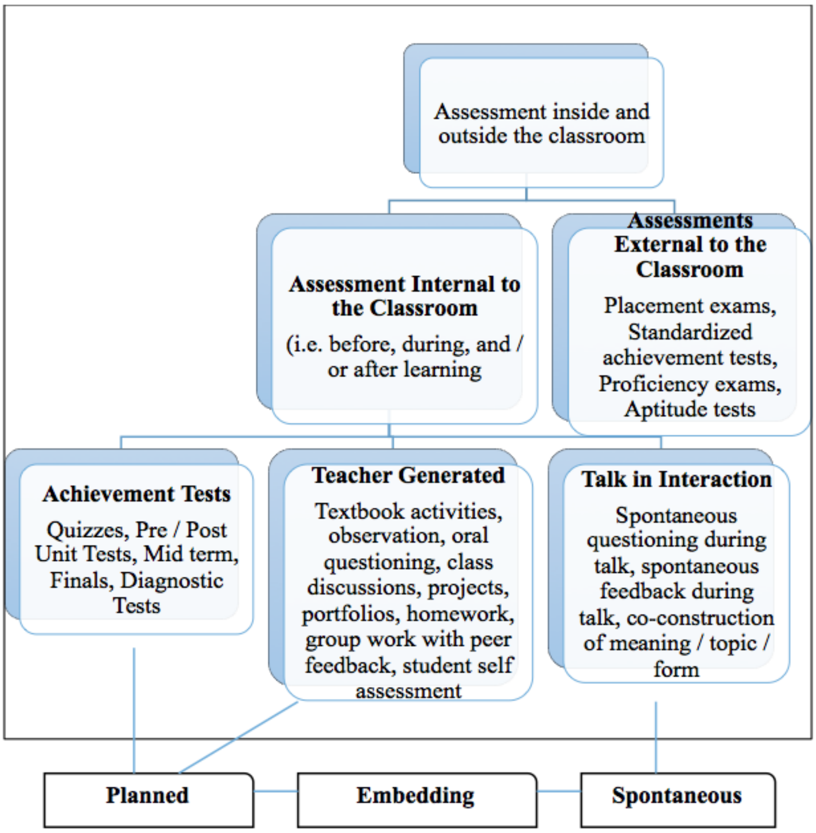
The Instructional Dimension of LOA is related to; Teacher's Content and Content Knowledge. Thus, it is important to consider the following question 'How much do instructors' pedagogical content knowledge influence the understanding of LOAs and choices regarding the following learning processes?'' The Interactional Dimension of LOA encapsulates the organization of LOA in an interactive manner. Lastly, The Affective Dimension of LOA defines learner's feelings and motivation level regarding learner's engagements in the assessment process. In other words, it is closely associated with the characteristics such as emotions, beliefs, personality, attitude, and motivation. To sum up, seven dimensions of LOA are illustrated in the below Figure 2.

Figure 2. *Dimensions of LOA framework (Reprinted from Turner & Purpura, 2014).*



The latter Figure 3, demonstrates the detailed implementation of 'Framework for LOA' proposed by Turner and Purpura (2014).

Figure 3. *Framework for learning-oriented assessment (Turner & Purpura, 2014).*


Assessment inside and outside the classroom
Assessment Internal to the Classroom
(i.e. before, during, and / or after learning
Assessments External to the Classroom
Placement exams, Standardized achievement tests, Proficiency exams, Aptitude tests
Achievement Tests
Quizzes, Pre / Post Unit Tests, Mid term, Finals, Diagnostic Tests
Teacher Generated
Textbook activities, observation, oral questioning, class discussions, projects, portfolios, homework, group work with peer feedback, student self assessment
Talk in Interaction
Spontaneous questioning during talk, spontaneous feedback during talk, co-construction of meaning / topic / form
Planned
Embedding
Spontaneous

The independent variables included External Assessments, Internal Assessments; planned assessments (achievement tests, teacher-generated), and spontaneous assessments (talk in interaction) in the context of the current research. Here are the LOA components for independent variables: Achievement Tests: pre-test, post-test, timed writing quizzes, self-regulated tasks: reflective diary, same day feedback, weekly personal response, portfolio, participation in weekly tutorials, patchwork texts, peer and group tasks: team projects (group critique and group assessment), mini projects (peer critique and peer assessment), in-class feedback, and computer-mediated collaborative writing.

To shed some light on the effectiveness of LOA in practice, the current research addressed the research question mentioned below:

Is there a significant difference between the pretest and posttest scores of the experimental and control groups?

2. METHOD

The method of the current study is a quasi-experimental design and it is a quantitative study. A quasi-experiment is a research design that aims to determine a cause-and-effect relationship, but unlike a true experiment, the groups involved are not randomly assigned. Quasi-experiments are usually carried out in real-world settings where it's hard or not feasible to randomly assign subjects. They're frequently used to assess the efficacy of a treatment such as a psychotherapeutic approach or educational intervention (Cook & Campell, 1979)

2.1. Participants

The participants were 40 students from the School of Foreign Languages in Turkey. They were in two intact groups in classes from the B2 level of language ability on the CEFR scale. Their age ranged between 18 and 24. The participants formed two randomly assigned groups of experimental and comparison groups (N=20 for each). Students' levels were determined according to the English Proficiency Exam of Erzurum Technical University, which is the exit/exemption test of the English Preparatory Program of School of Foreign Languages at a state university.

2.2. Instruments

The design of this study required several instruments for the study groups. Two 'argumentative essays' for both groups and several LOA-based assessments for the experimental group formed the basis for the data. Argumentative essay 1 served as the pretest and argumentative essay 2 as the post-test for both groups. The essays went through a double-check procedure according to rubrics and grading criteria developed by the institution. One of the raters was an EFL teacher from the institute and the other a teacher from Marmara University's School of Foreign Languages (Appendix A, B, C). Grading LOA tasks also followed the rubrics provided for each task and shared with the students.

2.3. Procedure

The present study was adopted a mixed-method approach with a combination of pre-test and post-test and quantitative design. Erzurum Technical University School of Foreign Languages students in upper-intermediate and advanced levels were asked to write compare-contrast essays, cause and effect essays, an argumentative essay, and an argumentative research paper. As for the assessment tool (pre / post-test) argumentative writing has been selected as a main writing performance in this study. There are several reasons why argumentative writing is selected as a main writing performance in this study. According to the scholars (Manzi, Flotts & Preiss, 2012; Paek & Kang 2017), argumentative writing can be considered as one of the most difficult and demanding types of writing when compared to other types of academic writing due to the following reasons: consequences of linking high cognitive skills along with the ability to use the language, sharing ideas on different contrasting views, writers' own point of view about the argument, and a well-designed critical angle (Krause & Brian, 1999). The above features make argumentative essay challenging for the author therefore, there are a couple of things that need to be considered; knowing how to interact and communicate with the audience, becoming aware of communicative nature of writing which is related to certain manner of considering and addressing views on a topic for or against and effort to change them.

The last very fundamental reason why argumentative writing is at the cornerstone of academia is that in the writing elements of their exam the globally accepted English proficiency tests: IELTS (International English Language Testing System) and TOEFL (Test of English as a Foreign Language) both use and administer argumentative essays. This notion demonstrates that ability to present, argue, justify or refute opinions are measurement criteria of a student's English writing proficiency.

Taking the above information into consideration, 'comparison group' is assessed through 'instruction following 'routine procedures for academic writing' which are; achievement tests: pre-test, post-test, argumentative essays, timed writing quizzes, and as for the 'experimental group' instruction following 'LOA Procedures'; achievement tests: pre-test, post-test, argumentative essays, timed writing quizzes, self-regulated tasks: reflective diary and reflective journal, same day feedback, weekly personal response, portfolio, participation in weekly tutorials, patchwork texts, peer and group tasks: team projects (group critique and group assessment), mini projects (peer critique and peer assessment), in-class feedback, and computer-mediated collaborative writing was administered. Here are the detailed descriptions of LOA components and procedures which are used as an intervention for the experiment group. Table 1 presents the two main categories of 'self-regulated and collaborative tasks' and 'assigned and assessed tasks'.

Table 1. *Self-regulated and collaborative tasks.*

	Self Regulated Tasks
Reflective Diary (RD)	Type of writing in several genres such as expectations from an academic writing class, impression, judgments, attitude regarding academic writing practices, procedures to help the efficiency of the course.
Same Day Feedback (SDF)	Questions formed by the teacher through an online platform on the same day they have in the class, also students are asked to offer input and critiques of each other's responses.
Weekly Personal Response (WPR)	Students prepare questions each week and upload them to the Blackboard System and answer every question posted by other students, combine them, and send them to the teacher.
Portfolio Assessment (PA)	Collecting students' work throughout the course to reflect on their effort, progress, and achievements. About essay drafts, paraphrasing, summarizing, editing, and citation.
Participation in Weekly Tutorials (PWT)	Tutorials are 15-minute, one-on-one workshops regarding academic writing compare / contrast essays, cause/effect essays, argumentative essays, and the argumentative research paper in which students receive assistance and feedback.
Patchwork Text Assessment	Learners were asked to fulfill regular short writing tasks; patches including various themes and genres throughout the module. The teacher constantly checked the writing and gives formative feedback to help students produce a reflective, 'stitching together of the patches. PT provides students with continuous productivity, and collective assessment along with learning via 'metacognitive self-reflection.

Table 1. *Continues.*

	Peer and Group Tasks
Team Project (Group Critique and Group Assessment) (TP)	Groups are required to write reaction papers to selected articles by using the academic writing skills they have learned throughout the semester including the joint writing abilities of learners. Assessing both individual efforts, contributions to group work, and our level of involvement in performing a group task was observed through ‘assessment criteria’.
Mini Projects (Peer Critique and Peer Assessment) (MP)	Students assess and evaluate their classmate’s work and have their work assessed by peers. Also, peer involvement personalized the learning experience, potentially motivating continued learning processes.
In-class Feedback (ICF)	Students were required to criticize, give feedback, edit and reflect upon each other’s writing tasks in in-class activities.
Computer-Mediated Collaborative Writing (CMCW)	This writing task is implemented in a web platform where learners discuss the writing tasks, co-build and revise paragraphs and collectively creates a solitary online text via joint endeavors with the help of technological tools like Google Docs and Blackboard (Online Education Platform).

Here is the detailed table for weekly ‘LOA procedures’ and ‘routine procedures’ for both the experimental group and comparison group. Thus, Table 2 demonstrates the weekly instructions for the experimental and comparison groups.

Table 2. *Weekly LOA procedures and routine procedures.*

	LOA Procedures	Experimental Group	Routine Procedures Comparison Group
	Daily Tasks:	*Weekly Tasks:*	
WEEK 1: Researched Essay	Same Day Feedback 1	PRE-TEST / Argumentative Essay Weekly Personal Response 1 Participation in Weekly Tutorials 1 Mini Project 1	PRE-TEST / Argumentative Essay In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book
WEEK 2: Comparison-Contrast Essay	Reflective Diary 1 Same Day Feedback 2 In-class Feedback 1	Patchwork Text 1 Computer-Mediated Collaborative Writing 1 Portfolio 1	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book
WEEK 3: Comparison-Contrast Essay	Same Day Feedback 4	Weekly Personal Response 2 Participation in Weekly Tutorials 2 Mini Projects 2	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book
WEEK 4: Cause / Effect Essay	Same Day Feedback 4 In-class Feedback 2	Patchwork Text 2 Computer-Mediated Collaborative Writing 2 Portfolio 2 *Comparison-Contrast Essay Mid-Term*	In-class instruction, feedback Activities / Tasks fromEffective Academic Writing Book *Comparison-Contrast Essay Mid-Term*
WEEK 5: Cause / Effect Essay	Reflective Diary 2 Same Day Feedback 5	Weekly Personal Response 3 Participation in Weekly Tutorials 3 Mini Project 3 *Timed Writing Quiz*	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book *Timed Writing Quiz*
WEEK 6: Argumentative Essay	Same Day Feedback 6 In-class Feedback 3	Patchwork Text 3 Computer-Mediated Collaborative Writing 3 Portfolio 3	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book

Table 2. *Continues.*

WEEK 7: Argumentative Essay	Same Day Feedback 7	Weekly Personal Response 4 Participation in Weekly Tutorials 4 Mini Project 4	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book
WEEK 8: Argumentative Essay	Reflective Diary 3 Same Day Feedback 8 In-class Feedback 4	Patchwork Text 4 Computer-Mediated Collaborative Writing 4 Portfolio 4 *Timed Writing Quiz*	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book *Timed Writing Quiz*
WEEK 9: Classification Essay	Same Day Feedback 9	Weekly Personal Response 5 Participation in Weekly Tutorials 5 Mini Project 5	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book
WEEK 10: Classification Essay	Same Day Feedback 10 In-class Feedback 5	Patchwork Text 5 Computer-Mediated Collaborative Writing 5 Portfolio 5	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book
WEEK 11: Reaction Essay	Same Day Feedback 11	Weekly Personal Response 6 Participation in Weekly Tutorials 6 Mini Project 6	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book
WEEK 12: Reaction Essay	Same Day Feedback 12 In-class Feedback 6	Patchwork Text 6 Computer-Mediated Collaborative Writing 6 Portfolio 6 Team Project *POSTTEST / Argumentative Essay Writing Final Exam*	In-class instruction, feedback Activities / Tasks from Effective Academic Writing Book *POSTTEST / Argumentative Essay Writing Final Exam*

As can be seen from Table 2 above, experimental and comparison groups had received different intervention. The experimental group proceeded with below mentioned 'LOA procedures'; achievement tests: pre-test, post-test, comparison-contrast essay, cause-effect essays and argumentative essays, timed writing quizzes, self-regulated tasks: reflective diary and reflective journal, same day feedback, weekly personal response, portfolio, participation in weekly tutorials, patchwork texts, peer and group tasks: team projects (group critique and group assessment), mini projects (peer critique and peer assessment), in-class feedback, and computer-mediated collaborative writing. However, the comparison group proceeded with pre-test, post-test, comparison-contrast essay, cause-effect essays, and argumentative essays and timed writing quizzes. In addition to these, the comparison group had received in class instruction and feedback.

The data for this research came from four sources; first, the rating scores from the pretest that included scores from the argumentative essay 1 were performed by both groups; second, the scores of regular assessments were assigned during the course for both groups, third, the scores from LOA tasks for the experimental group, fourth the end of the process, the scores of posttests on argumentative essay 2 for both groups. Informed consent form is taken by the participants along with the ethical from institution.

Here is the explanation of the procedure for data collection, the intervention lasted for 12 weeks. Pre-test (argumentative essay) for both comparison and experimental groups was administered in the first week. The control group went through a routine process of writing instruction and regular feedback provided by teacher. The experimental group, however, received additional assessments following LOA procedures; Self-Regulated and Collaborative Tasks mentioned above. At the end of the semester (the 12th week) post-test (Argumentative Essay) for both Comparison and Experimental groups was administered.

2.4. Data Analysis

To answer research question quantitative data was analyzed through SPSS. Test of normality, Kolmogorov-Smirnov, Mann-Whitney Tests and Wilcoxon Signed Ranks Test were used to estimate the significance of instruction following 'Routine Procedures' and 'LOA procedures' effect on academic writing.

3. RESULTS

To answer the first question of whether LOA has any effect on students' academic writing ability, test of normality, Kolmogorov-Smirnov, Mann-Whitney Tests and Wilcoxon Signed Ranks Test were performed. The test of normality, Kolmogorov-Smirnov statistics are presented in Table 3. The test of normality, Kolmogorov-Smirnov test results between the study groups presented here.

Table 3. *The tests of normality.*

	Tests of Normality					
	Kolmogorov-Smirnov[a]			Shapiro Wilk		
	Statistic	*df*	Sig.	Statistic	*df*	Sig.
PRE-TEST	.130	41	.078	.924	41	.009
POST-TEST	.121	41	.137	.937	41	.025

The table above shows that the two tests of normality revealed complicated results and did not agree with each other. Kolmogorov-Smirnov test supported normal distribution (p>0.05), while Shapiro-Wilk test indicated not normal distribution of the data (p<0.05). To avoid statistical weaknesses and risks, data will be counted as with non-normal distribution and non-parametric tests will be used for the data analysis. Non-parametric tests can be used for the analysis of both normal and not normal data distribution. Table 4 and Table 5 show results of the Mann-Whitney Test in order to see the significance difference between experimental and comparison groups on the pretest and posttest.

Table 4. *Mann-Whitney test results: Descriptive statistics.*

	Ranks /Descriptive statistics			
	Intervention	N	Mean Rank	Sum of Ranks
PRE-TEST	EXP	20	22.75	455.00
	COMP	21	19.33	406.00
	Total	41		
POST-TEST	EXP	20	29.70	594.00
	COMP	21	12.71	267.00
	Total	41		

Table 5. *Mann-Whitney test results: Test statistics.*

	Test Statistics[a]	
	PRE-TEST	POST-TEST
Mann-Whitney U	175.000	36.000
Wilcoxon W	406.000	267.000
Z	-.914	-4.546
Asymp. Sig. (2-tailed)	.361	.000

a. Grouping Variable: Intervention.

The results from Mann-Whitney Test in the Table 4 and Table 5 above indicate that there is no significant difference (*p*>0.05) between experimental and comparison group for the pretests, which supports the homogeneity of the sample. However, a significant difference (*p*<0.05) is found between experimental and comparison group for the posttest scores. Descriptive statistics results in Table 4 clarifies that experimental group has outperformed the comparison group with a significant difference in the posttest scores. That is why, it can be assumed that the LOA intervention significantly improved the writing scores of the experimental group. In other words, students' writing scores were higher in the post-test therefore, achievement was higher in the experimental group. Table 6 and Table 7 show results of the Wilcoxon Signed Ranks Test and Descriptive Statistics as well in order to see the significance difference between experimental and comparison groups on the pretest and posttest.

Table 6. *Wilcoxon Signed Ranks Signed Test: Ranks / Descriptive statistics.*

Ranks /Descriptive statistics				
		N	Mean Rank	Sum of Ranks
POSTEXP - PREEXP	Negative Ranks	0[a]	.00	.00
	Positive Ranks	20[b]	10.50	210.00
	Ties	0[c]		
	Total	20		
POSTCOMP - PRECOMP	Negative Ranks	0[d]	.00	.00
	Positive Ranks	21[e]	11.00	231.00
	Ties	0[f]		
	Total	21		

a. POSTEXP < PREEXP, b. POSTEXP > PREEXP, c. POSTEXP = PREEXP, d. POSTCOMP < PRECOMP, e. POSTCOMP > PRECOMP, f. POSTCOMP = PRECOMP.

Table 7. *Wilcoxon Signed Ranks Signed Test: Test statistics.*

Test Statistics[a]		
	POSTEXP - PREEXP	POSTCOMP - PRECOMP
Z	-3.926[b]	-4.044[b]
Asymp. Sig. (2-tailed)	.000	.000
Exact Sig. (2-tailed)	.000	.000

a. Wilcoxon Signed Ranks Test, b. Based on negative ranks.

According to Table 6 and Table 7 the findings of the Wilcoxon Signed Ranks Test for both groups of pre and post tests revealed that there is a significant difference between pre and post test results for both of the groups (*p*<0.05). Descriptive statistics indicated that post scores were significantly higher that pre-test scores for both of the groups. Therefore, the intervention had a positive effect on the experimental group, however, comparison group also reached significantly high progress without the intervention.

4. DISCUSSION and CONCLUSION

The findings of the current research are in support of the previous studies about computer-mediated collaborative writing, patchwork text assessment, portfolio assessment, self, peer, and group assessment and feedback as well as the principles of Turner and Purpura's (2014) LOA framework. Previous findings support the positive effect of interaction and computer-mediated writing on argumentative writing. These findings also indicate that regulation activities in collaborative writing foster learners' involvement, self-confidence, and responsibility (Chao &

Lo, 2011; Cho, 2017; Wang, 2019,). On other LOA tasks, Further, Wu, Petit and Chen (2015) studied the effect of online interactivity and discussion between EFL writing learners on a computer-mediated platform. These findings showed that learners benefited from online peer feedback specifically in essay writing assignments.

On another LOA task, the findings of the study corresponded with the earlier studies by Dalrymple and Smith (2008) who mention the positive effects of patchwork text regarding student interaction and participation. The findings of the current study were also in line with earlier studies about the effectiveness of portfolio assessment in writing (Eridafithri, 2015; Farahian & Avarzamani, 2018; Kathpalia & Heah, 2008; Lam, 2019; Romova & Andrew, 2011) who all reported the effectiveness of portfolios on academic writing specifically concerning peer collaboration, reflectivity, and feedback loops. Hence, the significance of the results of the study comes from combining these two significant aspects of ELT for understanding the effects of LOA on academic writing since LOA comprises research on academic writing as well. In this manner, it could be concluded that particularly in higher education, LOA is of utmost importance and should unquestionably be used to support and promote good learning. The existing studies on writing skills and LOA is detailed but failed to address both academic writing skills and implementation of LOA in higher education.

In addition to these, having mentioned the facets of computer-mediated collaborative writing facets above in the literature review, it can be implied that these facets are closely associated with stages of the LOA learning interaction model since WPR and TP were related to computer-mediated collaborative writing. With regard to stages of LOA learning interaction via technology, Jones and Seville (2016) proposed that delivery and mediation of assessment and learning tasks, capturing and recording data, tracking progress, individualization of learner's experience, enabling new forms of learning interactions and improving our understanding of learning are among the most important stages of LOA learning interaction via computer. As can be observed these correspond well with the ‘learning dimension’ since learners collaboratively engage and interact with each other. To provide an example, as mentioned before Storch (2019) defined collaborative writing in its broadest sense, collaborative writing is defined as the process of writing a text with multiple authors or writers (p. 2). It can then be said that collaborative writing consists of several themes like interaction among learners and editing phases of the writing process. As cited in Alghasab and Handley (2017), these concepts are again closely related to the Turner and Purpura’s (2014) LOA framework of ‘Learning Dimension’ since learners focus on self-regulation and ‘how they learn’.

Regarding the ‘Affective Dimension’ of Turner and Purpura’s (2014) LOA framework, it is again in consistent with Computer Mediated Writing since it promotes motivation as well. Thus, it is again related to SDF, RD, WPR and TP tasks of LOA. As the studies of Elola and Oskoz (2010); Storch, (2005) and Chen (2016) suggest computer-mediated collaborative writing is beneficial in promoting the acquisition of different language skills along with the motivation for learning.

Similarly, patchwork assessment fosters and promotes the concepts of learners’ self-reflection, peer feedback, self-regulation skills. As Wilson and Trevelyan (2012) claimed alternative components of patchwork text assessment encapsulate the flow of patches, resubmission of prior patches, summative feedback, collaboration, self-reflection. Therefore, these components of patchwork text assessment; collaboration, feedback, self-reflection are closely associated with the principles of Turner and Purpura’s LOA framework of ‘Learning Dimension’ since they are related to ‘how learners process learning and become responsible for their own learning’. Similarly, since patchwork text promotes student’s interaction and participation (Dalrymple & Smith, 2008) it aligns with the principles of Turner and Purpura’s LOA framework of ‘Learning Dimension’ as well. SDF, RD, WPR and TP tasks of LOA comprise patchwork assessment partly therefore, patchwork text also aligns with the principles of Turner and Purpura's LOA framework of 'elicitation dimension' which deals with the situations in which language is obtained and acquired. in this dimension learners' actions and progresses are observed and examined, hence 'elicitation dimension' is related to patchwork text principles of how students learn and observe their learning phase and pace since students have an opportunity to reflect, react and discuss in SDF, RD, WPR and TP tasks of LOA thanks to computer-mediated writing feedback (Winter, 2003).

Considering the principles and components of portfolio assessment along with SDF, RD, WPR, and TP tasks of LOA, portfolio assessment is associated and in consistency with the principles of Turner and Purpura's (2014) LOA framework of the 'learning dimension'. In portfolio assessment students become aware of their own learning and progress. In other words, according to Lam (2019) self-monitoring, self-reflection and self-assessment are the core element of portfolio assessment thus, the above-mentioned features of portfolio assessment highly coincide with the learner-centered teaching model of LOA as well as the 'learning dimension' of Turner and Purpura's (2014) LOA framework. Similarly, learners become independent and responsible learners with the help of portfolio (Arslan, 2014; Bader, Iversen & Varga, 2019; Eridafithri, 2015;) therefore, this is closely related to the principles of Turner and Purpura's LOA framework of 'Learning Dimension' since they are related to 'how learners process learning and becoming responsible for their own learning' in the process of SDF, RD, WPR and TP tasks of LOA respectively.

In this sense studies and literature reviewed on self / peer assessment and feedback are associated and in consistency with the principles of Turner and Purpura's (2014) LOA framework of the 'learning and Affective Dimension'.

The Affective Dimension of LOA defines learner's socio-mental inclinations with respect to how students experience and participate in the assessment process. In other words, it is closely associated with the characteristics like emotions, beliefs, personality, attitude, and motivation. therefore, self / peer assessment and feedback facilitate the affective dimension of learning by providing chances for learners to express their expectations from an academic writing class, impression, judgments, attitude regarding academic writing practices, procedures to help the efficiency especially for SDF, RD, WPR since learners are given chance to express themselves thanks to these LOA tasks. As Turner and Purpura (2013) claimed affective dimension is related to learner's socio-psychological aspects which is the learner's engagement in the process of assessment. Similarly, according to Katstra et al. (1987) study learners who receive peer feedback have more positive feelings and attitudes towards writing skills. In this respect, this is closely related to the 'affective dimension' of Turner and Purpura's (2014) LOA framework. Also, as the findings of Gielen et al. (2010) and Strijbos et al.'s (2010), study indicated that there were positive impacts of peer feedback on learners' learning outcomes. Lastly, regarding the principles of Turner and Purpura's (2014) LOA framework of the 'learning dimension', this is closely related to the Self / Peer Assessment and Feedback since these components are already embedded in the 'Learning Dimension'.

In sum, above mentioned studies regarding LOA, computer-mediated collaborative writing, patchwork text assessment, portfolio assessment, self, peer assessment and feedback have been carried out separately in the field. The current study on the effects of LOA on student's academic writing ability is carried out in order to fill the gap of cumulative different types of writing tasks as well as assessment. All in all, the findings of the current study touched upon several unresolved issues and for a more effective evaluation of academic writing in higher education, it would ideally be useful and significant.

4.1. Implication and Application

A significant implication of this study is the effectiveness of instruction within the LOA framework. Learners' high achievement on different LOA tasks implies that a move may start among EFL instructors, academic writing teachers, and course designers to consider adapting LOA activities for their contexts. Since the application of the LOA framework in the EFL context academic writing is gaining importance, popularity, and recognition, teacher education programs may want to include information that would prepare teachers to accommodate the changes (Mok, 2013).

4.2. Suggestions for Further Study

LOA is a newly emerging field. It can make significant changes in the education of the students. Therefore, it is applicable to multiple possibilities for research in multiple areas in which benefits of the LOA framework are investigated for other language skills in different contexts. Besides, its effect on other university courses, with other age range learners, at various levels of language ability are all fascinating areas of research. Another significant point worth mentioning would be related to the implementation of LOA in different online platforms. Since online and distance education have become popular, gained importance, and became part of our lives nowadays, a study of LOA administration and its effect on various skills via an online platform will serve as a base for future studies as well.

It should be also emphasized that a more comprehensible study of teacher's pedagogical practices on LOA referring both to pre-service and in-service teacher education practices would also be a thorough investigation of LOA in different aspects of English language teaching specifically in Turkish educational settings, meaning an EFL context. In addition to the above-mentioned aspects, a detailed and depth analysis of teacher education regarding implications of LOA; specifically, in terms of improving learners' assessment skills, differentiated instruction and fostering the feedback process can be explored as well as a further study. For example, Keppell (2006) and Carless's (2006) study that explored the principles of LOA in a teacher education context can be conducted in the school of foreign languages of Turkish universities with an emphasis on different feedback forms, peer learning, web-based platforms, project-based, and task-based learning and so on. To sum up, therefore it may be said that innovative learning platforms and e-assessment would be a comprehensible and pioneering area to be examined along with the LOA literacy of language teachers in the field of ELT.

4.3. Limitation of the Study

A number of limitations can emerge from the current study. To begin with, an argumentative essay was selected as the main writing performance due to the School of Foreign Languages Testing Policy; however, an expository essay would give more accurate information regarding student's academic writing skills' performance in this context since expository essay type comprises argumentative, cause-effect and compare-contrast essay types respectively.

Another limitation could be related to the number of participants. There were 40 students (participants) from the School of Foreign Languages. Future studies should include more participants to make further generalizations of the present findings reported in this dissertation.

Acknowledgments

The current research is produced from a Ph.D. dissertation.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Yeditepe University Humanities and Social Sciences Ethics Committee, 31.08.2020 - 1212020.

Authorship Contribution Statement

Haticetul Kubra Er: Investigation, Resources, Methodology Visualization, Software, Formal Analysis, and Writing-original draft. **Hossein Farhady**: Supervision, and Validation.

Orcid

Haticetul Kubra Er https://orcid.org/0000-0001-5995-6224
Hossein Farhady https://orcid.org/0000-0002-2019-6157

REFERENCES

Alghasab, M., & Handley, Z. (2017). Capturing (non-) collaboration in wiki-mediated collaborative writing activities: The need to examine discussion posts and editing acts in tandem. *Computer Assisted Language Learning*, *30*(7), 664-691. https://doi.org/10.1080/09588221.2017.1341928

Arslan, R. (2014). Integrating feedback into prospective English language teachers' writing process via blogs and portfolios. *The Turkish Online Journal of Educational Technology, 13*(1), 131-150.

Ashton, K., Salamoura, A., & Diaz, E. (2012). *The BEDA impact project: A preliminary investigation of a bilingual programme in Spain. Research notes, 50*, 34-41.

Bader, M., Burner, T., Iversen S., & Varga Z. (2019). Student perspectives on formative feedback as part of writing portfolios. *Assessment & Evaluation in Higher Education, 44*(7), 1017-1028. https://doi.org/10.1080/02602938.2018.1564811

Barker, (2013). Learning oriented assessment. Retrieved from: https://www.cambridgeenglish.org/research-and-validation/fitness-for-purpose/loa/

Bawarshi, A., & Reiff, M.J. (2010). *Genre: An introduction to the history, theory research, and pedagogy*. Parlor press/WAC Clearinghouse.

Beuningen, V., De Jong, N.H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning*, *62*, 1-41. https://doi.org/10.1111/j.1467-9922.2011.00674.x

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

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219–233.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International*, *44*(1), 57-66. https://doi.org/10.1080/14703290601081332

Carless, D. (2009). *Learning Oriented Assessment: Principles, practice and a project*. In Meyer, L.H, Davidson, H, Anderson, R, Fletcher, PM, Johnston & Rees, M. (Eds), Tertiary Assessment & Higher Education Student Outcomes: Policy, Practice & Research. Wellington.

Carless, D. (2014). Exploring learning-oriented assessment processes. *Higher Education*, *69*, pp.2-5. https://doi.org/10.1007/s10734-014-9816-z

Carless, D., Joughin, G., & Liu, N. (2006). *How assessment supports learning: Learning oriented assessment in action*. Hong Kong University Press.

Chao, Y.C.J., & Lo, H.C. (2011). Students' Perceptions of Wiki-Based Collaborative Writing for Learners of English as a Foreign Language. *Interactive Learning Environments*, *19*(4), 395-411, https://www.learntechlib.org/p/167206/.

Chen, T. (2016). Technology-supported peer feedback in ESL/EFL writing classes: A research synthesis. *Computer Assisted Language Learning*, *29*(2), 365-397, https://doi.org/10.1080/09588221.2014.960942

Cho, H. (2017). Synchronous web-based collaborative writing: Factors mediating interaction among second-language writers, University of Toronto. *Journal of Second Language Writing. 36*, 37-51, https://doi.org/10.1016/j.jslw.2017.05.013

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design & analysis issues in field settings*. Houghton Mifflin.

Dalrymple, R., & Smith, P. (2008). Patchwork Text: Enabling discursive writing and reflective practice on a foundation module in work-based learning. *Innovations in Education and Teaching International*, *45*(1), 47-54. https://doi.org/10.1080/14703290701757443

Elola, I., & Oskoz, A. (2010). Collaborative writing: Fostering foreign language and writing conventions development. *Language Learning & Technology*, *14*(3), 51–71, http://llt.msu.edu/vol14num3/elolaoskoz.pdf.

Eridafithri, E. (2015). The Application of portfolios to assess progress in writing of EFL students at secondary schools in banda aceh. *Studies in English Language and Education, 2*, 1-16. https://doi.org/10.24815/siele.v2i1.2231

Farahian M., & Avarzamani F. (2018). The impact of portfolio on EFL learners' metacognition and writing performance. *Cogent Education*, *5*(1). https://doi.org/10.1080/2331186X.2018.1450918

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007

Grabe, W., & Kaplan, R.B. (1996). *Theory and Practice of Writing: An Applied Linguistic Perspective*. Longman, New York, p. 91, 93, 114, 115, 202, 245.

Grabe, W., & Zhang C. (2016). Reading-writing relationships in first and second language academic literacy development. *Language Teaching*, *49*, 339-355 https://doi.org/10.1017/S0261444816000082

Green, A. (2017). Learning-oriented language test preparation materials: A contradiction in terms?. *Papers in Language Testing and Assessment*, *6*(1), 112-132. http://hdl.handle.net/10547/622430

Gibbons, S., & Kankkonen, B. (2011). Assessment as learning in physical education: Making assessment meaningful for secondary school students, *Physical and Health Education Journa*l, *76*(4), 6–12.

Hamp-Lyons, L. (2017). Language assessment literacy for language learning-oriented assessment. *Papers in Language Testing and Assessment*, *6*(1), 88-110. http://hdl.handle.net/10547/622445

Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing*, *30*, 31–44. https://doi.org/10.1016/j.jslw.2015.08.002

Hayes, J.R., & Chenoweth, N.A. (2001). Fluency in Writing: Generating Text in L1 and L2. Written Communication, *18*(1), 80–98. https://doi.org/10.1177/0741088301018001004

Hayes, J.R., & Flower, L.S. (1980). *Identifying the organization of writing processes*. In L. Gregg and E. R. Steinberg (Eds.), Cognitive Processes in Writing (pp. 3-30). Lawrence Erlbaum.

Horstmanshof, L., & Brownie, S. (2013). 'A scaffolded approach to discussion board use for formative assessment of academic writing skills'. *Assessment & Evaluation in Higher Education*, *38*(1)1, 61-73, https://www.learntechlib.org/p/132972/.

Huang, F.Q. (2003). On curriculum for learning: review from perspective of cultural Philosophy. *Peking University Education Review*, 4(90–94), 99.

Hume, A., & Coll, R. (2009). Assessment of learning, for learning, and as learning: New Zealand case studies. *Assessment in Education*, *16*(3), 269-290, https://doi.org/10.1080/09695940903319661

Hyland, K. (2014). *Writing: Texts, Processes and Practice*s, Routledge, Taylor and Francis Group.

Ibrahim, H.A. (2013). In Search for implementing Learning-Oriented Assessment in an EFL Setting. *World Journal of English Language*, *3*(4), 12-14, ISSN 1925-*f*0703, https://doi.org/10.5430/wjel.v3n4p11

Jones, N., & Seville, N. (2016). *Learning Oriented Assessment: A systematic approach* (Vol 45). Cambridge University Press, p.14.

Kathpalia, S., & Heah, C. (2008). Reflective writing: Insights into what lies beneath. *RELC Journal*, *39*(3), 300–317, https://doi.org/10.1177/0033688208096843

Katstra, J., Tollefson, N., & Gilbert, E. (1987). The effects of peer evaluation on attitude toward writing and writing fluency of ninth grade students. *Journal of Educational Research*, *80*(3), https://doi.org/10.1080/00220671.1987.10885745

Keppell, M., Au, E., Ma, A. & Chan, C. (2006). Peer learning and learning-oriented assessment in technology-enhanced environments. *Assessment & Evaluation in Higher Education*, *31*(4), 453-464. https://doi.org/10.1080/02602930600679159

Krause, K-L, & O'Brien, D. (1999). A sociolinguistic study of the argumentative writing of Chinese students. *Education Journal*, *27*(2), 43-64.

Lam R. (2019). Writing portfolio assessment in practice: individual, institutional, and systemic issues, Pedagogies: *An International Journal, 15*(3), 169-182. https://doi.org/10.1080/1554480X.2019.1696197

Manzi, J., Flotts, P., & Preiss, D.D. (2012). *Design of a college level test of written communication: Theoretical and methodological challenges*. In E.L. Grigorenco, E. Mambrino & D.D. Preiss (eds), Writing: A mosaic of new perspectives. Taylor and Francis Group, LLC, 385-400.

Martinez, M., & Lipson, J. (1989). Assessment for learning. *Educational Leadership*, *46*(7), 73-75. https://files.ascd.org/staticfiles/ascd/pdf/journals/ed_lead/el_198904_martinez.pdf

Mok, M.M.C. (2013). *Self-directed learning-oriented assessments in the Asia-Pacific*. Springer

Nicolás-Conesa, F., Roca de Larios, J., & Coyle, Y (2014). Development of EFL students' mental models of writing and their effects on performance. *Journal of Second Language Writing*. 24. http://dx.doi.org/10.1016/j.jslw.2014.02.004

Paek, J.K., & Kang, Y. (2017). Investigation of content features that determine Korean EFL learners' argumentative writing qualities. *English Teaching*, *72*(2), 101-122.

Purpura, J.E., & Turner, C.E. (2014). *A learning-oriented assessment approach to understanding the complexities of classroom-based language assessment, Presentation at the Roundtable on Learning-Oriented Assessment in Language Classrooms and Large-Scale Contexts*, Teachers College, Columbia University, New York.

Romova, Z., & Andrew, M. (2011). Teaching and assessing academic writing via the portfolio: Benefits for learners of English as an additional language. *Assessing Writing*, *16*, 111-122. https://doi.org/10.1016/j.asw.2011.02.005

Swales, J.M. (1990). *Genre Analysis: English in academic and research settings*. Cambridge University Press.

Stiggings, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan*, *87*(4), 324–328.

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, *14*, 153–173. https://doi.org/10.1016/j.jslw.2005.05.002

Storch, N. (2019). Collaborative writing. *Language Teaching*, *52*(1), 40-59. https://doi.org/10.1017/S0261444818000320

Strijbos, J.W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*(4), 291–303. https://doi.org/10.1016/j.learninstruc.2009.08.008

Trevelyan, R., & Wilson, A. (2012). Using Patchwork Texts in assessment: Clarifying and categorising choices in their use. *Assessment & Evaluation in Higher Education*, *37*(4), 487-498. https://doi.org/10.1080/02602938.2010.547928

Wang, L. (2019). Effects of regulation on interaction pattern in web-based collaborative writing activity, *Computer Assisted Language Learning*, https://doi.org/10.1080/09588221.2019.1667831

Westbury, I., Hopmann, S., & Riquarts, K. (2000). *Teaching as a reflective practice: The German didaktik tradition*. Routledge.

Wu, W.V., Petit, E., & Chen, C. (2015). EFL writing revision with blind expert and peer review using a CMC open forum. *Computer Assisted Language Learning*, *28*(1), 58-80. https://doi.org/10.1080/09588221.2014.937442

William, D., & Thompson, M. (2007). *Integrating assessment with learning: what will it take to make it work?* In: Dwyer, CA, (ed.) The Future of Assessment: Shaping Teaching and Learning. (pp. 53-82). Lawrence Erlbaum Associates.

Winter, R. (2003). Contextualizing the Patchwork Text: Addressing problems of course assessment in higher education. *Innovations in Education and Teaching International*, *40*(2), 112-122. https://doi.org/10.1080/1470329031000088978

Vallberg Roth, A.C. (2014). Bedömning i förskolans dokumentationspraktiker: Fenomen, begrepp ochreglering. Göteborgs universitet. *Institutionen för pedagogik och didaktik*, *19*(4–5), 403–437.

Zeng, W., Huang, F., Yu, L., & Chen, S. (2018). Towards a learning-oriented assessment to improve students’ learning a critical review of literature. *Educational Assessment Evaluation and Accountability*, *30*, 211–250. https://doi.org/10.1007/s11092-018-9281-9

APPENDIX

APPENDIX A: PRE-TEST

Student Name: ________ Overall Grade: ____/____

Please write an Argumentative Essay for the following topic:

- *The education system should be improved in parallel with the technological developments in communication.*

APPENDIX B: POST-TEST

Student Name: _______ Overall Grade: ____/____

Please write an Argumentative Essay for the following topic:

- *Increased media use creates behavior problems*

APPENDIX C: ESSAY GRADING CRITERIA*

	ORGANIZATION	CONTENT	GRAMMAR & PUNCTUATION	LEXIS
25	• The introduction begins with a hook or general statement • Introduction successfully narrows down to the thesis, the introduction ends with an explicit thesis statement • Body paragraphs contain clear topic sentences, elaborate on the thesis, are an appropriate length, and are well-connected with transition words • The conclusion summarizes the main points in the body or restates the thesis, and finishes with a concluding remark	• **Fully** addresses the question at hand • **All** main points are elaborated and explained thoroughly with sufficient supporting details that provide **full** reasoning and exemplification • Paragraphs are very clear, coherent, and unified	• Skillful command of language with almost no grammatical errors • Level appropriate and varied sentence structure • Very good use of punctuation and capitalization	• Wide range of level appropriate vocabulary • Almost no word formation errors and almost impeccable spelling
20	• The introduction has a hook or general statement but may not successfully connect to the thesis statement • Introduction somewhat successfully narrows down to the thesis and has a clear thesis statement • Body paragraphs have satisfactory topic sentences and elaborate on the thesis statement with sufficient use of transitional signals • The conclusion summarizes the main points but might have repeated the thesis word-for-word	• **Sufficiently** addresses the question at hand • Presents a developed and sufficient argument • Main points are supported with information that provides **adequate** reasoning and exemplification • Paragraphs are clear, coherent, and unified	• Good command of language with minor grammatical errors that do not impede understanding • Level appropriate sentence structure and adequate range • Good use of punctuation and capitalization	• Sufficient range of level appropriate vocabulary • Few word formation errors with mostly accurate spelling.
15	• The hook or general statements do not lead to the thesis statement / narrowing down not successful / may start too general or too specific • Attempt to create a thesis statement, but may be unclear or may not pose a stance • Topic sentences are unclear/weak or not well connected to the thesis • Body paragraphs are too short or not divided proportionately or not well-connected with insufficient use of transition signals • The arguments are somewhat reviewed in the conclusion / a new idea might be introduced	• **Somewhat** responds to the question at hand • There may be **more than one central argument** / some supporting ideas may be **irrelevant** • Content may have **inadequate** or **excessive** information or examples • Repetition of ideas either in the same paragraph or other paragraphs • Some **effort** may have been made to write coherently and clearly	• Some structures are accurate but sentence structure errors predominate / many minor errors. • Only a limited range of level appropriate sentence structure/ attempts to use level-appropriate sentence structures with some mistakes • Limited command of punctuation and capitalization	• Somewhat sufficient range of level appropriate vocabulary • Some major word formation errors and spelling errors that do not impede understanding
10	• There is an introduction but there is no hook or general statement, or general statements leading to thesis are irrelevant or non-existent. • There is a thesis statement but it is vague, or weak. • Topic sentences are non-existent or they are contradictory to the thesis and /or they do not correspond to the thesis. • Disproportionate paragraphs and use of only simple transition signals • The conclusion lacks a summary of the body / the arguments are not reviewed	• Barely responds to the question at hand. • The main argument may be too **vague**, **weak**, or **underdeveloped**/ Several arguments may have been made, but no central idea is in focus • Presents **inadequate** information with little or no supporting details • Limited clarity, coherence, or unity	• Weak command of language with many grammatical errors so much as to hinder comprehension • Sentence structures below level expectations / only simple sentences • Use of punctuation and capitalization below level expectations.	• Limited range of level appropriate vocabulary • Frequent errors of word forms that confuse meaning with many spelling errors
5	• Produces a simple written text (not in essay form) that lacks cohesion. Inappropriate paragraphing, no thesis statement, no conclusion	• Fails to respond to the question. • Produces a simple written text that shows minimal coverage of the assignment/task. No consistency, no unity. Not enough ideas or information to support ideas	• So many grammatical errors that comprehension is impossible	• Range and accuracy of lexis fall significantly short with too many errors in word formation and spelling
0	***NOT ENOUGH OF A SAMPLE TO GRADE***			

*Adapted from Istanbul Şehir University, School of Foreign Languages, Testing Department / IELTS Writing Grading Rubric / British Council / University of Cambridge

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 218–231

https://doi.org/10.21449/ijate.1249297

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

Automatic item generation for online measurement and evaluation: Turkish literature items

Ayfer Sayin[1,*], **Mark J. Gierl**[2]

[1]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]University of Alberta, Faculty of Education, Department of Educational Psychology, Alberta, Canada

ARTICLE HISTORY

Received: Feb. 09, 2023
Revised: May 24, 2023
Accepted: May 30, 2023

Keywords:

Automatic item generation,
Large item pool,
Computer-based testing,
Online measurement, and evaluation,
Turkish literature.

Abstract: Developments in the field of education have significantly affected test development processes, and computer-based test applications have been started in many institutions. In our country, research on the application of measurement and evaluation tools in the computer environment for use with distance education is gaining momentum. A large pool of items is required for computer-based testing applications that provide significant advantages to practitioners and test takers. Preparing a large pool of items also requires more effort in terms of time, effort, and cost. To overcome this problem, automatic item generation has been widely used by bringing together item development subject matter experts and computer technology. In the present research, the steps for implementing automatic item generation are explained through an example. In the research, which was based on the fundamental research method, first a total of 2560 items were generated using computer technology and SMEs in field of Turkish literature. In the second stage, 60 randomly selected items were examined. As a result of the research, it was determined that a large item pool could be created to be used in online measurement and evaluation applications using automatic item generation.

1. INTRODUCTION

Automatic item generation (AIG) is a research area where cognitive and psychometric theories and computer technology are used to create content for tests (Gierl & Haladyna, 2012; Gierl & Lai, 2013; Irvine & Kyllonen, 2002). AIG, which was first introduced by Bormuth (1969), helps to generate large numbers of items to item pools quickly and economically (Gierl & Haladyna, 2012; Gierl & Lai, 2013; Irvine & Kyllonen, 2002; Sinharay & Johnson, 2005). In this research, it is aimed to introduce the steps of AIG in detail, based on an item in the Field Proficiency Test (original acronym: AYT), the second session of the Higher Education Institutions Examination (original acronym: YKS) in Türkiye. In addition, the process of the examination of automatically generated items is explained.

It has been determined that there are difficulties in measurement and evaluation in the reports regarding the distance education process, which was rapidly passed in 2019 with the COVID-19 pandemic. Turkish Education Association Think Tank (original acronym: TEDMEM) stated that in this process, there were differences in testing practices and policies, especially in higher

*CONTACT: Ayfer Sayin ✉ ayfersayin@gazi.edu.tr Gazi Faculty of Education, Department of Educational Sciences, Division of Assessment and Evaluation in Education, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

education institutions; while some universities made pass-fail decisions, some wanted grades not to be included in the academic achievement grade point average (TEDMEM, 2020). These differences in testing practices between universities indicate that there is no universal agreement on measurement and evaluation practices for distance education. Similarly, although the courses are held remotely by the Ministry of National Education (original acronym: MEB) in Türkiye, mid-term exams were carried out face-to-face in Türkiye at the some grades. (MEB, 2021). Research have shown that most teachers cannot measure and evaluate in the distance education process (Saygı, 2021); shows that they also have a negative attitude toward online measurements and evaluation (Sarı & Nayır, 2020; Saygı, 2021). Teachers stated that they had difficulties in measurement and evaluation practices in the distance education process, especially due to security problems such as cheating (Adıgüzel, 2020; Kınalıoğlu & Güven, 2011). Cheating in test applications, insufficient evidence for the psychometric properties of the items; situations such as students plagiarizing in their homework, sharing the load by one or more students in group work; it also constitutes a problem of face-to-face measurements and evaluations. To prevent these problems, which are even more prominent in measurement and evaluation practices in distance education, individualized tests, different tests consisting of items of equivalent quality to each student can be applied. Measurement and evaluation problems in distance education are not limited to summative tests in which students' achievement is measured.

Traditional measurement and assessments and training focus on leveling to determine the extent to which students have achieved their goals rather than process assessments. Modern measurement and assessments, together with the process and the generation, come to the fore. At this point, it is necessary not only to apply for an exam to the students at the end of a certain period but also to make applications so that the students can see their deficiencies in the process. This process, in which formative tests are used, aims to support students' learning process by giving effective feedback (Bennett, 2011; Gierl & Lai, 2018). Similarly a more systematic and continuous assessment and evaluation approach should be adopted in the distance education process (Gaytan & McEwen, 2007). Balta and Türel (2013) state that if measurement and assessments are carried out continuously, students can be given feedback by considering their differences, which will significantly increase the quality of learning in distance education. For this, it is necessary both to implement practices to improve teachers' online measurement and evaluation skills and to create a digital infrastructure in schools.

Within the scope of the Movement to Increase Opportunities and Improve Technology (original acronym: FATİH) Project carried out by the MEB, smart boards were placed in the classrooms, and the way for students to use tablets in the education environment was paved (Artırma & Hareketi, 2020). Similarly, the Education Information Network (original acronym: EBA) created by the MEB has been widely used both inside and outside the school since 2012 (MEB, 2020). Adaptive learning environments added to the network in 2019, it is becoming even more effective. Measurement Selection and Placement Center (original acronym: ÖSYM) has been conducting some applications of the Foreign Language Exam in the computer environment as called e-YDS since 2014 in Türkiye (ÖSYM, 2022). It is considered that these applications will become more widespread in the coming years. However, in this electronic test, all candidates are presented with a single booklet (all items are the same even if the items and options’ place are different) at the same time. This is essentially a computerized version of a paper-and-pencil test and therefore does not provide the advantages of computer-based testing.

In today's world of rapidly developing technology, test development, and administration processes have begun to integrate with computer technology. The advantages of computer-based applications such as portability, usability, and the absence of time and space constraints have paved the way for the widespread use of computer applications in the field of measurement and evaluation (Chen et al., 2019; Gierl & Haladyna, 2012; Lane et al., 2015; Weber et al., 2003). For instance, the American College Test (ACT), which is one of the entrance exams for higher education institutions in the United States, Graduate Management Admission Test (GMAT) and Graduate Record Examinations (GRE) applied for graduate education; Transition tests (MCCQE, NCLE) in medical education applied in Canada are computer-based. The fact that computer-based tests provide flexible administration time and place for all students and that scores are calculated quickly, it is attractive not only for large-scale but also for in-class applications (Chen et al., 2019; Gierl et al., 2021). It is a rapid transition to computer-based tests in all classroom applications from preschool to secondary education, called K-12 (Gierl et al., 2022).

In classroom practices, teachers tend to require a long time to score the tests because of the large number of students in the classroom, the intensity of the lesson, and other works. When tests require a long time to score, students do not have enough time to follow their individual developments, to see their shortcomings, and to make applications to make up for them (Choi & Zhang, 2019; Clark & Rust, 2006). This reduces the effectiveness of feedback. Since computer-based test applications quickly explain the results to the students, they can follow their individual progress and manage their own learning processes (Corbett & Anderson, 2001; Zhu et al., 2020). In short, by taking advantage of digital education, it is necessary to establish a measurement and evaluation system that prioritizes the individual differences of the student, increases the motivation of the student as well as academic success, supports group work and communication, and allows interventions for the problems of distance education. For this, it is necessary to utilize technology not only in the application infrastructure but also in the creation of a large item pool.

While the test development process is difficult on its own, creating a large item pool requires much effort in terms of labor, time, and cost (Gierl & Lai, 2013; Kosh et al., 2019). Studies also demonstrate that teachers already have difficulties in writing items (Karatay & Dilekçi, 2019; Özyalçın & Kana, 2020). Especially in the distance education process, the teachers wanted to apply different tests consisting of equivalent items to the students for test security reasons. However beside to the cost of test development, an important problem is encountered in the process of writing an equivalent item to measure the same outcome: "subjectivity." An item is an expression of the knowledge, skills and competence of the SMEs who created the item. For this reason, the items created carry traces of the item author, and for this reason, item creation is expressed as "art" according to some researchers (Rodriguez, 2005). However, equivalent items, especially those in different forms, should have comparable content and psychometric properties. However, while there are differences even among the items created by the same item developer, it is challenging for different item developers to ensure the equivalence of the items. Haladyna and Rodriguez (2013) states that test development procedures are applied by most item writer, therefore, asking for items is a science, but the items still contain subjectivities from their authors (Gierl & Haladyna, 2012). Therefore, it is necessary to combine experts and technology in test development and to take advantage of the innovations that computer-based testing systems bring to the content, administration environment and assessment aspects of tests, as well as to the way items are written (Davey, 2011). As one of them, automatic item generation can be used as an alternative item development process to create a large item pool. In addition, in classroom applications, a large item pool can be obtained by overcoming this problem with AIG. Simultaneously, the reliability of the test can be ensured by conducting content coding studies at AIG (Gierl et al., 2022).

1.1. Automatic Item Generation (AIG)

AIG, is an item generation method that combines models, content expertise and computer technology to create large and efficient test banks in a short time (Gierl et al., 2021). AIG which has started to spread to psychology, education, and computer sciences, is mainly carried out for two purposes: The first is to generate equivalent items with similar difficulty and psychometric properties, and the second is to create an item pool with items with different item difficulty ranges (Sinharay & Johnson, 2005). In other words, it is aimed to generate the desired quality and number of items in the item pool. Considering that time, effort, and cost must develop a test (Gierl et al., 2016), AIG provides significant cost savings while creating a large item pool (Kosh et al., 2019). Another advantage of AIG is that it allows quick and effective feedback on test results to students through practice (Gierl & Lai, 2018). In this way, students can follow the development of their individual learning.

Generally, the AIG process is carried out in two different ways: artificial intelligence-based and template-based (Gierl & Lai, 2013). In this research, the stages of template-based automatic item generation are introduced in an applied way through an example. Template-based AIG has three stages: In the first stage (i) a "cognitive model" is developed for the AIG, in the second stage (ii) an "item model" is developed in which the cognitive model content is embedded to generate new items. In the third and final stage (iii), computer algorithms are used to place the content of the cognitive model developed in the first stage into the item model developed in the second stage (Gierl & Lai, 2013). As a result, the template-based automatic item generation method, which includes these three stages, if the generated items have an equivalent and predictable difficulty level, the result can be considered successful.

2. METHOD

2.1. Design

In this research was conducted within the framework of the qualitative research model. Since the research will be based on experiments and theory aiming to acquire new information about the foundations of facts and observable facts, it can be considered fundamental research. Fundamental research is studies conducted to examine, analyze, strengthen a theory related to a certain field or to put forward a new theory (Kaptan, 1998).

2.2. Participants

In the research, the opinions of five experts, including a measurement and evaluation expert and four academicians in the field of Turkish language and literature education, were taken in collecting the validity and reliability evidence of the models developed for AIG.

2.3. Data Tools

In the research items were generated in the field of Turkish literature with AIG. AIG usually starts with a parent item. To demonstrate its applicability in this study, the YKS application with the highest number of candidates in Türkiye was chosen, and the 12th item of the AYT in the second session of YKS in 2018 was identified as the parent item. It is shown in Figure 1.

Figure 1. *12th item for 2018 in the field proficiency test.*

12. He is an 18th century Classic Turkish literature poet. He is considered one of the important representatives of the Sebk-i Hindî. He became the owner of a divan at a young age. He proved his success in mesnevi writing with his allegoric work Hüsnü Aşk.

Which of the following is the poet mentioned in this passage?

A) Süleyman Çelebi B) Ahmet Pasha

C) Sehi Bey D) Taşlıcalı Yahya

E) Âşık Pasha

The test items in the YKS application are open access. The main item in Figure 1 was accessed from https://dokuman.osym.gov.tr/pdfdokuman/2018/YKS/TSK/ayt_yks_2018.pdf OSYM's own website. Based on the parent item, first the "cognitive model" and "item model" were created. This process is introduced in the findings section.

2.4. Analysis

The generated items were analyzed according to SMEs opinion and similarity index. SMEs' opinions on the models and items

- scientific accuracy,
- language and expression,
- item difficulty

examined in this regard.

Cosine Similarity Index (CSI) also was used to determine the similarity of automatically generated items (Gierl & Lai, 2013). CSI refers to the similarity between the vectors of the texts in the two items. It is calculated using the cosine of the angle between two vectors in the multidimensional space of unique words. In other words, CSI is a word similarity measure calculated using an algorithm based on the text-vector indexing technique. CSI is calculated with the formula in 1.1 (Bayardo et al., 2007).

$$\cos(\vec{A},\vec{B}) = \frac{\vec{A}\,.\,\vec{B}}{\|\vec{A}\|\,\|\vec{B}\|} \qquad (1.1)$$

In the equation in 1.1, A, and B are expressed in a binary vector of word formations. First, the length of the binary vector is found by determining the total unique number of words in texts A and B. Then, vectorization is performed for each text, depending on whether the words are in the A and B texts. The CSI value takes a value between 0 and 1; the closer to 0, the less similarity; it increases as it approaches 1. When CSI is 0, it means that there is no common word in the items generated, and when it is 1, it means that all words are common.

3. FINDINGS

The results of the research are reported in line with AIG's steps: (i) development of a cognitive model, (ii) development of an item model, (iii) generation of the items with computer technology.

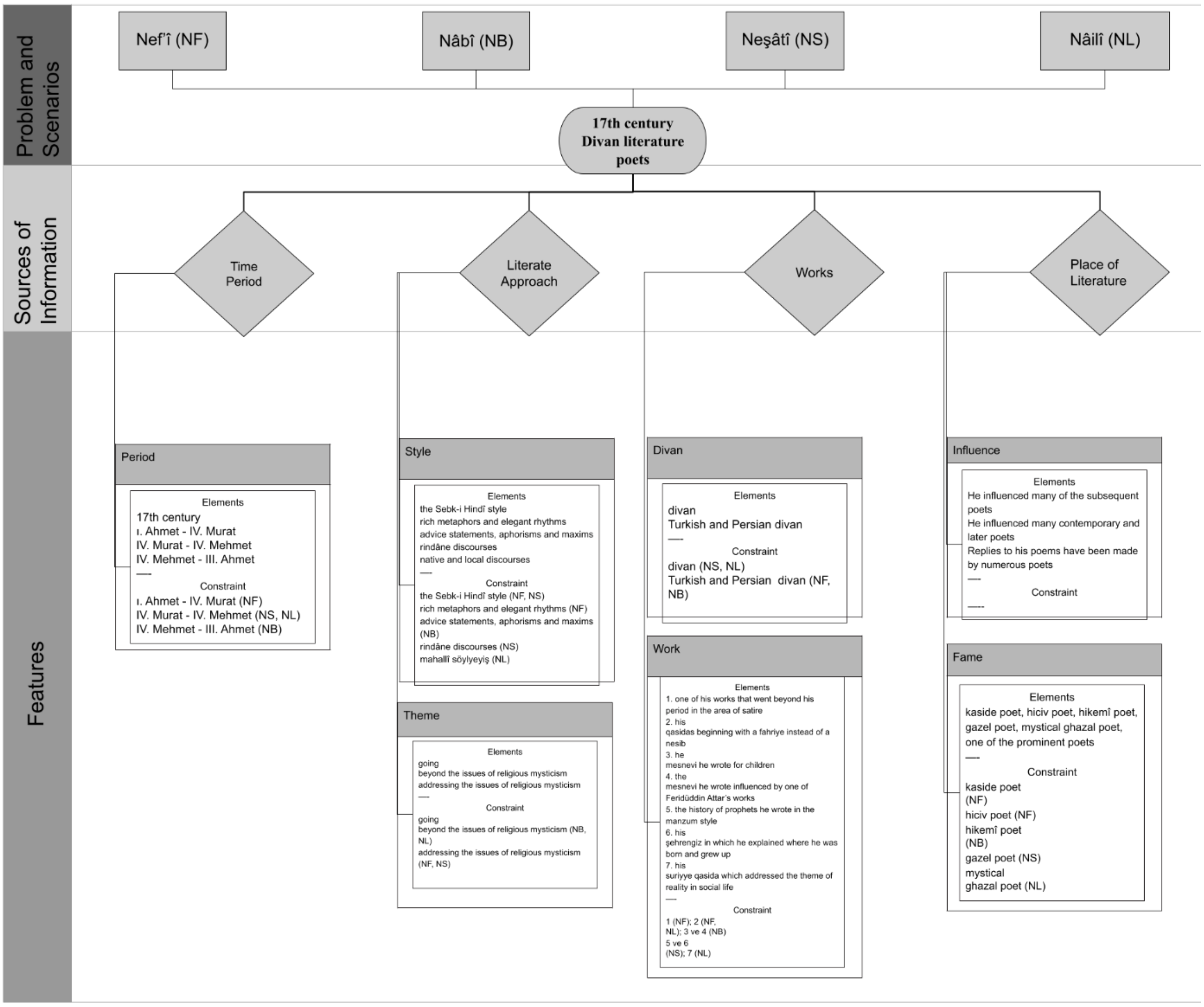
3.1. Cognitive Model Development

A cognitive model is an organizational diagram that includes the essential information, skills, and proficiencies and how they are used to solve a certain item. In this organizational diagram, the sources of information necessary to answer an item and the elements within each source of information are identified (Gierl et al., 2021). Thus, not only does the information needed to establish a cognitive template be identified, but effective feedback can also be given to students after an exam administration (Chen et al., 2019; Gierl et al., 2016).

At the first stage, the feature aimed at measuring the 12th item of the 2018 AYT (ÖSYM, 2018), was determined. Within the scope of the Turkish language and literature course, the item that focuses on "divan literature poets" is "Knows the poet's life and view to evaluate the relationship between poet and poetry." at A.1.10 in the curriculum (MEB, 2018). After the general features were identified, the next stage was to define the necessary information, skills, and content essential to answer the 12th item and for this reason the item was examined. The item is answered based on the key features of the "period" in which the poet lived, the "place" of the poet within the divan literature, the situation of being a "divan", the poet's "fame" in the tradition of poems, and the prominent "work" of the poet. The expert opinions of four academicians in Turkish language and literature education revealed that the features of poet's "place" in the literature and the poet's "fame" were very similar and overlapping. Considering that Sebk-i Hindî in the sample item is a feature of style, the "place" feature was expressed as two separate features: “style” and “theme”. Moreover, the “influence” feature in similar items in the AYT was added to finalize the key features.

After the key features to form the cognitive model were identified, then the correct responses and the elements and limitations for each correct response were defined (Gierl et al., 2021), so that the cognitive model was developed. In the present study, four poets from the 17th century divan literature were chosen as samples for the cognitive model; however, another model in which all the centuries are taken into consideration, for example, can also be formed. The elements and limitations for the key features previously mentioned for each poet were defined and placed into an organizational structure. For example, although all poets lived in the 17th century, the rulers of their time differed. While Nef’î lived during the period of four padishahs from Ahmet I to Murat IV, Nâbî lived in the periods of Mehmet IV to Ahmet III. The Encyclopedia of Islam of the Religious Foundation of Türkiye (orig. TDV) was used as the main source for determining the characteristics of Divan literature poets (TDV, 2022). After revisions were made in the cognitive model developed based on the expert opinions obtained from four academicians in Turkish language and literature education, the cognitive model was finalized (see Figure 2).

Figure 2. *Cognitive model development in the first step of AIG.*


Problem and Scenarios
Sources of Information
Features
Nef’î (NF)
Nâbî (NB)
Neşâtî (NS)
Nâilî (NL)
17th century Divan literature poets
Time Period
Literate Approach
Works
Place of Literature
Period
Elements
17th century
ı. Ahmet - IV. Murat
IV. Murat - IV. Mehmet
IV. Mehmet - III. Ahmet
Constraint
ı. Ahmet - IV. Murat (NF)
IV. Murat - IV. Mehmet (NS, NL)
IV. Mehmet - III. Ahmet (NB)
Style
Elements
the Sebk-i Hindî style
rich metaphors and elegant rhythms
advice statements, aphorisms and maxims
rindâne discourses
native and local discourses
Constraint
the Sebk-i Hindî style (NF, NS)
rich metaphors and elegant rhythms (NF)
advice statements, aphorisms and maxims (NB)
rindâne discourses (NS)
mahallî söylyeyiş (NL)
Theme
Elements
going
beyond the issues of religious mysticism
addressing the issues of religious mysticism
Constraint
going
beyond the issues of religious mysticism (NB, NL)
addressing the issues of religious mysticism (NF, NS)
Divan
Elements
divan
Turkish and Persian divan
Constraint
divan (NS, NL)
Turkish and Persian divan (NF, NB)
Work
Elements
1. one of his works that went beyond his period in the area of satire
2. his
qasidas beginning with a fahriye instead of a nesib
3. he
mesnevi he wrote for children
4. the
mesnevi he wrote influenced by one of Feridüddin Attar’s works
5. the history of prophets he wrote in the manzum style
6. his
şehrengiz in which he explained where he was born and grew up
7. his
suriyye qasida which addressed the theme of reality in social life
Constraint
1 (NF); 2 (NF, NL); 3 ve 4 (NB)
5 ve 6
(NS); 7 (NL)
Influence
Elements
He influenced many of the subsequent poets
He influenced many contemporary and later poets
Replies to his poems have been made by numerous poets
Constraint
Fame
Elements
kaside poet, hiciv poet, hikemî poet, gazel poet, mystical ghazal poet, one of the prominent poets
Constraint
kaside poet
(NF)
hiciv poet (NF)
hikemî poet
(NB)
gazel poet (NS)
mystical
ghazal poet (NL)

3.2. Item Model Development

In the second stage, an "item model" is developed in AIG. An item model defined where the key features will be placed. It includes basic properties and elements that can be changed to generate new items (Gierl et al., 2021). This place can occur in the stem of the item, in the question prompt and in the options (correct answer and distracters). In this stage, ancillary features including text, pictures, tables, graphs, and diagrams can also be added to the item model, and random variables that can be manipulated, although not necessary, to solve the problem (Gierl & Lai, 2013).

The item model can be created in a one-layer or n-layered manner (Gierl & Lai, 2013). The purpose of item generation using the one-layer model is to generate few items by manipulating only the basic variables in the cognitive model. While fewer items are generated according to the one-layer item model, many more items can be generated compared to the n-layered model. Since the manipulations in a layered model are limited only by the number of elements, the similarity of the generated items is very high. This causes the generated items to be called "clones." In the n-layered model, the language used in the body, item sentence and options by using the syntactic structures of the language; structured hierarchically. Using the possibilities of the language, the content, key features, and elements can be manipulated by embedding them into each other. In this way, much more items can be generated from a cognitive model, and less similar items can be obtained (Gierl et al., 2021). An example of both one-layer and n-layer models can be seen in Table 1.

Table 1. *Examples of one and n-layer models.*

Stem (one layer)	He is a Classical Turkish literature poet who lived in the <**period**>. He is considered among important poets who skilfully used the <**style**>. The poet has <**divan**>. <**Influence**> He is considered as a <**fame**> in the divan poetry tradition. His works were <**theme**> of religious mysticism and it was proved his skill in writing with <**work**>.
Stem (n-layer)	*Period, style and theme + work, divan and fame* **Period, style and theme** 1. He is a Classical Turkish literature poet who lived during the <**Period**>. He is shown to be among the important poets who used the <**theme**> religious mysticism and <**style**> skilfully in his poems. 2. He is among the poets who used the <**theme**> religious mysticism and <**style**> skilfully in their poems. He is one of the divan literature poets who lived in the <**period**>. **Work, divan and fame** 1. The poet has the <**divan>** divan. <**Influence**> In the divan poetry tradition is referred to as a <**fame**> . He proved his skill in writing with his work <**work**>. 2. <**Influence**> He gained fame in the period he lived and in divan literature as a <**fame**>. The poet has the <**divan**> divan. The poet demonstrated his skill in <**work**>. 3. The poet has the <**divan**> divan and he has become identified with <**work**>. <**Influence**> He is referred to as <**fame**> in divan poetry. 4. <**Influence**> He displayed his skill in writing with <**work**> . The poet has the <**divan**> divan. In divan literature, it is known as <**fame**>.
Item Promtp	**Which of the following is the poet mentioned in this passage?**
Correct Options	Nef'î, Nâbî, Neşâtî, Nâilî

3.4. Generating Items Using Computer Technology

In the third and final stage of AIG, the model content created in the first step using computer technology is placed in the item model developed in the second stage, and item generation is carried out (Gierl et al., 2021). Different software are developed for item generation in the literature: Math Test Creation Assistant (Singley & Bennett, 2002), ModelCreator (Higgins et al., 2005), Item Distiller (Higgins, 2007), IGOR (Gierl & Lai, 2012), EAQC (Gutl et al., 2011), MARTEN (https://www.mghlpartners.com/software). In the present study, items were generated using scripts written in Phyton. After the third stage of AIG was successfully completed, 320 items in a one-layer model and 2560 items in a n-layer model were automatically generated. Samples of the generated items are presented in Table 2. The original items in Turkish are added in the Appendix.

Table 2. *Samples of the generated items.*

Sample items generated with the one-layer model	**Sample items generated with the n-layer model**
1. He is a Classical Turkish literature poet who lived in the 17th century. He is considered among important poets who skilfully used the Sebk-i Hindî style. The poet has Turkish and Persian divan. He influenced many contemporary and later poets. He is considered as a kasida poet in the divan poetry tradition. His works were going beyond the issues of religious mysticism and it was proved his skill in writing with his qasidas beginning with a fahriye instead of a nesib. **Which of the following is the poet mentioned in this passage?** A) Nef’î* B) Nesîmî C) Nedim D) Şeyh Gâlip E) Nâbî	**1.** He is one of the divan literature poets who lived in the periods of four padishahs between IV. Murad, Sultan İbrâhim & IV. Mehmet. He is among the poets who used the going beyond the issues of religious mysticism and Sebk-i Hindî style skilfully in their poems. The poet has a divan where he collects his poems and is identified with the history of the prophets, which he wrote in verse. He influenced many contemporary and later poets. In divan literature, it is known as one of the prominent poets. **Which of the following is the poet mentioned in this passage?** A) Bâkî B) Nâbî C) Şeyh Galip D) Neşatî* E) Nedim
127. He is a Classical Turkish literature poet who lived in the periods of four padishahs between Mehmet IV and Ahmet III. He is considered among important poets who skilfully used the native andd local discourses. The poet has Turkish and Persian divan. He is considered as a hikemî poet in the divan poetry tradition. His works were addressing the issues of religious mysticism and it was proved his skill in writing with the mesnevi he wrote influenced by one of Feridüddin Attar’s works. **Which of the following is the poet mentioned in this passage?** A) Neşati B) Necati Bey C) Nâbî* D) Gülşehrî E) Nef’î	1368. The poet is shown to be among the important poets who used the addressing the issues of religious mysticism and hikemî poet skilfully in his poems. He is a Classical Turkish literature poet who lived during the 17th century. Replies to his poems have been made by numerous poets. He became famous as a mystical ghazal poet in the period he lived and in divan literature. The poet has a divan where he collects his poems. He showed his poetic skill with his syria eulogy, which is about the reality in social life. **Which of the following is the poet mentioned in this passage?** A) Fuzûlî B) Nâilî* C) Nedim D) Şeyh Galip E) Nef’î
205. He is a Classical Turkish literature poet who lived in the 17th century. He is considered among important poets who skilfully used the rich metaphors and elegant rhythms. The poet has Turkish and Persian divan. He is considered as a hiciv poet in the divan poetry tradition. His works were going beyond the issues of religious mysticism and it was proved his skill in writing with his qasidas beginning with a fahriye instead of a nesib. **Which of the following is the poet mentioned in this passage?** A) Nesîmî B) Nedim C) Şeyh Galip D) Nâbî E) Nef’î*	2054. He is shown to be among the important poets who used the addressing the issues of religious mysticism and the native andd local discourses skilfully in his poems. He is the one of the Classical Turkish literature poet who lived in the 17th century. He showed his mastery in writing with his masnavi written for children. The poet has Turkish and Persian divan. He is known as a hikemi poet in Divan literature. **Which of the following is the poet mentioned in this passage?** A) Neşati B) Necati Bey C) Gülşehrî D) Nâbî* E) Nef’î

3.4. Examining the generated items by AIG

After the three steps of AIG were carried out, SMEs' opinion was used to evaluate the generated items, and then the similarity ratios of the generated items were calculated. At this stage, the opinions of four SMEs in the field of Turkish and literature education were taken. 60 items randomly selected among the automatically generated items, and they were examined by SME. In the items the experts examined, they reported that there were no errors in the linguistic accuracy and that the information given to answer the items was sufficient and clear. The experts whose opinions were received within the scope of the study revised the wordings of some of the automatically generated items, and the item model was redeveloped by taking these revisions into consideration. According to the four academicians, who stated that the difficulty level of the items varied, the easiest item was the one with the correct answer Nef'î and the most difficult item was the one with the correct answer Nâilî. The experts stated that the difficulty level of the items with the same correct answer did not significantly vary, and they were equivalent.

In the second stage of examination, Cosine Similarity Index (CSI) was used to determine the similarity of automatically generated items. The CSI value of the items generated with the one-layer model within the scope of the research was between 0.66 and 0.99; the CSI value of the items generated with the n-layered model was calculated between 0.27 and 0.99.

4. DISCUSSION and CONCLUSION

The advantages of computer-based applications, such as portability, usability, and lack of time and space restrictions, have paved the way for the widespread use of computer applications in the field of measurement and evaluation (Weber et al., 2003). Computer-based test systems bring innovations in the form of item preparation as well as the content they offer, the application environment and the evaluation dimension of the tests (Parshall et al., 2002). In the test development process with the help of computers, automatic systems that can generate different items from the same item pattern have gained importance (Gierl & Haladyna, 2012). One of these innovations is AIG, where computers are integrated into the test development process. In this research, we introduce how to automatically generate high-stakes test items for university entrance exams in Türkiye. Although AIG practices are largely in English (Embretson & Yang, 2006; Kosh et al., 2019; Lai et al., 2016), there are also practices in Korean (Choi et al., 2018; Gierl et al., 2021) and German (Arendasy & Sommer, 2012). The results of this research are important in terms of showing that AIG is both suitable for Turkish language items and can be used in high-stakes tests.

Additionally, in the literature, it is seen that there are studies in the field of medicine and mathematics in template-based automatic item generation (Colvin, 2014; Lai et al., 2016; Singley & Bennett, 2002; Sun et al., 2019). AIG should be used in psychological testing areas where cognitive models are involved, and individuals' reasoning skills should also be measured (Hommel et al., 2022; Sun et al., 2019; Yang et al., 2021) and cognitive ability items can be developed in the reasoning areas (Freund et al., 2008; Poinstingl, 2009). In the previous study, AIG was used to generate items in the field of Turkish literature. To indicate that automatic item generation is suitable for test development in different disciplines, automatic item generation is introduced through a literature item within the scope of 2018 AYT in this study. AIG can also be used in different disciplines.

A cognitive model was created in the first stage of AIG, and an item model was developed in the second stage. After the third stage, 320 items in a one-layer model and 2560 items in a n-layer model were automatically generated. In the first and second stages of template-based AIG, experts such as item writer, SME, measurement, and evaluation expert, as in traditional item writing, took part. This study demonstrates that unlike the traditional item writing process, the

SMEs were not involved in writing or examining the item, but in the process of creating the models that would form the item and examining these models. Gierl et al. (2016) stated that experts are still needed in the AIG process. However, it takes a very long time and intensive labor for an expert group of 5 experts to create 2560 items. Studies show that AIG significantly reduces the cost in terms of time, labor, and cost in the test development process (Alves et al., 2010; Gierl et al., 2021; Kosh et al., 2019).

Since evaluating the generated items is important, items were examined based on SMEs opinions and the CSI in this study. According to SMEs opinion, items with the same correct answer show equivalent characteristics; in the items in which different poets were asked, it was determined that items with different difficulty indexes were generated. Similarly based on the results obtained after the pre-application, Gierl et al. (2016) determined that the generated items are in different difficulty ranges. Ryoo et al. (2022) also stated that the item difficulties of the generated items are the similar and different. Therefore the result of this study is in consistency with the fundamental aim of AIG (Sinharay & Johnson, 2005). This shows that the items generated with AIG can be used in tests prepared for different purposes (summative, formative and diagnostic assessment or in-class and large-scale etc.), in other words, it is a common area of use.

It has been determined that the CSI value of the items generated with the 1-layered model CSI value from 0,66 to 0,99, while the n-layered model is calculated between 0,27 and 0,99. It means that the items generated by the n-layer model are less similar to each other and they are not clones. This result is consistent with other research results (Gierl & Lai, 2012). It is expected that the CSI values of the items generated with the n-layered model are low, and it is recommended to use the n-layered model for AIG (Gierl & Lai, 2013).

In this study, the examination of the automatically generated items was carried out by SMEs' opinions and by calculating the cosine similarity index. Pre-tests of the automatically generated items can be made, and the results of the pre-application and examination can also be carried out by calibrating. Using them is just as important as creating a large pool of items. Further studies can be conducted on the use of the item pool. In this study, template-based AIG approach is introduced. Artificial intelligence-based AIG studies can also be conducted. Simultaneously, international studies in automatic item generation depend on the changing data type (Chen et al., 2019), calibration shapes (Bai, 2019), and difficulty estimation methods (Chen et al., 2019; Gierl et al., 2016).

Effectiveness in classroom tests can be increased by creating item pools based on AIG on the EBA platform, which is widely used in schools created by the MoNE. A measurement and evaluation system can be created that all teachers can contribute and use in this process. In large-scale test applications, mobile computer-based applications can be disseminated, and AIG can be used in applications.

Acknowledgments

This research was supported by 2219-International Postdoctoral Research Fellowship Program for Turkish Citizens and the Scientific and Technological Research Council of Türkiye.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Gazi University, E-77082166-604.01.02-609178.

Authorship Contribution Statement

Ayfer Sayin: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Mark J. Gierl:** Methodology, Supervision, and Validation.

Orcid

Ayfer Sayin https://orcid.org/0000-0003-1357-5674
Mark J. Gierl https://orcid.org/0000-0002-2653-1761

REFERENCES

Adıgüzel, A. (2020). Teachers' views on distance education and evaluation of student success in the pandemic process. *Milli Eğitim Dergisi*, *49*(1), 253-271. https://doi.org/10.37669/milliegitim.781998

Alves, C.B., Gierl, M.J., & Lai, H. (2010). Using automated item generation to promote principled test design and development. *American Educational Research Association, Denver, CO, USA*.

Arendasy, M.E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and individual differences*, *22*(1), 112-117.

Artırma, F., & Hareketi, T.İ. (2020). *FATİH projesi*. Retrieved from: http://fatihprojesi.meb.gov.tr

Bai, Y. (2019). *Cognitive Diagnostic Models-based Automatic Item Generation: Item Feature Exploration and Calibration Model Selection*. Columbia University.

Balta, Y., & Türel, Y. (2013). An examination on various measurement and evaluation methods used in online distance education. *Turkish Studies-International Periodical For The Languages, Literature and History of Turkish or Turkic*, *8*(3), 37-45. http://dx.doi.org/10.7827/TurkishStudies.427

Bayardo, R.J., Ma, Y., & Srikant, R. (2007). Scaling up all pairs similarity search. Proceedings of the 16th international conference on World Wide Web.

Bennett, R.E. (2011). Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, *18*(1), 5-25.

Chen, B., Zilles, C., West, M., & Bretl, T. (2019). Effect of discrete and continuous parameter variation on difficulty in automatic item generation. Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, Proceedings, Part I 20,

Choi, J., Kim, H., & Pak, S. (2018). Evaluation of Automatic Item Generation Utilities in Formative Assessment Application for Korean High School Students. *Journal of Educational Issues*, *4*(1), 68-89.

Choi, J., & Zhang, X. (2019). Computerized item modeling practices using computer adaptive formative assessment automatic item generation system: A tutorial. *The Quantitative Methods for Psychology*, *15*(3), 214-225.

Clark, C.M., & Rust, F.O.C. (2006). Learning-centered assessment in teacher education. *Studies in Educational Evaluation*, *32*(1), 73-82.

Colvin, K.F. (2014). *Effect of automatic item generation on ability estimates in a multistage test*. University of Massachusetts Amherst.

Corbett, A.T., & Anderson, J.R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. Proceedings of the SIGCHI conference on Human factors in computing systems,

Davey, T. (2011). A Guide to Computer Adaptive Testing Systems. *Council of Chief State School Officers*.

Embretson, S., & Yang, X. (2006). 23 Automatic item generation and cognitive psychology. *Handbook of statistics*, *26*, 747-768.

Freund, P.A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied psychological measurement*, *32*(3), 195-210.

Gaytan, J., & McEwen, B.C. (2007). Effective online instructional and assessment strategies. *The American journal of distance education*, *21*(3), 117-132.

Gierl, M.J., & Haladyna, T.M. (2012). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M.J., & Lai, H. (2012). The role of item models in automatic item generation. *International journal of testing*, *12*(3), 273-298.

Gierl, M.J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*(3), 36-50.

Gierl, M.J., & Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied psychological measurement*, *42*(1), 42-57.

Gierl, M.J., Lai, H., Pugh, D., Touchie, C., Boulais, A.-P., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education*, *29*(3), 196-210.

Gierl, M.J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.

Gierl, M.J., Shin, J., Firoozi, T., & Lai, H. (2022). Using content coding and automatic item generation to improve test security. Frontiers in Education,

Gutl, C., Lankmayr, K., Weinhofer, J., & Hofler, M. (2011). Enhanced Automatic Question Creator--EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education. *Electronic Journal of e-Learning*, *9*(1), 23-38.

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

Higgins, D. (2007). Item Distiller: Text retrieval for computer-assisted test item creation. *Educational Testing Service Research Memorandum (RM-07-05). Princeton, NJ: Educational Testing Service*.

Higgins, D., Futagi, Y., & Deane, P. (2005). Multilingual generalization of the ModelCreator software for math item generation. *ETS Research Report Series*, *2005*(1), i-38.

Hommel, B.E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *psychometrika*, *87*(2), 749-772.

Irvine, S., & Kyllonen, P. (2002). *Generating items for cognitive tests: Theory and practice*. Erlbaum.

Kaptan, S. (1998). *Bilimsel araştırma teknikleri ve istatiksel yöntemleri.* Tekışık Ofset Tesisleri.

Karatay, H., & Dilekçi, A. (2019). Competencies of turkish teachers in measuring and evaluating language skills. *Milli Eğitim Dergisi*, *48*(1), 685-716.

Kınalıoğlu, İ.H., & Güven, Ş. (2011). Issues and solutions on measurement of student achievement in distance education. *XIII. Akademik Bilişim Konferansı Bildiriler*, 637-644.

Kosh, A.E., Simpson, M.A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice*, *38*(1), 48-53.

Lai, H., Gierl, M.J., Touchie, C., Pugh, D., Boulais, A.-P., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and learning in medicine*, *28*(2), 166-173.

Lane, S., Raymond, M.R., Haladyna, T.M., & Downing, S.M. (2015). Test development process. In *Handbook of test development* (pp. 19-34). Routledge.

MEB. (2018). *Ortaöğretim Türk dili ve edebiyatı dersi (9, 10, 11 ve 12. sınıflar) öğretim programı*. http://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=353

MEB. (2020). *EBA Yeni Dönem*. https://yegitek.meb.gov.tr/www/egitim-bilisim-aginin-eba-yeni-donem-lansmani-gerceklesti/icerik/2999

MEB. (2021). *Çevrim içi sınav*. https://www.meb.gov.tr/yuz-yuze-egitime-bir-haftalik-aradan-sonra-devam/haber/24621/tr

ÖSYM. (2018). *AYT örnek soruları*. https://www.osym.gov.tr/TR,13680/2018.html

ÖSYM. (2022). *E-YDS*. https://www.osym.gov.tr/TR,25238/2023.html

Özyalçın, K.E., & Kana, F. (2020). An evaluation on the skills of writing sub-text questions of teachers of Turkish as a foreign language. *Çukurova University Journal of Turkology Research (ÇÜTAD)*, *5*(2), 488-506.

Parshall, C.G., Spray, J.A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. Springer Science & Business Media.

Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychological Test and Assessment Modeling*, *51*(2), 123.

Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3-13.

Ryoo, J.H., Park, S., Suh, H., Choi, J., & Kwon, J. (2022). Development of a New Measure of Cognitive Ability Using Automatic Item Generation and Its Psychometric Properties. *SAGE Open*, *12*(2), 21582440221095016.

Sarı, T., & Nayır, F. (2020). Education in the pandemic period: Challenges and opportunities. *Electronic Turkish Studies*, *15*(4). http://dx.doi.org/10.7827/TurkishStudies.44335

Saygı, H. (2021). Problems encountered by classroom teachers in the covid-19 pandemic distance education process. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *7*(2), 109-129. https://doi.org/10.51948/auad.841632

Singley, M.K., & Bennett, R.E. (2002). *Item generation and beyond: Applications of schema theory to mathematics assessment*. Generating Items for Cognitive Tests: Theory and Practice., Nov, 1998, Educational Testing Service, Princeton, NJ, US; This chapter was presented at the aforementioned conference.,

Sinharay, S., & Johnson, M. (2005). Analysis of Data from an Admissions Test with Item Models. Research Report. ETS RR-05-06. *ETS Research Report Series*.

Sun, L., Liu, Y., & Luo, F. (2019). Automatic generation of number series reasoning items of high difficulty. *Frontiers in Psychology*, *10*, 884.

TDV. (2022). Türk İslam Ansiklopedisi. In https://islamansiklopedisi.org.tr/

TEDMEM. (2020). *2020 eğitim değerlendirme raporu* (TEDMEM Değerlendirme Dizisi [2020 education evaluation report], Issue.

Weber, B., Schneider, B., Fritze, J., Gille, B., Hornung, S., Kühner, T., & Maurer, K. (2003). Acceptance of computerized compared to paper-and-pencil assessment in psychiatric inpatients. *Computers in Human Behavior*, *19*(1), 81-93.

Yang, A.C., Chen, I.Y., Flanagan, B., & Ogata, H. (2021). Automatic generation of cloze items for repeated testing to improve reading comprehension. *Educational Technology & Society*, *24*(3), 147-158.

Zhu, M., Liu, O.L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668.

APPENDIX: SAMPLES OF THE GENERATEDITEMS IN TURKISH

Tek katmanlı modelle üretilen maddeler	**Çok katmanlı modelle üretilen maddeler**
1. XVII. yüzyılda yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde Sebk-i Hindî üslubunu kullanan önemli şairler arasında gösterilir. Şairin Türkçe ve Farsça divanı bulunmaktadır. Kendisinden sonra gelen pek çok şairi etkilemiştir. Divan şiiri geleneğinde kaside şairi olarak kabul edilen şair; şiirlerinde dinî tasavvufî konuların dışına çıkmış, yazarlık gücünü nesib yerine fahriye ile başlayan kasideleriyle kanıtlamıştır. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Nef'î* B) Nesîmî C) Nedim D) Şeyh Gâlip E) Nâbî	**1.** IV. Murad, Sultan İbrâhim ve IV. Mehmet arasındaki dört farklı padişah döneminde yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde dinî tasavvufî konuların dışına çıkan şair, Sebk-i Hindî üslubunu ustalıkla kullanan önemli şairler arasında gösterilir. Şairin şiirlerini topladığı bir divanı vardır. Manzum tarzda kaleme aldığı peygamberler tarihi eseri ile özdeşleşen şair, dönemindeki ve kendisinden sonra gelen pek çok şairi etkilemiştir. Divan şiirinde gazel ustası olarak adından söz ettirmiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Bâkî B) Nâbî C) Şeyh Galip D) Neşatî* E) Nedim
127. IV. Mehmet'ten III. Ahmet'e dört padişah döneminde yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde hikmet ve darbımeselleri ustalıkla kullanan önemli şairler arasında gösterilir. Şairin Türkçe ve Farsça divanı bulunmaktadır. Divan şiiri geleneğinde hikemî şairi olarak kabul edilir. Şiirlerinde dinî tasavvufî konuları ele alan şair, yazarlık gücünü Feridüddin Attar'ın bir eserinden esinlenerek kaleme aldığı mesnevisi ile ispat etmiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Neşati B) Necati Bey C) Nâbî* D) Gülşehrî E) Nef'î	1368. Şiirlerinde dinî tasavvufî konuları ele alan şair, şarkılarında yerli ve mahallî söyleyişleri ustalıkla kullanan şairler arasında yer alır. XVII. yüzyılda yaşamış divan edebiyatı şairlerinden biridir. Pek çok şair tarafından şiirlerine nazireler yazılmıştır. Yaşadığı dönemde ve divan edebiyatında tasavvufî gazel şairi olarak ün yapmıştır. Şairlik maharetini sosyal yaşamdaki gerçekliği konu alan suriyye kasidesi ile göstermiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Fuzûlî B) Nâilî* C) Nedim D) Şeyh Galip E) Nef'î
205. XVII. yüzyılda yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde zengin mecazları ve ihtişamlı ahenkleri ustalıkla kullanan önemli şairler arasında gösterilir. Şairin Türkçe ve Farsça divanı bulunmaktadır. Divan şiiri geleneğinde hiciv şairi olarak kabul edilir. Şiirlerinde dinî tasavvufî konuların dışına çıkan şair, yazarlık gücünü hiciv alanında dönemini aşan bir eseri ile ispat etmiştir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Nesîmî B) Nedim C) Şeyh Galip D) Nâbî E) Nef'î*	2054. Şiirlerinde dinî tasavvufî konuları ele alan şair, öğüt ifadelerini ustalıkla kullanan şairler arasında gösterilir. XVII. yüzyılda yaşamış divan edebiyatı şairlerinden biridir. Yazmadaki ustalığını çocuklar için yazdığı mesnevisi ile göstermiştir. Türkçe ve Farsça divanı bulunan şair, divan edebiyatında hikemî şairi olarak bilinir. **Bu parçada sözü edilen şair, aşağıdakilerden hangisidir?** A) Neşati B) Necati Bey C) Gülşehrî D) Nâbî* E) Nef'î

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 232–256

https://doi.org/10.21449/ijate.1263121

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

The development of a scale of self-efficacy to work on forgiveness in counseling

Meryem Vural-Batik [1,*], Selda Örs-Özdil [2], Necla Afyonkale-Talay [2]

[1]Ondokuz Mayis University, Faculty of Education, Department of Special Education, Türkiye
[2]Mayis University, Faculty of Education, Department of Educational Sciences, Türkiye
[3]Psikoodak Psychology and Family Counseling Center, Samsun, Türkiye

ARTICLE HISTORY

Received: Mar. 10, 2023
Revised: May 23, 2023
Accepted: Jan. 04, 2023

Keywords:
Forgiveness,
Self-efficacy,
Counselor efficacy,
Validity,
Reliability.

Abstract: It is said that working on forgiveness in psychological counseling will significantly benefit the individual, taking into account the good consequences of forgiving on the individual. This study aimed to develop a measurement tool for determining self-efficacy to work on forgiveness in counseling (SSWOFIC). The most commonly regarded forgiveness process model, Enright's Forgiveness Process Model, served as the foundation for the creation of this measurement tool. 285 counselors provided information for the Exploratory Factor Analysis (EFA) and 258 counselors provided information for the Confirmatory Factor Analysis (CFA). For the content validity of the scale, eight specialists were contacted. EFA revealed a single factor structure with 41 items that accounted for 61.7% of the overall variation. It was found that all of the SSWOFIC's items were discriminative and had a high level of factor loading value in the pertinent factor. To ascertain if the structure identified by EFA was confirmed or not, CFA was carried out. The one-factor structure was confirmed, as evidenced by the resulting model's fit indices. The computed Cronbach's Alpha and McDonald's Omega reliability coefficients were 0.99, and the Split-Half method's results for the Guttman and Spearman-Brown coefficients were 0.96. The SSWOFIC results demonstrated the validity and reliability of the scale, which consists of a single component and 41 items. The established scale will make it possible to conduct studies to ascertain the level of self-efficacy of psychological counselors with regard to this matter and to examine this feature in terms of other variables.

1. INTRODUCTION

In recent years, the idea of "forgiveness" has drawn attention in the study of psychology (Bugay, 2010). It is acknowledged as a significant issue in the field of counseling, particularly with the growth of positive psychology (Bugay & Demir, 2012; Ergüner Tekinalp & Terzi, 2012). Forgiveness is recognized as a human virtue in positive psychology (Seligman & Csikszentmihalyi, 2000). There are several definitions of forgiveness in the world of psychology, but the one provided by Enright (1996) is generally accepted. Enright (1996) described forgiveness as the voluntary renunciation of feelings like wrath, unfavorable

*CONTACT: Meryem Vural-Batik ✉ meryem.vural@omu.edu.tr Ondokuz Mayis University, Faculty of Education, Department of Special Education, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

judgment, and indifference in favor of feelings like love, generosity, and compassion for another person who has wrongfully injured them. The forgiving person renounces the right to be angry with and take revenge on the person who hurt them (Enright & Coyle, 1998). Accepting what happened, letting go of anger, and feeling good are all necessary for forgiveness (Enright, 2001).

There are various things that influence forgiveness. These variables can be classified as personal, interpersonal, and environmental variables (Hoyt & McCullough, 2005). Individual characteristics are cited as influencing factors for forgiveness. According to McCullough and Hoyt (2002), some people are more capable of forgiving others than others (Bellah et al*.,* 2003). Personality traits such as emotional maturity, common sense, resilience, empathy against aggression (Kamat et al*.,* 2006), tolerance, and extraversion (Ross et al*.,* 2004) increase the tendency to forgive. Relational factors affecting forgiveness include characteristics related to the relationship between the harmed person and the harmed person. The identity of the harming person (Bugay & Demir, 2012), the closeness of the harming person (Alpay, 2009), the nature of the relationship (Bugay & Demir, 2011), the hierarchical status of the two parties (Aquino et al*.,* 2001), the attitude of the harming person after the mistake and the willingness to apologize (Eaton et al*.,* 2007) affect forgiveness. The traits connected to the fault are the contextual elements influencing forgiveness (the situation or event that requires forgiveness). The factors that affect forgiveness include the mistake's topic, its intentionality, its intention to injure, its severity (magnitude), its result, its compensability, and its repetition (Bugay & Demir, 2011; Mullet & Girard, 2000).

Forgiving the person who harmed them has positive physical, psychological, social, and spiritual effects on individuals. Forgiveness has a positive effect on physical health and people who can forgive others have fewer symptoms of illness (Toussaint et al*.,* 2001). On the contrary, it is stated that people who have difficulty in forgiveness have more negative emotions and experience more physical problems because they have a more stressful life (Witvliet et al*.,* 2001). It is understood that the effect of forgiveness on physical health is indirect. The positive effects of forgiveness on psychological health are quite numerous. Forgiveness increases subjective well-being (Asıcı, 2018; Balcı-Çelik & Öztürk-Serter, 2017), psychological well-being (Lawler-Row & Piferi, 2006; Tse & Yip, 2009) and life satisfaction (Kaleta & Mroz, 2018; Öztörel, 2018) and protects mental health (Göztepe-Gümüş, 2015; Şentepe, 2016). On the other hand, cognitive distortions (Besim, 2017; Gündüz, 2014), anger (Aslan, 2016; Topbaşoğlu, 2016), and rumination (Oral, 2016; Özgür & Eldeleklioğlu, 2017) seem to be more intense in individuals with low forgiveness. Forgiveness helps the person who has been harmed to accept the truth, accept himself/herself and others without prejudice, and develop problem-solving skills (Şener & Çetinkaya, 2015). Thus, negative emotions such as rumination, anger, depression, and anxiety related to the unforgiven event/situation decrease; positive states such as utilizing social support, coping skills, and psychological well-being increase (Gürbüz, 2016). Another benefit of forgiveness emerges in one's social relationships. When the person who has been harmed chooses to forgive, they become psychologically stronger and can express themselves in more positive ways (Tüccar, 2015). Interpersonal relationships of people who prefer forgiveness are more regular (Toussaint & Webb, 2005), and interpersonal harmony increases (Tse & Yip, 2009). For this reason, individuals with high levels of forgiveness are more satisfied with their family and work lives and friendships (Gürbüz, 2016). Forgiveness also has spiritual effects on the individual. In most religions, the importance of being forgiving is emphasized, and people who have been harmed are advised not to hold grudges, give up the desire for revenge and forgive them. When the person who has been harmed is forgiving, he/she may feel peace, thinking that he/she is acting by religious suggestions (Van Tongeren et al*.,* 2015). Considering the physical, psychological, social, and spiritual effects of forgiveness on individuals, the importance of forgiveness is understood.

When the positive effects of forgiveness on individuals are examined, it is seen that it also serves the purposes of the psychological counseling process. Psychological counseling is a psychological assistance process provided by professionals under theories and principles to help individuals to better define themselves, realize their potential, make choices by taking decisions, especially in certain periods of their lives, to develop appropriate solutions by recognizing their problems and to solve their problems (Korkut, 2007). In the psychological counseling process, angry counselees (Murray, 2002) blame themselves or others (Menahem & Love, 2013) and/or experience problems in different areas of life due to many challenging events such as divorce, deception, neglect, abuse, abuse, fraud, criticism, obstruction, death of a loved one, illness or disasters are frequently encountered. In such cases, individuals' emotional and cognitive balances may be negatively affected and problems may occur in the social sphere. In the psychological counseling process, counselees' experiencing forgiveness enables them to restructure their impaired emotional and cognitive balances (Gordon et al., 2000) and thus contributes to their psychological healing process (Wade et al., 2005). In this respect, forgiveness is seen as a therapeutic technique that facilitates healing to achieve the counseling's therapeutic goals (Berecz, 2001; Murray, 2002; Wade et al., 2005). When forgiveness is worked on in the psychological counseling process, counselees' awareness of the event, how the event affected them, and their feelings and thoughts increases; counselees can continue their lives more healthily by choosing one of the situations of forgiveness or unforgiveness (Thompson et al., 2005). Considering the relationship between these changes in individuals' lives and the outcomes of the counseling process, it can be said that it would be important and useful to study forgiveness in counseling.

In the literature, it is observed that many forgiveness interventions have been carried out and that process-based interventions have the most effective results (Baskin & Enright, 2004). In studies conducted to improve forgiveness, it has been revealed that interventions based on the Forgiveness Process Model (Asıcı, 2018; Bugay & Demir, 2012; Ertürk, 2019; Freedman, 2018; Freedman & Knupp, 2003; Hilbert, 2015; Ji, 2013; Vural-Batık & Afyonkale-Talay, 2021) positively affect feelings, thoughts, and attitudes towards forgiveness and contribute to an increase in the tendency to forgive. In the Forgiveness Process Model, it is emphasized that it is a process for an individual to let go of negative feelings, face past experiences and painful feelings, look at the person who hurt him/her from a different perspective and choose to let go of feelings of revenge and anger. According to Enright's Forgiveness Process Model (see Table 1), which is one of the forgiveness process models, forgiveness takes place in four phases (uncovering, deciding, working, and deepening) consisting of 20 units in total (Baskin & Enright, 2004; Enright, 2001; Enright & Fitzgibbons, 2000). In the uncovering phase, there are issues related to becoming aware of the defense mechanisms used and the anger experienced, enabling them to face negative emotions, and realizing the harm of the negative emotions. This phase, in which the individual questions the pain experienced, its importance in his/her life, whether it threatens his/her life, and expresses the pain experienced, can be an emotionally painful process for the individual. In the decision phase, it is aimed to accept that the efforts made so far have not worked, to want to forgive at the cognitive level, and to decide to forgive. The individual thinks about forgiveness and develops awareness about what forgiveness is and is not in this phase. Forgiveness is not fully realized, but the individual's desire for revenge decreases. The next phase is the work phase, and there are issues related to accepting the pain, reshaping the negativity experienced, looking at it from a different perspective, and re-evaluating it. The individual begins to feel compassion for the person who harmed them, to develop a different perspective by empathizing with them, and to see them as human beings beyond the mistake they made in this phase. In the last phase which is the deepening phase, the aim is to realize the meaning of pain, and realize the freedom of forgiveness. In this phase of the forgiveness process, the individual consciously gives up emotions such as anger, resentment, and revenge that may negatively affect the health status of the person with the new perspective he/she has gained; he/she derives positive goals and meanings from the injustice and pain he/she has experienced. The individual makes sense of and internalizes the concept of forgiveness in all aspects (Enright, 1996; Enright, 2001; Enright & Fitzgibbons, 2000). The main purpose of these stages is to enable the individual to forgive the person who harmed them and thus help them to strengthen their psychological health and continue their lives with a positive perspective (Satıcı, 2016).

Table 1. *Stages of the forgiveness process.*

Phase 1: Uncovering
Unit 1: How do you avoid facing anger?
Unit 2: Have you faced your anger?
Unit 3: Are you afraid to admit you are ashamed?
Unit 4: Does your anger affect your health?
Unit 5: Do you keep thinking about the situation/offender you have been hurt by?
Unit 6: Do you compare your situation with that of the offender?
Unit 7: Does this hurt have a lasting impact on your life?
Unit 8: Has this hurt changed your worldview?
Phase 2: Decision
Unit 9: Accepting that the work done so far has not worked.
Unit 10: Willingness to begin the process of forgiveness.
Unit 11: Deciding to forgive
Phase 3: Work
Unit 12: Trying to understand.
Unit 13: Compassion and empathy work
Unit 14: Accepting pain.
Unit 15: Giving the offender a moral gift
Phase 4: Deepen
Unit 16: Recognizing the meaning of pain.
Unit 17: Recognizing the need to forgive oneself.
Unit 18: Realizing that you are not alone.
Unit 19: Realizing the meaning of your life.
Unit 20: Realizing the freedom of forgiveness

For counselors to work on forgiveness in the counseling process, it is important to know the meaning and importance of forgiveness, the factors affecting forgiveness, and the forgiveness processes (Menahem & Love, 2013). When working on forgiveness, counselors should first help counselees to understand the forgiveness process correctly. If the counselee is willing to forgive, counselors should explain what forgiveness is and is not and provide the necessary information about forgiveness (Rotter, 2001). For example, many counselees may think that forgiveness is synonymous with forgetting and reconciliation, so they may not be willing to forgive. In such cases, counselors need to be capable of providing their counselees with the right information about forgiveness (İkiz et al., 2015). Whether an individual has weak or strong self-efficacy beliefs has an impact on the individual's performance or behavior (Zimmerman, 2000). Albert Bandura (1977) defines self-efficacy, which is based on the Social Learning Theory, as the degree of belief that a person has in himself/herself about whether he/she can do a job successfully or not. Strong self-efficacy belief is a behavior that increases the motivation of an individual to cope with a problem when faced with any problem and enables him/her to make an effort (Pamukçu & Demir, 2013). Counselors' belief that they can help their counselees is an effective factor in determining their performance in the counseling process (Cormier & Nurious, 2003). Studies on counselors' self-efficacy perceptions are mostly related to counseling self-efficacy (Aktaş & Zorbaz, 2018; Akşab & Türk, 2022; Bingöl, 2018; Fırıncı-Kodaz & Vural-Batık, 2018; Pamukçu & Kağnıcı, 2013; Sarıkaya, 2017; Sarpdağ, 2019; Yayla & İkiz, 2017), special education self-efficacy (Aksoy & Diken, 2009; Arşit, 2019; Bayar & Doğan, 2021; Derin-Kılıç & Er, 2021; Vural-Batık & Fırıncı-Kodaz, 2018) and consultation self-efficacy (Bozkur & Kaya, 2021). In studies conducted with counselor candidates, the effects of the courses and supervision taken in undergraduate education on counseling self-efficacy and counseling skills were examined (Atik, 2017; Aydın, 2020; Koçyiğit- Özyiğit, 2019; Pamukçu & Kağnıcı, 2017; Şeker, 2019; Ülker-Tümlü, 2019). In the national and international literature, there is no study on self-efficacy to work on forgiveness in counseling. The most important reason for this may be that there is no measurement tool in the literature to determine the ability to work on forgiveness. There are a limited number of studies that qualitatively examined counselors' attitudes toward forgiveness (İkiz et al., 2015; Konstam et al., 2010. In a study examining the beliefs of counselor candidates about forgiveness, it was determined that counselor candidates had some knowledge about the meaning of forgiveness; however, they did not know what real forgiveness was, and they saw forgiveness not as a personality trait but as a conditional process in interpersonal relationships (İkiz et al., 2015). Konstam et al. (2010), in their study to determine the attitudes of mental health professionals towards forgiveness and their practices related to forgiveness in the counseling process, found that counselors with more positive attitudes towards forgiveness were more likely to encourage their counselees to talk about forgiveness. The lack of a measurement tool to determine the self-efficacy to work on forgiveness in counseling in the literature limits the research on this subject. This study aimed to develop a measurement tool to determine self-efficacy to practice forgiveness in counseling. Enright's Forgiveness Process Model, which is the most widely accepted forgiveness process model, was taken as the basis for the development of this measurement tool. It is thought that the development of this measurement tool will enable studies to be conducted to determine the self-efficacy to work on forgiveness in counseling. In addition, it is hoped that it will contribute to the literature by enabling the development of training programs to increase counselors' self-efficacy to work on forgiveness in counseling, determining the effectiveness of these programs, and using them in counselor training.

2. METHOD

This study aimed to develop a scale to determine the level of self-efficacy to work on forgiveness in counseling and to conduct validity and reliability analyses. In this context, the research is a scale development study. Information about the study groups and the steps followed in the development process of the scale are given below.

2.1. Study Groups

In the process of developing the self-efficacy scale for working on forgiveness in counseling, data were collected from two different study groups to conduct exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

The data were collected online from counselors working in various institutions in the 2022-2023 academic year. In scale development studies, the study group should be as heterogeneous as possible in terms of the trait to be measured (Erkuş, 2012). In this way, the scale can be examined in terms of its ability to measure individuals at different levels in terms of the measured trait. For this purpose, care was taken to ensure that the data collected through convenience sampling consisted of psychological counselors with different working years, working in different school types/institutions and at different levels. Firstly, EFA was conducted with the data obtained from psychological counselors. In the second stage, the data collected from psychological counselors were used for CFA. Data were collected from 285 people for AFA and as a result of the examination of the assumptions, some data were excluded from the analysis and analyzes were made on the data set of 258 people. Data were collected from a separate group of 258 people for DFA and as a result of the examination of the assumptions, some data were excluded from the analysis and analyzes were performed on a data set of 234 people. Information about the study groups in which the analyses were conducted is presented in Table 2.

Table 2. *Distribution of the study group according to demographic variables*

Data from the sample for EFA N_1 = 258			Data from the sample for CFA N_2 =234		
Gender	*f*	%	Gender	*f*	%
Female	216	83.7	Female	188	80.3
Male	42	16.3	Male	46	19.7
Age	*f*	%	Age	*f*	%
21-30	101	39.2	21-30	104	44.4
31-40	102	39.5	31-40	86	36.8
41+ Age	55	21.3	41+ Age	44	18.8
Seniority	*f*	%		*f*	%
1 – 5 Year	65	25.2	1 – 5 Year	83	35.5
6 – 10 Year	76	29.5	6 – 10 Year	56	23.9
11 – 15 Year	43	16.7	11 – 15 Year	39	16.7
16 – 20 Year	30	11.6	16 – 20 Year	24	10.2
21+ Year	44	17.0	21+ Year	32	13.7
Institution of Duty	*f*	%	Institution of Duty	***f***	**%**
Preschool	16	6.2	Preschool	11	4.7
Primary School	60	23.3	Primary School	39	16.7
Middle School	81	31.4	Middle School	90	38.5
High School	61	23.6	High School	48	20.5
Special Education School	11	4.3	Special Education School	5	2.1
Guidance Research Center	15	5.8	Guidance Research Center	29	12.4
Other (ASP, BİLSEM, Hospital, etc.)	14	5.4	Other (ASP, BİLSEM, Hospital, etc.)	12	5.1
Education Status	*f*	%	Education Status	***f***	***%***
Undergraduate	194	75.2	Undergraduate	165	70.5
Master's Degree	61	23.6	Master's Degree	65	27.8
PhD	3	1.2	PhD	4	1.7
Receiving Forgiveness Education	*f*	%	Receiving Forgiveness Education	***f***	***%***
Yes	12	4.7	Yes	15	6.4
No	246	95.3	No	219	93.6
Reading Resources on Forgiveness	*f*	%	Reading Resources on Forgiveness	***f***	***%***
Yes	114	44.2	Yes	97	41.5
No	144	55.8	No	137	58.5

When Table 2 is examined, it is seen that the majority of the individuals in both study groups are women and counselors with undergraduate education. It can be said that the study groups have a heterogeneous structure in terms of age, the institution of duty, and different years of employment variables. It is seen that more than 90% of the individuals in both study groups have not received any training on working with forgiveness in counseling and more than 50% of them have not read any resources on this subject.

2.2. Scale Development Process

In this scale development study to determine the self-efficacy of counselors to work on forgiveness in counseling, a literature review was conducted and no measurement tool developed for this purpose was found. To determine the items of the measurement tool, theories/models related to forgiveness in the literature (Enright, 2001; Hargrave & Sells, 1997; Worthington, 2001) and Bandura's (2006) guide for developing self-efficacy scales were examined. In line with the examinations, a pool of 44 items was created to cover these stages by taking into account the four stages in "Enright's Forgiveness Process Model" (Enright, 2001), one of the forgiveness process models, in writing the items of the measurement tool. Before the items were submitted to the expert opinion, a meeting was held by the researchers to examine whether the items were appropriate in terms of language and expression, comprehensibility, and scientific suitability, necessary corrections were made and a draft form of 41 items was created.

The 41-item draft form was e-mailed in excel format to three faculty members with a PhD in counseling and guidance counseling, two faculty members with a PhD in psychology, and three faculty members with a PhD in measurement and evaluation. While creating the item evaluation excel form for the experts, firstly, explanations about the purpose of the scale were given; then the experts were asked to evaluate the items in terms of suitability for the purpose, suitability in terms of language and expression, comprehensibility and suitability for the sub-dimension they wanted to measure. The experts were asked to give their opinions on the appropriateness of each item by using a triple rating as "appropriate", "should be improved", or "unnecessary"; they were asked to explain the items that were deemed unnecessary or should be improved and to write a suggestion for correction, if any. In line with the opinions of the experts, the content validity ratio (CVR) for each item and content validity index (CVI) for the scale were calculated using excel, taking into account Lawshe's (1975) analysis method. Table 3 shows the CVR values calculated for each item and the CVI value obtained from the whole scale.

Table 3. *Lawshe's analysis results.*

Items	CVR	Items	CVR	Items	CVR	Items	CVR
1	1.00	12	1.00	23	1.00	34	1.00
2	1.00	13	1.00	24	0.75	35	0.75
3	1.00	14	1.00	25	1.00	36	1.00
4	1.00	15	1.00	26	1.00	37	1.00
5	1.00	16	1.00	27	1.00	38	1.00
6	0.75	17	1.00	28	0.75	39	0.75
7	1.00	18	0.75	29	1.00	40	1.00
8	1.00	19	1.00	30	1.00	41	1.00
9	1.00	20	1.00	31	1.00		
10	1.00	21	1.00	32	1.00		
11	1.00	22	1.00	33	1.00		

Content Validity Index (CVI): .92

In Table 3, it is seen that the CVR values calculated based on the opinions of eight experts on the items are 0.75 and above. In this study, the CVR critical values in Ayre and Scally's (2014) study were taken into consideration for the acceptable critical value for an item to be included in the scale. In that study, it was determined that the CVR critical value was 0.75 at α=0.05 significance level for eight experts. As a result of the analysis, 41 items were decided to be included in the scale and the CVI value of these items was calculated as 0.92. In the expert opinions, adjustments were made on the relevant items in line with the suggestions of the experts who marked the items as necessary and wrote suggestions, and the final form was given to the measurement tool. For example, the meanings of terms that are unlikely to be known, such as rumination and regulation, are given in parentheses. Sentences containing more than one statement have been changed, and spelling errors in items have been corrected.

The items that were decided to be included in the scale were examined for the last time by a faculty member who is an expert in the field of Turkish teaching in terms of item comprehensibility and compliance with Turkish grammar rules. So, the 41-item draft form was made ready for the pre-test application. The counselors were asked to rate the extent to which the items in the scale reflected themselves on a scale of 1-5, and the response categories of the items were formed as "1-Not at all", "2-Little reflects", "3-Moderately reflects", "4-Very much reflects", "5-Totally reflects".

To check whether the items were comprehensible, clear, and explicit for the target group, a face-to-face pretest was conducted with 14 counselors. The counselors found the trial form mostly clear and understandable. However, four participants stated that they needed to read three items several times to understand them. These items were transformed into a more simplified structure before the actual implementation. After obtaining ethical approval from Ondokuz Mayıs University Social Sciences and Humanities Research and Publication Ethics Committee (Decision number: 2022-1080), data were first collected from 285 counselors for EFA in December 2022, and then CFA was conducted on the data collected from 258 counselors in February 2023 to test the accuracy of the construct obtained. The data were obtained through Google Forms, which provided the consent of the psychological counselors.

2.3. Data Analysis

EFA was first performed using the data gathered from the initial study group of 285 participants. To determine whether the data were appropriate for factor analysis, the assumptions of univariate and multivariate outliers, missing values, univariate and multivariate normalcy, multicollinearity, and enough sample size were examined. There were no missing values in the data set. All individuals' z scores were calculated and values between -2.88 and +2.59 were obtained in order to find outliers. No data were discovered to be outside of the -3 and +3 range (Tabachnick & Fidell, 2009). The P-P graph, skewness, and kurtosis coefficients, as well as the assumption of normality in each item score (univariate), were used to assess the results. The item scores in the data set were found to have kurtosis and skewness values between -1.00 and +1.00. This demonstrates that the item scores comply with the characteristic of normal distribution. The collinearity problem was examined by Pearson Product Moment Correlation between the items and it was determined that there was no multicollinearity problem (r<0.90).

The multivariate outliers and multivariate normalcy were investigated using the program created by Aybek (2021) in R Shiny. 27 multivariate outliers were discovered throughout the application's examination, and those data were eliminated. The data set did not meet the multivariate normality assumption, according to the Henze-Zirkler multivariate normality test results in R Shiny (p.001). The analysis was done on the 258 person data set that was downloaded from the program and was free of multivariate outliers. Also, the Bartlett Sphericity Test and Kaiser-Meyer-Olkin (KMO) tests were employed to determine whether the sample size and data were appropriate for factor analysis. The data are eligible for factor analysis because the Barlett Sphericity Test is significant and the KMO value is close to 1.

In Likert scales, if the assumption of multiple normalities is violated, the Principal Axis Factors (PAF) calculation method should be preferred among the factor extraction methods. It is stated that the PAF method is a powerful enough method for factor extraction and is widely used in many cases (Costello & Osborne, 2005; Phakiti, Costa, Plonsky, & Starfield, 2018; as cited in Şencan & Fidan, 2020). In this study, since the multivariate normality assumption was not met in the data set, the Principal Axis Factoring extraction technique was selected from the factor extraction methods. In deciding the number of factors of the scale, the parallel analysis method was taken as a basis, and the slope accumulation graph, eigenvalues, and explained variance ratios were taken into consideration. Since a single-factor structure was determined, no rotation technique was used.

To determine whether the single-factor structure of the scale determined as a result of EFA was confirmed or not, CFA was performed on the data collected from 258 participants. As in EFA, assumptions were first tested to determine the suitability of the data for factor analysis. There were no missing values in the data set. It was determined that the z scores of all individuals were between -3 and +3, the kurtosis and skewness values of the item scores were between -1.00 and +1.00, and the Pearson Product Moment Correlation calculated between the items was less than 0.90. Therefore, it can be stated that univariate outlier, normality, and multicollinearity assumptions are met in the data set. As a result of the multivariate normality and multivariate outlier analysis in R Shiny, 24 multivariate outliers were found and these data were deleted. Subsequent analyses were conducted on a data set of 234 participants. The Henze-Zirkler multivariate normality test result showed that the assumption of multivariate normality was not met in the data set (p<.001). Different methods can be used for parameter estimation of the CFA model. In the software used for CFA, unless a different method is specified, estimations are made according to the maximum likelihood (ML) method. However, to use the ML method, the data must meet the assumption of multivariate normality (Tabachnick & Fidell, 2009). Koğar and Yılmaz Koğar (2015), in their research comparing different estimation methods, stated that the Unweighted Least Squares (ULS) method gives more effective results than other methods when the multivariate normality assumption is not met. Therefore, ULS, one of the estimation methods, was used in this study.

To determine the discrimination level of the items between those who have and those who do not have the characteristics they want to measure, item discriminations were examined with corrected item-total test correlation and t-test comparisons of 27% lower and upper groups. In addition, to provide evidence for construct validity, the difference between the scores obtained from the scale by individuals who had and had not read resources on forgiveness was examined with an unrelated samples t-test. For the reliability of the scale, Cronbach's Alpha and McDonald's Omega coefficients and the coefficients obtained from the Split-half method were calculated. Jamovi 2.3.21, IBM SPSS Statistic 22, LISREL.8.51 package programs were used in data analysis and the R Shiny application was used. The significance level was set as .05 in the statistical analysis.

3. FINDINGS

In this section, EFA and CFA results of the developed scale, followed by reliability analyses and item statistics are presented respectively.

3.1. EFA Results

On the data obtained from EFA firstly, item-total test correlations and the difference between the item mean scores of the 27% lower and upper groups was examined. High item-total test correlations indicate that the items in the measurement tool measure a similar feature and that the internal consistency of the test is high. The findings obtained as a result of item analysis are given in Table 4.

Table 4. *Results of the ıtem analysis of the EFA study group.*

Item No	Corrected Item- Total Correlation	Upper and lower 27% *t* value	Item No	Corrected Item- Total Correlation	Upper and lower 27% *t* value
M1	0.55	8.35	M22	0.85	17.09
M2	0.65	10.58	M23	0.85	17.22
M3	0.65	10.99	M24	0.86	17.48
M4	0.60	10.70	M25	0.70	12.83
M5	0.77	14.77	M26	0.83	15.71
M6	0.72	13.26	M27	0.87	19.19
M7	0.78	13.10	M28	0.88	19.85
M8	0.75	14.36	M29	0.76	14.26
M9	0.81	15.58	M30	0.82	15.49
M10	0.81	16.49	M31	0.82	15.73
M11	0.78	14.44	M32	0.85	16.38
M12	0.83	17.11	M33	0.86	17.80
M13	0.88	20.75	M34	0.84	17.38
M14	0.84	17.58	M35	0.83	16.77
M15	0.82	17.60	M36	0.73	12.58
M16	0.85	17.28	M37	0.81	15.35
M17	0.84	16.65	M38	0.85	16.41
M18	0.79	14.04	M39	0.78	16.22
M19	0.79	15.35	M40	0.85	17.02
M20	0.88	18.53	M41	0.87	17.37
M21	0.85	16.85			

According to Table 4, the corrected item-total test correlation values ranged between 0.55 and 0.88. The fact that the corrected item-total correlations are greater than the threshold value of 0.30 indicates that the items adequately measure the desired construct and that the items are sufficient in terms of distinguishing the feature to be measured. High item-total test correlation indicates that the scale may be unidimensional. When the difference between the item mean scores of the 27% lower and upper groups was examined, it was releaved that the difference between the mean scores of the lower and upper groups was significant at the 0.001 level in all items. Significant *t* values for the differences between the lower and upper groups are considered evidence for the discrimination of the item (Erkuş, 2012). Accordingly, it can be said that all of the items in the scale are discriminative.

After it was examined that the corrected item-total correlations, the results of Barlett and Kaiser-Meyer-Olkin (KMO) analyses conducted to check the suitability of the data for factor analysis are given in Table 5 below.

Table 5. *Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphericity test results.*

Statistic		Value
Kaiser-Meyer-Olkin (KMO)		.97
Bartlett's sphericity	χ^2	12746
	df	820
	p	<.001

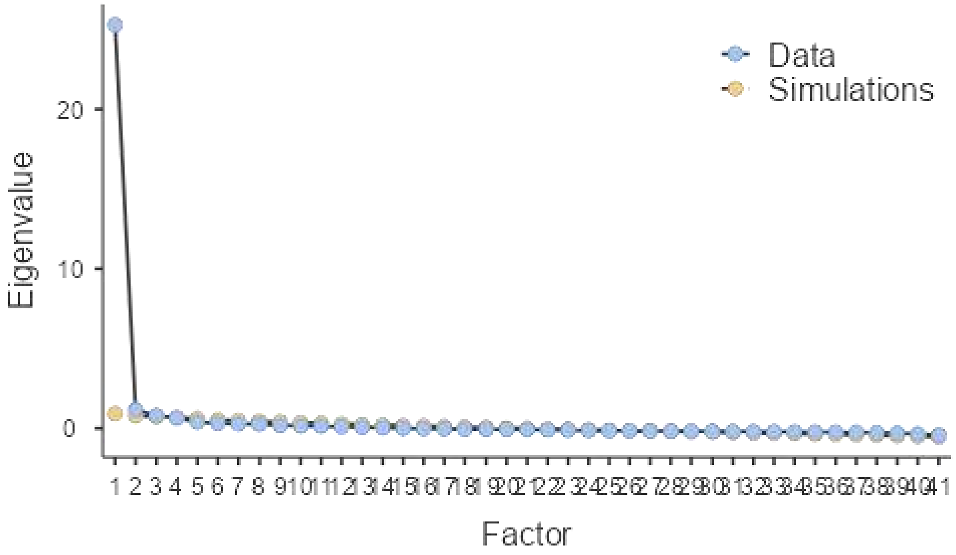
When the suitability of the data for EFA was examined, it was determined that the KMO value was 0.97 and the Barlett Sphericity test result (χ^2= 12746, *df*=820, *p*<.001) was significant. Thus, the data were found to be suitable for factor analysis. As a result of the EFA conducted without limiting the dimension to explore the factor structure of the scale, it was seen that there were two factors with eigenvalues above 1. The factor eigenvalues obtained as a result of the analysis and the explained variance rates are given in Table 6.

Table 6. *Factor eigenvalues and explained variance.*

Factor	SS Loadings	% of Variance
1	26.73	65.19
2	1.24	3.03

The slope accumulation graph and explained variance ratios indicate that the scale exhibits a single-factor structure. The slope accumulation graph obtained according to the parallel analysis method is given in Figure 1. The parallel analysis method also reveals that the scale shows a single-factor structure.

Figure 1. *Scree plot.*



As a result of the EFA, which was limited to a single factor, the variance explained was 65.2% of the total variance. After it was decided that the scale showed a single-factor structure, the factor loadings of the items were analyzed. Table 7 shows the factor loadings of the 41 items in the scale.

Table 7 shows that the factor loadings of the items vary between .515-.871. Factor loadings of .60 and above are considered to be high (Kline, 2005). Therefore, no item was removed from the scale.

Table 7. *Factor loadings of the items on the dimension.*

Items		Factor Loadings
M1	I can define the concept of forgiveness within the framework of the literature.	0.515
M2	I can explain to the counselee the difference of forgiveness from concepts such as forgetting, excusing, and turning a blind eye.	0.669
M3	I can enable the counselee to open up about unforgiven experiences.	0.649
M4	I can explain the stages of the forgiveness process within the framework of the literature.	0.579
M5	I can explain the stages of the forgiveness process within the framework of the literature.	0.746
M6	I can make the counselee feel the desire to try forgiveness.	0.729
M7	I can enable the counselee to make a self-assessment of the level of forgiveness at the beginning of counseling.	0.769
M8	I can recognize the defense mechanisms used by the counselee in case of unforgiveness.	0.761
M9	I can work with the counselee about the defense mechanisms used in case of unforgiveness.	0.804
M10	I can help the counselee to realize his/her feelings about unforgiven experiences.	0.805
M11	I can explain to the counselee the possible negative effects of anger related to unforgiven experiences on health.	0.794
M12	I can ensure that anger related to unforgiven experiences is revealed in the therapeutic process.	0.817
M13	I can make the counselee aware of the effects of unforgiveness in his/her life.	0.836
M14	I can bring awareness to the counselee about rumination (repetitive negative internal conversations) related to unforgiven experiences.	0.826
M15	I can help the counselee cope with rumination about unforgiven experiences.	0.805
M16	I can make the counselee realize the dysfunctional thoughts about comparing his/her situation with the person he/she has not forgiven.	0.836
M17	I can make the counselee aware of how unforgiven experiences affect his/her philosophy of life.	0.801
M18	I can realize that the counselee cannot regulate (regulate) his/her emotions.	0.763
M19	I can use various interventions for the counselee to achieve emotional regulation.	0.762
M20	I can help the counselee to make a self-assessment of their readiness to decide to forgive.	0.871
M21	I can help the counselee to recognize their dysfunctional strategies for the experiences they cannot forgive.	0.826
M22	I can encourage the counselee to want to start the forgiveness process.	0.842
M23	I can work with the counselee to decide to forgive.	0.819
M24	I can make the counselee aware of his/her thoughts, feelings, and behaviors related to the experiences he/she cannot forgive.	0.831
M25	I can explain the relationship between thoughts, feelings, and behaviors related to forgiveness to the counselee within the framework of the Cognitive-Behavioral Therapy approach.	0.706

M26	I can bring awareness to the counselee's automatic thoughts about the experiences that the counselee cannot forgive.	0.832
M27	I can enable the counselee to develop a perspective that will facilitate forgiveness regarding unforgivable experiences.	0.866
M28	I can work with the counselee to generate new/alternative thoughts that will facilitate the forgiveness process.	0.860
M29	I can define the concept of compassion within the framework of literature.	0.704
M30	I can work with the counselee to feel compassion for the person they cannot forgive.	0.822
M31	I can work with the counselee to empathize with the person they cannot forgive.	0.827
M32	I can work with the counselee to accept the pain related to unforgiveness.	0.841
M33	I can work with the counselee to discover the meaning of the pain felt related to the experiences they cannot forgive.	0.852
M34	I can help the counselee discover that he/she is not the only one who has experienced situations that require forgiveness.	0.827
M35	I can help the counselee to realize that he/she also needs forgiveness.	0.822
M36	I can define the concept of reconciliation within the framework of the literature.	0.648
M37	I can explain the reconciliation process to the counselee.	0.751
M38	I can enable the counselee to express forgiveness clearly.	0.829
M39	I can enable the counselee to express forgiveness indirectly such as imagination and artistic activities.	0.759
M40	I can make the counselee aware of the positive emotions felt as a result of forgiveness.	0.827
M41	I can enable the counselee to self-evaluate the results of the forgiveness experience.	0.813

3.2. CFA Results

The findings of the CFA conducted to confirm the structure of the single-factor scale that emerged as a result of EFA are presented in Figure 2 and Table 8 below. Accordingly, the standardized factor loadings of the items in the relevant factor and the error variances of the items are shown. After obtaining the path diagram, the significance of the standardized factor loading values of the items under the factors should be checked first. It was observed that the *t* values of 11 items (M13, M16, M24, M27, M28, M32, M33, M35, M38, M40, M41) were less than 1.96, that is, they were not significant at a .05 significance level. Although it is recommended to exclude items with insignificant *t* values from the analysis within the framework of the structural equation, it is stated that the error variances and factor loading values of the items should be checked before making this decision (Çokluk et al., 2021). When the factor loading values obtained as a result of the analysis are examined, it is observed that the standardized factor loading values of all items are between 0.58 and 0.90. An error variance above 0.90 weakens the fit of the model to the data and it is stated that observed variables with very high error variance can be removed from the model (Çokluk et al., 2021; Kline, 2011). It is seen that the error variances of all items are considerably smaller than 0.90. Since 41 items in the scale had high factor loading values both as a result of EFA and CFA and the error variances were low as a result of CFA, it was decided that no item should be excluded from the analysis.

Figure 2. *Factor loadings of the items revealed by CFA results.*

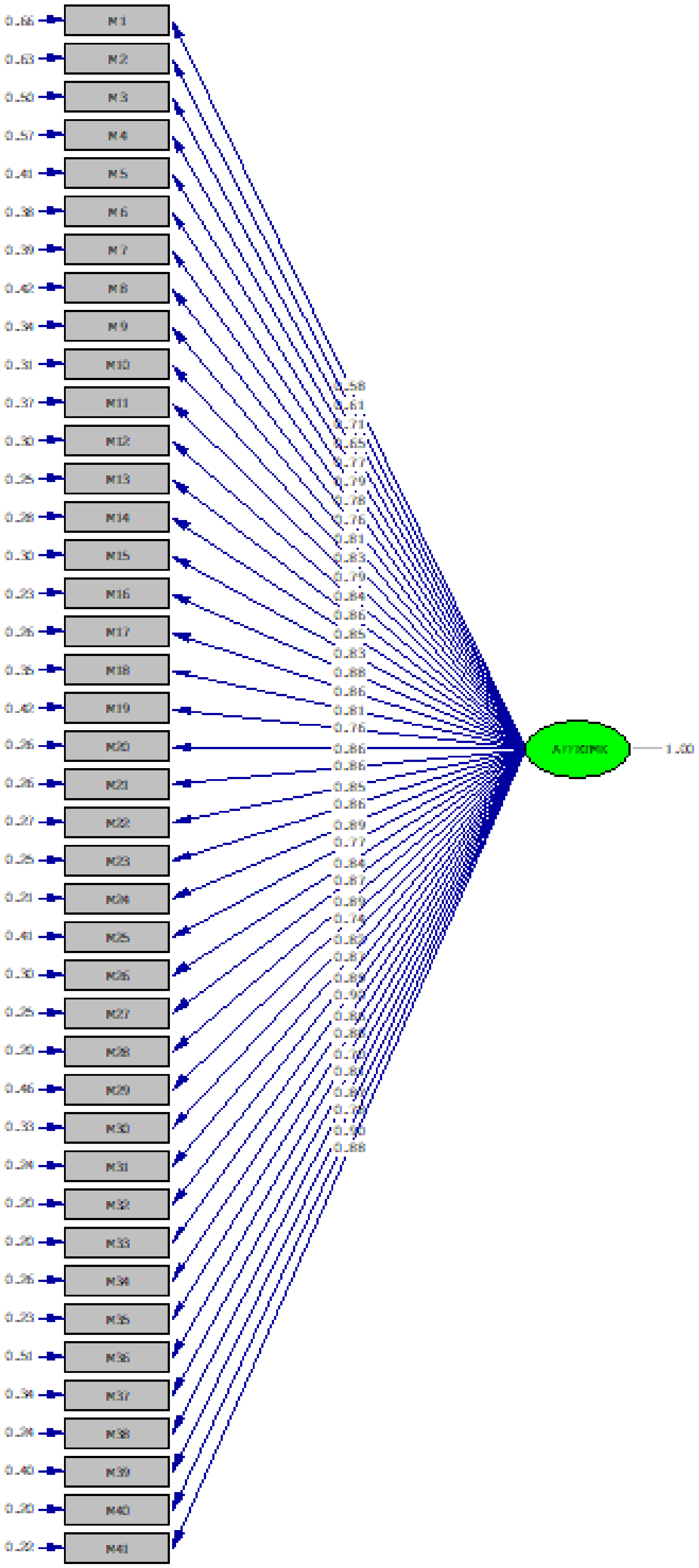


Table 8. *Standardized Factor loadings and SH of the items.*

Item No	Standardized Factor Loadings	SH	Item No	Standardized Factor Loadings	SH
M1	0.58	0.66	M22	0.85	0.27
M2	0.61	0.63	M23	0.86	0.25
M3	0.71	0.50	M24	0.89	0.21
M4	0.65	0.57	M25	0.77	0.41
M5	0.77	0.41	M26	0.84	0.30
M6	0.79	0.38	M27	0.87	0.25
M7	0.78	0.39	M28	0.89	0.20
M8	0.76	0.42	M29	0.74	0.46
M9	0.81	0.34	M30	0.82	0.33
M10	0.83	0.31	M31	0.87	0.24
M11	0.79	0.37	M32	0.89	0.20
M12	0.84	0.30	M33	0.90	0.20
M13	0.86	0.25	M34	0.86	0.26
M14	0.85	0.28	M35	0.88	0.23
M15	0.83	0.30	M36	0.70	0.51
M16	0.88	0.23	M37	0.81	0.34
M17	0.86	0.26	M38	0.87	0.24
M18	0.81	0.35	M39	0.78	0.40
M19	0.76	0.42	M40	0.90	0.20
M20	0.86	0.26	M41	0.88	0.22
M21	0.86	0.26			

After examining the coefficients obtained as a result of CFA, the goodness-of-fit indices produced to evaluate the model as a whole were examined. Goodness-of-fit index values for model-data fit are given in Table 9.

Table 9. *The goodness of Fit Index Values for the Model.*

χ^2	*sd*	χ^2/sd	AGFI	GFI	CFI	NFI	NNFI	RMSEA	SRMR
3242.55	779	4.16	0.99	1.00	1.00	1.00	1.00	0.12	0.045

When Table 9 is examined, it is seen that the χ^2 value is 3242.55 and the value obtained by dividing the χ^2 value by the degrees of freedom is 4.16. A value of 5 or below indicates an acceptable fit (Kline, 2011). In addition, it is suggested that the evaluation of the model in confirmatory factor analysis should not be based on a single value (especially χ^2) but on multiple fit indices. Accordingly, when the fit indices of the scale are examined, AGFI, GFI, CFI, NFI, and NNFI values above 0.95 are indicative of an excellent fit. RMSEA and SRMR values between 0.05 and 0.08 indicate good fit, and values between 0.80 and 0.10 indicate acceptable fit. It is seen that the RMSEA value obtained is close to 0.10 acceptable fit and the SRMR value is below 0.05. When all the analysis results and goodness of fit values obtained with CFA are evaluated together, it may be said that the one-factor structure of the scale consisting of 41 items generally fits the data well and the scale structure is confirmed.

3.3. Item Analysis and Validity Analysis Based on Group Differences

To determine the discrimination levels of the items in the SSWOFIC, the total scores obtained from the scale were determined and 27% lower-upper group (Nlower: 64 and Nupper: 62) comparisons were made. Pearson Product Moment Correlation Coefficient was used to calculate the corrected item sub-dimension total correlation, and an unrelated sample *t*-test was used for 27% lower-upper group comparisons. The findings obtained as a result of item analysis are given in Table 10.

According to Table 10, the corrected item-total test correlation values ranged between 0.59 and 0.89. The fact that the corrected item-total correlations are greater than the threshold value of 0.30 indicates that the items adequately measure the desired construct and that the items are sufficient in terms of distinguishing the feature to be measured. When the difference between the item mean scores of the 27% lower and upper groups was examined, it was releaved that the difference between the mean scores of the lower and upper groups was significant at the 0.001 level in all items. Significant *t* values for the differences between the lower and upper groups are considered evidence for the discrimination of the item (Erkuş, 2012). Accordingly, it can be said that all of the items in the scale are discriminative.

Table 10. *Results of the item analysis of the SSWOFIC.*

Item No	Corrected Item- Total Correlation	Upper and lower 27% *t* value	Item No	Corrected Item- Total Correlation	Upper and lower 27% *t* value
M1	0.59	8.94	M22	0.85	17.31
M2	0.62	9.99	M23	0.86	19.06
M3	0.71	11.62	M24	0.88	19.21
M4	0.65	10.72	M25	0.76	16.00
M5	0.77	15.56	M26	0.83	17.60
M6	0.79	14.56	M27	0.86	18.04
M7	0.78	15.05	M28	0.89	20.21
M8	0.76	13.76	M29	0.74	13.92
M9	0.81	16.29	M30	0.81	16.98
M10	0.83	16.63	M31	0.87	19.33
M11	0.79	14.34	M32	0.89	20.58
M12	0.83	16.87	M33	0.89	20.91
M13	0.86	16.42	M34	0.86	18.91
M14	0.84	19.93	M35	0.87	18.41
M15	0.83	18.49	M36	0.70	12.35
M16	0.87	24.99	M37	0.81	17.30
M17	0.86	20.78	M38	0.86	18.27
M18	0.80	16.24	M39	0.77	15.17
M19	0.76	14.03	M40	0.89	19.18
M20	0.85	19.25	M41	0.88	20.47
M21	0.86	18.55			

Unrelated samples t-test was used to determine whether the self-efficacy levels of psychological counselors to study forgiveness differed according to whether they read a source about forgiveness. The findings obtained as a result of the analysis are presented in Table 11.

Table 11. *The independent t-test results.*

Group	N	Mean	*SD*	*t*	*p*
Reading resources on forgiveness	97	136.45	26.14	3.18	0.002
Not reading resources on forgiveness	137	124.28	30.62		

Table 11 shows that the self-efficacy levels of teachers who read resources on forgiveness were statistically higher than those who did not (p<0.05). Considering that this finding is expected, it can be said that the scale accurately measures the related construct.

3.4. Reliability Analysis Results

Cronbach's alpha and McDonald's omega coefficients were calculated for the reliability of the SSWOFIC. The Cronbach's alpha and McDonald's omega coefficients of the single-factor 41-item scale were 0.99. After the reliability coefficients of the whole scale were calculated, the internal consistency reliability of the scale was also calculated with the Split-half method. The Cronbach's alpha coefficient of 21 items in the first half was 0.97 and the Cronbach's alpha coefficient of 20 items in the second half was 0.98. It can be said that the internal consistency coefficient values of the two groups formed with the Split-Half method are close to each other and very good. With this method, Guttman and Spearman-Brown coefficients were found to be 0.96. These findings show that the scale as a whole has a high level of reliability.

4. DISCUSSION and CONCLUSION

It is clear that forgiveness serves the purposes of counseling when taking into account its beneficial impacts on the individual. In order to accomplish the objectives of counseling, forgiveness is viewed as a therapeutic technique that promotes healing (Berecz, 2001; Murray, 2002; Wade et al., 2005). Counselors must understand what forgiveness is and how it works in order to work on it during counseling sessions (Menahem & Love, 2013). The performance of counselors in the counseling process is significantly influenced by their confidence in their ability to assist their counselees (Cormier & Nurious, 2003). Yet, one significant gap in the literature was the absence of a measurement method to assess one's capacity to work on forgiveness in counseling. A useful scale with good validity and reliability was sought to measure self-efficacy to practice forgiveness in counseling in light of this deficit.

The majority of scales employed in studies on the self-efficacy of psychological counselors relate to counseling skills (Aktaş & Zorbaz, 2018; Akşab & Türk, 2022; Bingöl, 2018; Fırıncı-Kodaz & Vural-Batık, 2018; Pamukçu & Kağnıcı, 2013; Sarıkaya, 2017; Sarpdağ, 2019; Yayla & İkiz, 2017). Also, there are scales to measure counselors' self-efficacy in consultations and special education (Aksoy & Diken, 2009; Arşit, 2019; Bayar & Doğan, 2021; Derin-Kılıç & Er, 2021; Vural-Batık & Fırıncı-Kodaz, 2018). (Bozkur & Kaya, 2021). The statements on these scales that refer to counseling abilities were a key source for the scale created for the current investigation.

The construction of the SSWOFIC took into account both Bandura's (2006) self-efficacy scale development guide and Enright's Forgiveness Process Model (Enright, 2001). An item pool with comments regarding approaches and counseling abilities to assist the counselee in the four-phase forgiveness process was created. Eight experts reviewed the 44 items for content validity. According to the experts' suggestions, a 41-item draft form was created.

Data were collected from two different study groups for the validity and reliability analysis of the scale. As a result of the EFA conducted in the first study group, a single-factor structure with 41 items was obtained. This single-factor structure explained 65.2% of the total variance. The factor loadings of all items were high, so no item was removed from the scale. CFA was conducted in the second study group to determine whether this structure was confirmed or not.

The fit indices obtained as a result of the analysis were found to be high. To determine the discrimination of the items in the scale, 27% lower and upper groups were analyzed. As a result of the 27% lower and upper groups analysis of all items, the t value was found to be significant and the discrimination values were high. The corrected item-total test correlation values of the items indicate that the scale has high item discrimination and high validity. To provide evidence for the construct validity of the scale, the difference between the scores of the groups who read and did not read resources on forgiveness was examined and a statistically significant difference was found. To determine the reliability of the scale, reliability coefficients were calculated using Cronbach's Alpha, McDonald's Omega and Split-Half methods. It was determined that the reliability of the scale was high. The final version of the developed scale is given in the Appendix. As a result, a scale with high validity and reliability was introduced to the literature.

The SSWOFIC, whose validity and reliability have been established, can be used by practitioners and researchers for a variety of applications. To find out if a counselee has the self-efficacy to work on forgiveness in counseling, research can be done. With the use of a scale, studies can be used to identify counselors who have a low opinion of their own efficacy in working with forgiveness. Training programs can then be developed to raise this perspective, and the success of these programs can be assessed. The fact that only psychological counselors were included in the study is one of its shortcomings. If this scale, which was created by gathering information from psychological counselors, is validated for psychologists, a study can be done to find out. Studies on the scale's validity and reliability can also be done on psychologists. In addition, the small number of male participants in the study group is one of the limitations of this study. For this reason, it may be recommended to carry out validity and reliability studies on different study groups of the research and to perform multiple group analyzes.

Acknowledgments

Thank to Atakum Guidance and Research Center for the support given by during the data collection process.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ondokuz Mayıs University Social Sciences and Humanities Research and Publication Ethics Committee - Document numbered: 2022-1080.

Authorship Contribution Statement

Meryem Vural-Batik: Conception, Design, Materials, Data Collection and Processing, Interpretation, Literature Review, Writing- original draft, Critical Review. **Selda Örs-Özdil**: Conception, Design, Supervision, Materials, Methodology, Analysis, Writing- original draft, Critical Review. **Necla Afyonkale-Talay**: Conception, Design, Supervision, Materials, Data Collection, Literature Review, Writing- original draft.

Orcid

Meryem Vural-Batik https://orcid.org/0000-0002-7836-7289
Selda Örs-Özdil https://orcid.org/0000-0002-7134-5896
Necla Afyonkale-Talay https://orcid.org/0000-0002-9835-2340

REFERENCES

Aksoy, V., & Diken, İ.H. (2009). Rehber Öğretmen Özel Eğitim Öz Yeterlik Ölçeği: Geçerlik ve güvenirlik çalışması [School Counselors' Self-Efficacy Scale regarding Special

Education (SCSSSE): Validity and reliability results]. *Ankara University Faculty of Educational Sciences Journal of Special Education, 10*(1), 29-37. https://doi.org/10.1501/Ozlegt_0000000131

Akşab, G., & Türk, F. (2022). Psikolojik danışmanların psikolojik danışma öz-yeterlik algılarına ilişkin bir derleme çalışması [A review on counseling self-efficacy perceptions of psychological counselors]. *The Journal of School Counseling, 5*(1), 1-40. https://dergipark.org.tr/en/download/article-file/1685048

Aktaş, E.F., & Zorbaz, S.D. (2018). Okul psikolojik danışmanlarının çocukla psikolojik danışma yeterliklerine ilişkin görüşleri [School counselor's opinions about their competence of child counseling], *Inonu University Journal of the Faculty of Education, 19*(1), 245-256. https://doi.org/10.17679/inuefd.416003

Alpay, A. (2009). *Yakın ilişkilerde bağışlama: Bağışlamanın bağlanma, benlik saygısı, empati ve kıskançlık değişkenleri yönünden incelenmesi [Forgiveness in close relationship: The investigatement of forgiveness in terms of attachment, sels-esteem, empathy and romantic jealousy]* (250105) [Master thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Aquino, K., Tripp, T.M., & Bies, R.J. (2001). How employees respond to personal offense: The effect of blame attribution, victim status, and offender status on revenge and reconciliation in the workplace. *Journal of Applied Psychology, 86,* 52–59. https://doi.org/10.1037/0021-9010.86.1.52

Arşit, M.H. (2019). *Özel eğitim okulları ile rehberlik ve araştırma merkezlerinde görev yapan rehberlik öğretmenlerinin özel eğitimde rehberlik ve psikolojik danışma hizmetlerine ilişkin özyeterliliklerinin incelenmesi [An examination of self-efficacy of school counsellors working in special education schools with counseling and research centers on guidance and psychological counseling services in special education]* (540386). [Master thesis, Biruni University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Asıcı, E. (2018). *Affetme odaklı grup rehberliğinin ergenlerin saldırganlık ve öznel iyi oluşları üzerindeki etkisi [The effect of forgiveness focused group guidance on aggression and subjective well-being of adolescents]* [511610). [Doctoral dissertation, Dokuz Eylül University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Aslan, G.A. (2016). *Üniversite öğrencilerinin affetme davranışları, yaşam doyumları ve sürekli öfke düzeyleri arasındaki ilişkilerin incelenmesi [The relationship between the forgiveness behaviour, life satisfaction and trait-anger of the university students]* [450116). [Master thesis, Gazi University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Atik, Z. (2017). *Psikolojik danışman adaylarının bireyle psikolojik danışma uygulaması ve süpervizyonuna ilişkin değerlendirmeleri [Counselor candidates' evaluation of individual counseling practicum and supervision]* (470017). [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Aybek, E.C. (2021). Data preparation for factor analysis. https://shiny.eptlab.com/dp2fa/

Aydın, F. (2022). *Algılanan süpervizör tarzları ile süpervizyon doyumu arasındaki ilişkide psikolojik danışma öz-yeterliğinin aracı rolü [The mediator role of counseling self-efficacy in the relationship between perceived supervisory styles and satisfaction with supervision]* (711068). [Doctoral dissertation, Trabzon University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Ayre, C., & Scally, A.J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development, 47*(1), 79-86. https://doi.org/10.1177/0748175613513808

Balcı-Çelik, S., & Öztürk-Serter, G. (2017). Üniversite öğrencilerinin romantik ilişkilerinde affetmenin öznel iyi oluşları üzerindeki rolü [The role of forgiveness on subjective well-being of university students in their romantic relationships]. *International Journal of Human Sciences, 14*(4), 3990-4001. https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/4874/2389

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares, & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age Publishing.

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

Baskin, T.W., & Enright, R.D. (2004). Intervention studies on forgiveness: A meta analysis. *Journal of Counseling and Development, 82*, 79-90. https://doi.org/10.1002/j.1556-6678.2004.tb00288.x

Bayar, Ö., & Doğan, T. (2021). Okul psikolojik danışmanı adaylarının özel eğitim öz-yeterlik algıları ve yeterlilik düzeyleri: Bir karma yöntem çalışması [School counselor candidates' special education self-efficacy perceptions and levels of proficiency: A mixed-method study]. *Ankara University Faculty of Educational Sciences Journal of Special Education, 22*(2), 369-394. https://doi.org/10.21565/ozelegitimdergisi.695682

Bellah, C.G., Bellah, L.D., & Johnson, J.L. (2003). A look at dispositional vengefulness from the three and five-factor models of personality. *Individual Differences Research, 1*, 6- 16.

Berecz, J.M. (2001). All that glitters is not gold: Bad forgiveness in counseling and preaching. *Pastoral Psychology, 49*(4), 253-275. https://digitalcommons.andrews.edu/pubs/2176

Besim, G. (2017). *Üniversite öğrencilerinde affetme, bitirilmemiş işler ve öfke [Forgiveness, unfinished business and anger among university students]* (483377). [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Bingöl, T.Y. (2018). Okul psikolojik danışman adaylarının genel özyeterlik ve özel eğitimde rehberlik ve psikolojik danışmanlığa ilişkin öz-yeterlik inançları arasındaki ilişkinin incelenmesi [Investigation of the relationship between general self-efficacy of school psychological counselor candidates and self-efficacy beliefs of that regarding special education]. *OPUS International Journal of Society Researches, 8*(15), 40-40. https://doi.org/10.26466/opus.444225

Bozkur, B., & Kaya, A. (2021). Adaptation of the Consultation Self-Efficacy Scale into Turkish and investigation of the consultation self-efficacy of school counselors. *Pegem Journal of Education and Instruction, 11*(1), 49-96. https://files.eric.ed.gov/fulltext/EJ1286904.pdf

Bugay, A. (2010). *Kendini affetmeyi yordayan sosyobilişsel, duygusal ve davranışsal faktörlerin incelenmesi [Kendini affetmeyi yordayan sosyo-bilişsel, duygusal, davranışsal faktörlerin incelenmesi]* (277716). [Doctoral dissertation, Orta Doğu Teknik University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Bugay, A., & Demir, A. (2011, 3 Sep). *Yaşam doyumu ile affetme arasındaki ilişkide ruminasyon eğiliminin aracı rolü.* XI. National Psychological Counseling and Guidance Congress, Turkey.

Bugay, A., & Demir, A. (2012). Affetme arttırılabilir mi?: Affetmeyi geliştirme grubu [Can be forgiveness increased?: Forgiveness Enrichment Group]. *Turkish Psychological Counseling and Guidance Journal, 4*(37), 96-106.

Cormier, S., & Nurious, P.S. (2003). *Interviewing and change strategies for helpers: Fundamental skills and cognitive behavioral interventions* (5th ed.). Pacific Grove: Thomson-Brooks/Cole.

Çokluk, O., Şekercioğlu, G., & Büyüköztürk, Ş. (2021). *Sosyal bilimler için çok değişkenli istatistik, SPSS ve LISREL uygulamları* (6th ed.) [Multivariate statistics SPSS and LISREL applications for social sciences]. PegemA.

Derin- Kılıç, A., & Er, E. (2021). Rehber Öğretmenleri ile rehber öğretmen adaylarının özel eğitimde rehberlik ve psikolojik danışma hizmetlerine ilişkin öz-yeterlik algı düzeylerinin incelenmesi [An investigation of the self-efficacy perception levels of school counselors and school counselor candidates regarding guidance and psychological counselling services in special education]. *Akdeniz Journal of Education, 4*(2), 26-44. https://dergipark.org.tr/en/download/article-file/1738911

Eaton, J., Struthers, C.W., Shomrony, A., & Santelli, G. (2007). When apologies fail: The moderating effect of implicit and explicit self-esteem on apology and forgiveness. *Self and Identity*, *6*(2-3), 209-222. https://doi.org/10.1080/15298860601118819

Enright, R.D. (1996). Counseling within the forgiveness triad: On forgiving, receiving forgiveness, and self forgiveness. *Counseling and Values*, *40*(2), 107-126. https://doi.org/10.1002/j.2161-007X.1996.tb00844.x

Enright, R.D. (2001). *Forgiveness is a choice: A step by-step process for resolving anger and restoring hope.* Washington, DC: American Psychological Association.

Enright, R.D., & Coyle, C.T. (1998). Researching the process model of forgiveness within psychological interventions. In E. L. Worthington Jr. (Eds), *Dimensions of forgiveness: Psychological research and theological perspectives* (pp. 139-161). Philadelphia: Templeton Foundation Press.

Enright, R.D., & Fitzgibbons, R.P. (2000). *Helping clients forgive: An empirical guide for resolving anger and restoring hope.* Washington, DC: American Psychological Association.

Ergüner-Tekinalp, B., & Terzi, Ş. (2012). Terapötik bir araç olarak bağışlama: İyileştirici bir etken olarak bağışlama olgusunun psikolojik danışma sürecinde kullanımı [Forgiveness as a therapeutic tool: Using forgiveness for healing in the counseling process]. *Education and Science, 37*(166), 14-24. http://213.14.10.181/index.php/EB/article/view/405/435

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme I: Temel kavramlar ve işlemler.* [Measurement and scale development in psychology I: Basic concepts and operations.]. Pegem Akademi Yayınları.

Ertürk, K. (2019). *Lise öğrencilerinde affetme becerisi geliştirmeye yönelik psiko-eğitim programının affetme ve yaşam doyumu üzerindeki etkisi [The effect of psychoeducation program to improve forgiveness skill on forgiveness and life satisfaction of high school students]* (555976). [Master thesis, Sakarya University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Fırıncı Kodaz, A., & Vural Batık, M. (2018). Kurum deneyiminin psikolojik danışman adaylarının okul psikolojik danışmanı öz yeterlik algılarına etkisi [The Effect of Field Practice on Psychological Counselor Candidates' Counseling Self-Efficacy Levels]. *OPUS – International Journal of Society Researches, 8*(15), 902-929. https://doi.org/10.26466/opus.425811

Freedman, S. (2018). Forgiveness as an educational goal with at-risk adolescents. *Journal of Moral Education*, 47(4), 415-431. https://doi.org/10.1080/03057240.2017.1399869

Freedman, S., & Knupp, A. (2003). The impact of forgiveness on adolescent adjustment to parental divorce. *Journal of Divorce & Remarriage, 39*(1-2), 135-165. https://doi.org/10.1300/J087v39n01_08

Gordon, K.C., Baucom, D.H., & Snyder, D.K. (2000). The use of forgiveness in marital therapy. In M. McCullough, K.I. Pargament ve C.E. Thoresen (Eds.), *Forgiveness: Theory, research, and practice* (pp. 203-227). Guilford Press.

Göztepe-Gümüş, I. (2015). *Evli bireylerde bağışlama, tekrarlayıcı düşünme düzeyi ile ruh sağlığı ve evlilik uyumu arasındaki ilişkiler[The relations among forgiveness, repetitive thinking, marital adjustment, and mental health]* (393366). [Master thesis, Ankara University]. Council of Higher EducationThesis Center. https://tez.yok.gov.tr

Gündüz, Ö. (2014). *Üniversite öğrencilerinde affetmeyi yordayan değişkenlerin belirlenmesi [Determining the variables that predict forgiveness among university students]* (370311). [Master thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Gürbüz, E. (2016). *Evlilik içinde aldatılan bireylerin affetmelerini yordamada bağlanma stilleri ve psikolojik sağlamlığın rolü [The role of attachment styles and resilience as predictor of forgiveness among individuals who were betrayed in marriage]* (437524). [Master thesis, Bahçeşehir University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Hargrave, T.D., & Sells, J.N. (1997). The development of a Forgiveness Scale. *Journal of Marital and Family Therapy. 23*(1), 41-62. https://doi.org/10.1111/j.1752-0606.1997.tb00230.x

Harris, A.H.S., Luskin, F., Norman, S.B., Standartd, S. Bruning, J., & Hilbert, H.K.E. (2015). The impact and evaluation of forgiveness education with early adolescents. [Master thesis, University of Northern Iowa]. https://scholarworks.uni.edu/cgi/viewcontent.cgi?article=1188&context=hpt

Hoyt, W.T., & McCullough, M.E. (2005). Issues in the multimodal measurement of forgiveness. In E. L. Worthington Jr. (Ed.), In *The handbook of forgiveness* (pp. 109-123). Taylor & Francis Group.

İkiz, F.E., Mete-Otlu, B., & Asıcı, E. (2015). Beliefs of counselor trainees about forgiveness. *Educational Sciences: Theory & Practice*, *15*(2), 463-479. https://doi.org/10.12738/estp.2015.2.2205

Ji, M., Tao, L., & Zhu, T. (2016). Piloting forgiveness education: A comparison of the impact of two brief forgiveness education programmes among Chinese college students. *Asia-Pacific Education Researcher, 25*(3), 483–492. https://doi.org/10.1007/s40299-016-0273-6

Kaleta, K., & Mroz, J. (2018). Forgiveness and life satisfaction across different age groups in adults. *Personality and Individual Differences, 120,* 17-23. https://doi.org/10.1016/j.paid.2017.08.008

Kamat, V.L., Jones, W.H., & Row, K.L. (2006). Assessing forgiveness as a dimension of personality. *Individual Differences Research, 4*, 322 330. https://doi.org/10.1016/j.paid.2017.08.008

Kline, R.B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). The Guilford Press.

Koçyiğit-Özyiğit, M. (2019). *Bireyle psikolojik danışma uygulaması dersinde grup süpervizyonu sürecinin incelenmesi: Bir durum çalışması [An investigation of group supervision process of "Individual counseling practice course": A case study]* (545739). [Doctoral dissertation, Ege University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Koğar, H., & Yılmaz-Koğar, E. (2015. Comparison of different estimation methods for categorical and ordinal data in confirmatory factor analysis. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 351-364. https://dx.doi.org/10.21031/epod.94857

Konstam, V., Marx, F., Schurer, J., Harrington, A., Lombardo, N.E., & Deveney, S. (2010). Forgiving: What mental health counselors are telling us. *Journal of Mental Health Counseling*, *22*(*3*), 253–267.

Korkut, F. (2007). *Okul temelli önleyici rehberlik ve psikolojik danışma [School-based preventive guidance and psychological counseling]* (2. bs.). Anı Yayıncılık

Lawler-Row, K.A., & Piferi, R.L. (2006). The forgiving personality: Describing a life well lived?. *Personality and Individual Differences, 41,* 1009-1020. https://doi.org/10.1016/j.paid.2006.04.007

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel psychology, 28*(4), 563-575.

McCullough, M.E., & Hoyt, W.T. (2002). Transgression-related motivational dispositions: Personality substrates of forgiveness and their links to the Big Five. *Personality and Social Psychology Bulletin*, *28*, 1556-1573. https://doi.org/10.1177/014616702237583

Menahem, S., & Love, M. (2013). Forgiveness in psychotherapy: The key to healing. *Journal of Clinical Psychology: In Session, 69*(8), 829-835. https://doi.org/10.1002/jclp.22018

Mullet, E., & Girard, M. (2000). Developmental and cognitive points of view on forgiveness. In M.E. McCullough, K.I. Pargament, & C.E. Thoresen (Eds.), *Forgiveness: Theory, research and practice* (pp. 111–132). Guilford Press.

Murray, R.J. (2002). The therapeutic use of forgiveness in healing intergenerational pain. *Counseling and Values*, *46*, 188-198. https://doi.org/10.1002/j.2161-007X.2002.tb00212.x

Oral, T. (2016). *Üniversite öğrencilerinin affetme düzeylerinin öz-anlayış, kişilerarası hataya ilişkin ruminasyon ve kişilik özellikleri açısından incelenmesi [The investigation of university students' forgiveness levels in terms of self-compassion, rumination about an interpersonal offense and personality traits]* (428811). [Doctoral dissertation, Necmettin Erbakan University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Özgür, H., & Eldeleklioğlu, J. (2017). REACH affetme modelinin Türk kültürü üzerinde etkililiğinin incelenmesi [Analyzing the effect of REACH Forgiveness Model on the Turkish culture]. *The Journal of Happiness & Well-Being, 5*(1), 98-112.

Öztörel, İ. (2018). *Psikolojik danışman adaylarının psikolojik sağlamlık, yaşam doyumu ve affetme düzeylerinin incelenmesi [Investigation of psychological resilience, life satisfaction and forgiveness levels of counselor candidates]*. [Master thesis, Yakın Doğu University].

Pamukçu, B., & Demir, A. (2013). Psikolojik Danışma Öz-Yeterlik Ölçeği Türkçe Formu'nun geçerlik ve güvenirlik çalışması [The validity and reliability study of the Turkish version of Counseling Self-Efficacy Scale]. *Turkish Psychological Counseling and Guidance Journal, 5*(40), 212–221.

Pamukçu, B., & Kağnıcı, D.Y. (2017). Beceriye Dayalı Grupla Psikolojik Danışma Eğitimi'nin grupla psikolojik danışma becerilerine etkisinin incelenmesi [The examination of the skilled group counselor training model's effect on group counseling skills]. *Elektronik Sosyal Bilimler Dergisi, 16* (61), 448-465. https://doi.org/10.17755/esosder.304685

Ross, S.R., Kendall, A.C., Matters, K.G., Rye, M.S., & Wrobel, T.A. (2004). A personological examination of self- and other-forgiveness in the five factor model. *Journal of Personality Assessment, 82,* 207-214. https://doi.org/10.1207/s15327752jpa8202_8

Rotter, J.C. (2001). Letting go: Forgiveness in counseling. *The Family Journal: Counseling and Therapy for Couples and Families, 9*(2), 174-177. https://doi.org/10.1177/1066480701092012

Sarıkaya, Y. (2017). *Süpervizör rolleri, tarzları ve süpervizyon terapötik ittifakının psikolojik danışma öz yeterliği ile ilişkisi [The relationship between supervisor roles, styles, supervisory working alliance and counseling self-efficacy]* (481309). [Doctoral dissertation, Gaziosmanpaşa & Giresun University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Sarpdağ, M. (2019). *Psikolojik danışman adaylarının psikolojik danışma becerilerinin yordayıcıları: psikolojik danışma özyetkinliği, duygu yönetimi ve kişilik özellikleri [The predictors of counseling skills of counselors trainess: Counseling self-efficacy, emotion* *management, and personality traits]* (555328). [Master thesis, Mehmet Akif Ersoy University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Satıcı, S.A. (2016). *Üniversite öğrencilerinin affetme, intikam, sosyal bağlılık ve öznel iyi oluşları: farklı yapısal modellerin denenmesi üzerine bir araştırma [The predictors of counseling skills of counselors trainess: Counseling self-efficacy, emotion management, and personality traits]* (432440). [Doctoral dissertation, Anadolu University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Seligman, M.E.P., & Csikszentmihalyi, M. (2000). Positive psychology. *American Psychologist, 55*(1), 4-15. https://doi.org/10.1037//0003-066X.55.1.5.

Şeker, U. 2019. *Psikolojik danışman adaylarının problem çözme becerilerinin ve mesleki öz-yeterlik algılarının özerklik, süpervizyon yaşantıları ve meslek etiği ile ilişkilerinin incelenmesi [Investigating of the counselor candidates' problem solving skills and counseling self-efficacy perceptions in relation to autonomy, supervision experiences and counseling ethics]* (579216). [Master thesis, Anadolu University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Şencan, H., & Fidan, Y. (2020). Likert verilerinin kullanıldığı keşfedici faktör analizlerinde normallik varsayımı ve faktör çıkarma üzerindeki etkisinin SPSS, factor ve prelis yazılımlarıyla sınanması. [Normality assumption in the exploratory factor analysis with likert scale data and testing its effect on factor extraction]. *Business & Management Studies: An International Journal, 8*(1), 640-687. http://dx.doi.org/10.15295/bmij.v8i1.1395

Şener, E., & Çetinkaya, F.F. (2015). Bir liderlik özelliği olarak affetme ve örgütsel düzeyde etkileri üzerine bir inceleme [Forgiveness as a leadership feature and a study on its effects on organizational level]. *Journal of Business Research, 7*(4), 24-42. https://openaccess.ahievran.edu.tr/

Şentepe, A. (2016). *Ruh sağlığı belirtilerinin yordayıcısı olarak affetme ve dindarlık ilişkisi [The relationship between forgiveness and religiosity as predictors of mental health symptoms]* (445979). [Doctoral dissertation, Sakarya University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Tabachnick, B.G., & Fidell, L.S. (2009). *Using multivariate statistics.* Allyn and Bacon Publishing

Thompson, L.Y., Snyder, C.R., Hoffman, L., Michael, S.T., Rasmussen, H.N., Billings, L.S., et al., (2005). Dispositional forgiveness of self, others, and situations. *Journal of Personality, 73*(2), 313–359. https://doi.org/10.1111/j.1467-6494.2005.00311.x

Topbaşoğlu, T. (2016). *Yaşam doyumunun yordayıcısı olarak öfke ve affetme: Affetmenin düzenleyici rolü [Anger and forgiveness as predictors of life satisfaction: The moderator role of forgiveness]* (445684). [Master thesis, Pamukkale University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Toussaint, L.L., Williams, D.R., Musick, M.A., & Everson, S.A. (2001). Forgiveness and health: Age differences in a U.S. probability sample. *Journal of Adult Development, 8*, 249–257. http://dx.doi.org/10.1023/A:1011394629736

Tse, W.S., & Yip, T.H.J. (2009). Relationship among dispositional forgiveness of others, interpersonal adjustment and psychological well being: Implication for interpersonal theory of depression. *Personality and Individual Differences*, *46*, 365–368.

Ülker-Tümlü, G. (2019). *Bireyle Psikolojik danışma Uygulaması Süpervizyonunda Ayrıştırıcı Süpervizyon Modeline'ne Dayalı Grup Süpervizyonu Sürecinin Yapılandırılması [Structuring the group supervision process in the supervision of individual counseling practicum based on the discrimination model]* (542987). [Doctoral dissertation, Anadolu University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr

Van Tongeren, D.R., Green, J.D., Hook, J.N., Davis, D.E., Davis, J.L., & Ramos, M. (2015). Forgiveness increases meaning in life. *Social Psychological and Personality Science, 6*, 47–55. http://dx.doi.org/10.1177/1948550614541298

Vural-Batık, M., & Afyonkale-Talay, N. (2021). The effect of a group psychoeducation program for ımproving forgiveness on the forgiveness levels of psychological counselor candidates. *Pamukkale University Journal of Education, 51*, 1-32. https://doi.org/10.9779/pauefd.%20686232

Vural-Batık, M., & Fırıncı-Kodaz, A. (2018). Kurum deneyiminin psikolojik danışman adaylarının özel eğitim öz yeterlik algılarına etkisi [The effect of the ınternship course on counselor trainees' sense of selfefficacy regarding special education]. *Ondokuz Mayis University Journal of Education Faculty, 37*(1), 209-222. https://doi.org/10.7822/omuefd.327621

Wade, N.G., Bailey, D., & Shaffer, P. (2005). Helping clients heal: Does forgiveness make a difference? *Professional Psychology: Research and Practice, 36*(6), 634-641. https://doi.org/10.1037/0735-7028.36.6.634

Witvliet, C.V.O., Ludwig, T.E., & Vander Laan, K.L. (2001). Granting forgiveness or harboring grudges: implications for emotion, physiology, and health. *Psychological Science, 121*, 117–123.

Worthington, E.L. (2001). *Five Steps to Forgiveness: The Art and Science of Forgiving*. The United States: Crown.

Yayla, E., & İkiz, F.E. (2017). Psikolojik danışmanların etkili nitelikleri ile danışma öz-yeterlik düzeyleri arasındaki ilişki [The relation between counselors' effective characteristics and counseling self-efficacy levels]. *Turkish Psychological Counseling and Guidance Journal, 48*(7), 31–44.

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

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 257–278

https://doi.org/10.21449/ijate.1210479

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

Application of the Rasch model in streamlining an instrument measuring depression among college students

Sherwin Balbuena [1,*]

[1]Dr. Emilio B. Espinosa Sr. Memorial State College of Agriculture and Technology, Masbate, Philippines

ARTICLE HISTORY

Received: Nov. 26, 2022
Revised: May 14, 2023
Accepted: May 14, 2023

Keywords:
Rasch,
Depression,
Precision,
Measurement,
USDI.

Abstract: Depression is a latent characteristic that is measured through self-reported or clinician-mediated instruments such as scales and inventories. The precision of depression estimates largely depends on the validity of the items used and on the truthfulness of people responding to these items. The existing methodology in instrumentation based on a factor-analytic approach has limited applicability, especially in the detection of sources of measurement error in item- and person-level analyses. While there are probabilistic approaches such as the use of Item Response Theory and the Rasch model in validating instruments, there are no definite guidelines on the sequence of steps to follow. This study explored the suitability of the Rasch model in assessing and streamlining the University Student Depression Inventory (USDI) using a sequential strategy based on the item response model assumptions, which involves fitting the data to the model through the elimination of misfits, analyzing retained items, and constructing measures. The strategy was applied to two sets of survey data collected from the same population of college students enrolled in a Philippine university but in different semesters. Results showed that the Rasch procedure was able to detect misfit items and persons, which guided decisions regarding the removal of problematic items and persons while preserving the reliability of the original scale. The methodology used was found to be replicable, as the analyses for the two datasets yielded comparable results in terms of number of items retained, item estimates and severity ordering, and distribution of student depression measures.

1. INTRODUCTION

Latent characteristics are human traits, constructs, or attributes that are neither directly observable nor tangible (e.g., feelings, affect, intelligence, etc.). As such, it is made manifest by eliciting responses from human subjects through interviews, tests, or self-reports. Usually, an individual responds to questions or items, and his/her responses are considered outer or observable indicators of this inner but unobservable human condition. So, the existence and quantity of the latent characteristic must be deduced from those observed, manifest responses (Bond & Fox, 2013).

In practice, a latent characteristic is indirectly measured through the administration of instruments. Prior to administration, these instruments undergo a rigorous development and

*CONTACT: Sherwin Balbuena ✉ balbuenasherwine@debesmscat.edu.ph Dr. Emilio B. Espinosa Sr. Memorial State College of Agriculture and Technology, Masbate, Philippines

e-ISSN: 2148-7456 /© **IJATE** 2023

validation process. At the early stages of the process, an item pool was developed based on the concept deduced from an underlying construct or latent characteristic, and each item was designed to capture information about the construct. Then the initial pool of items was piloted with a sample of the target population. The responses of this sample are usually subjected to factor analysis to assess the psychometric properties of the items. The items that load highly on the factors extracted based on eigenvalues (amounts of variance contributed) are retained and are therefore chosen to make up a scale. This process is based on a factor-analytic (FA) approach.

The FA method, also known as Classical Test Theory (CTT), is based on the premise that a test is valid and reliable if it comprises items that have high loadings with known variables related to the latent trait being assessed and if the responses to given items are consistent. Due to the presence of unpredictable items and person responses, it is more error-prone since it does not offer item- and person-level metrics to identify items that did not function as planned and to detect response sets. Furthermore, the estimation of individual abilities is test-dependent, while the estimation of item difficulties and discrimination is sample-dependent (Kohli et al., 2015).

Probabilistic approaches to instrument item analysis have emerged in recent years. One example is the use of item response theory (IRT) or the Rasch model (RM) to investigate the psychometric properties of constructed items and to validate and refine existing tests, questionnaires, or scales. Using this new approach, the response of a person to an item is modeled using a logistic function relating the person's underlying ability and the item's difficulty. As applied in scaling, an initial pool of items is created based on the underlying latent characteristic. The items are designed to gather information about the attribute at different severities in the latent continuum (a linear scale where a person can be identified as having less or more of the attribute). Next, the constructed scale is administered to a sample from the intended population. The responses of the sample are analyzed for model fit using fit statistics for both items and persons, and those items that do not deviate significantly from model expectations are retained in the new scale. This is one advantage of IRT, since it provides sample-free and test-free measures by estimating item and person parameters separately using conditional maximum likelihood methods and by requiring that response data fit the model.

The amount of the latent characteristic is quantified based on the outcome of the person's response to instruments. Measures of the latent characteristic of interest are obtained by summing up the person's ordinal item responses. The adequacy of the quantification of latent characteristics based on raw scores depend on the length of the questionnaire. In the factor-analytic point of view, the more items you include, the more valid and reliable the instrument becomes, as it gathers more information about the latent concept of interest. However, instruments with high reliability indices may contain redundant items (Boyle, 1985). Furthermore, using longer questionnaires would increase respondent burden, which would subsequently lead to low response rates (Stanton et al., 2002) and poor data quality (Galesic and Bosnjak, 2009; Maloney et al., 2011). More importantly, the validity of the data largely depends on the truthfulness of persons responding to questionnaire items and on the appropriateness of the items included in the questionnaire. Analysis of flawed data due to invalid responses and items would certainly produce invalid results and inferences about the latent characteristic.

An example of a latent characteristic that is currently attracting attention is depression. Depression is a common mental health problem that can impair an individual's functioning at home, work, or school (WHO, 2017). It is a medical condition characterized by a set of behavioral, cognitive, social, and biological symptoms (Hyde et al., 2008). Its severity ranges from a mild feeling of sadness to serious suicidal thoughts (Olsen et al., 2003; Forkmann et al., 2013; Balsamo et al., 2014). Depressive symptoms often manifest even at an early age, can be recurrent, and, if left untreated, will lead to the development of severe mental disorders (Hankin, 2006). Diagnosing the early signs of depression and providing appropriate interventions (e.g., counseling) can potentially prevent the progression of the disease, which is less costly than treating patients with severe depression (O'Connell et al., 2009).

The diagnosis of depression relies on self-reported instruments and diagnostic interviews, where the presence of a sufficient number of symptoms qualifies an individual as depressed. Over the past five decades, various depression inventories such as Beck's Depression Inventory (BDI), Patient Health Questionnaire (PHQ), and others have been developed, validated, and used to assess depression levels. These inventories have been employed in both general and specific patient populations, providing valid measures of depression. Typically, the patient's responses are analyzed using predetermined cutoff points to determine depression presence and severity. Different measurement frameworks, including CTT and IRT, have been employed to estimate depression levels using these instruments, yielding comparable measures (e.g., Stansbury et al., 2006; Shea et al., 2009; Balsamo et al., 2014; Wongpakaran et al., 2019).

The prevalence of depressive disorder is estimated using either self-reported instruments or a clinician-rating scales. The use of self-reported instruments has been found to overestimate the proportion of depressed individuals measured by diagnostic interviews. The lower point prevalence of depression obtained in diagnostic interviews might be due to the stringent criteria being adopted by clinicians in screening depressed individuals and could be associated with the socio-demographic characteristics of patients. Hence, the use of both methods in estimating depression prevalence was recommended (Lim et al., 2018).

Obtaining measures of depression is important for informing clinicians or researchers about this latent disease and its prevalence. However, measuring depression can be difficult, as it is done through a series of thorough observations and interviews with the patient. In the absence of psychiatric experts or clinicians to confirm the presence or absence of the disease, treatment is sometimes delayed, and undetected mild cases progress to severe cases. Alternative sources of depression measures are needed to detect not only the severe cases but also those at risk and provide a more inclusive mental health assessment. There are many available depression scales, but most of them are lengthy. We need to provide a rigorous statistical methodology by which we can assess and streamline these scales to efficiently measure depression for clinical use and research purposes.

To date, procedures for instrument assessment using IRT have been varied across fields of inquiry. In many health studies, using the Rasch model in instrument short-form development and psychometric validation is referred to as Rasch analysis. The analysis involves testing the following: (a) the data’s fit to the model; (b) the appropriateness of response format for polytomous items; (c) differential item functioning; (d) targeting of persons and items; (e) reliability; (f) local independence; and (g) unidimensionality (Tennant & Conaghan, 2007). Unfortunately, many studies applying Rasch analysis did not provide a definite sequence of steps to follow in assessing item properties, which could be used by researchers as a guide in streamlining existing instruments. Hence, there is a need to develop a robust statistical procedure for instrument quality assessment using Rasch analysis.

The general objective of this research was to evaluate the applicability of the Rasch model for assessing and analyzing an instrument measuring student depression for possible item reduction without compromising validity and reliability.

The specific objectives were as follows:

1. To determine the suitability of the Rasch model in the construction of scales for measuring depression in students;

2. To streamline the questionnaire items given the same precision level as that of the original instrument;
3. To develop an appropriate procedure for analyzing questionnaire items, which is applicable in evaluating the quality of instruments measuring other latent characteristics; and
4. To test the replicability of the procedure when applied to two datasets derived from the same population.

1.1. The Rasch Model

This study used the following assumptions of the Rasch model in forming the basis for deciding which items to discard, retain or review: unidimensionality (assessed using Rasch fit statistics and Martin-Löf test), local stochastic independence (detected using item residual correlations), and no differential item functioning (determined using ordinal logistic regression in IRT with gender as the only reference group). Person fit was also investigated to identify persons with aberrant response patterns to be excluded from the analysis. The procedure developed in this study was empirically applied to a dataset derived from two surveys on the mental health of Filipino college students. However, it is assumed that this procedure will be sufficiently robust when applied to streamlining instruments measuring other types of latent characteristics.

Named after its originator Georg Rasch (1960), a Danish mathematician, the Rasch model is a statistical model that is used to analyze data from educational and psychological measurement instruments, such as tests, surveys, and questionnaires. It is a type of item response theory model, which is a framework for understanding how individuals respond to specific items on a measurement instrument. The Rasch model is based on the assumption that the difficulty of an item and the ability of the person responding to the item are related in a specific way. The model specifies a mathematical relationship between the two, which allows researchers to estimate an individual's ability level based on their responses to a series of items (Bond & Fox, 2013). The idea of invariant measurement, which is developed through specific objectivity, governs the model. Item difficulty may be evaluated independently of the people included in the sample, and individual ability can be estimated irrespective of the test items (Wright & Linacre, 1987).

One feature of the Rasch model is that it provides estimates of person measure (or person ability) and item location (or item difficulty). The item and person estimates can be calibrated on the same logit scale. Furthermore, unlike traditional ordinal scales using unequal-spaced intervals of scores to rank the underlying ability from less to more, Rasch scaling permits equal spacing of intervals (Yu, 2011). This is done by finding the logarithm of the original (raw score-based) scaling used. Theoretically, the difference between two discrete raw scores (which are actually sums of affirmative or ordinal item responses) is not meaningful. Hence, Rasch scaling resolves this problem by converting discrete raw scores into continuous logit measures after fitting the data to the model, which can be used not only to determine who has more or less of the ability but also to compare the relative distances between person abilities.

Essentially, the Rasch model is based on the theory about the construct of latent characteristics under scrutiny. The construction of items is guided by a thorough understanding and definition of the latent concepts involved and of the behavioral manifestations (i.e., responses) that represent the construct. The order of item locations can be empirically ascertained by estimating the item difficulty parameters after fitting the response data to the Rasch model.

1.2. Measuring Depression in Students

Detection of depressive disorder in an individual is done using self-reported instruments and diagnostic interviews, where the presence of a sufficient number of symptoms qualifies the individual as depressed. Diagnosis of depression is also done through the analysis of patient’s self-reported symptomatology in interviews. For the past five decades, several clinical or research depression inventories have been developed, validated, and then used to come up with measures of the depression level. Some inventories commonly used for the general population include Beck's Depression Inventory (BDI; Beck et al., 1961), Patient Health Questionnaire (PHQ; Spitzer et al., 1999), Hospital Anxiety and Depression Scale (HADS; Zigmond and Snaith, 1983), Depression Anxiety Stress Scale (DASS; Lovibond and Lovibond, 1995), Centre for Epidemiological Studies - Depression Scale (CES-D; Radloff, 1977), Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960), and Zung Self-Rating Depression Scale (Zung, 1965). Some of these inventories have already been used in specific patient populations. Disease-specific depression scales, which are new versions of the above inventories, have been administered and found to provide valid measures of disease-related depression levels.

An instrument called the University Student Depression Inventory (USDI; Khawaja and Bryden, 2006) is used to screen for student depression, a non-disease-related kind of depression. The measure was initially created and tested on a sample of university students from Australia. Following principal component factor analysis using oblique and orthogonal rotation techniques, the developers of USDI identified three (3) factors from an original set of 125 produced items. The final instrument (found in Appendix A) was composed of 30 items classified into three sub-scales: lethargy (LG; 9 items), cognitive-emotional (CE; 14 items), and academic motivation (AM; 7 items). To obtain a measure of student depression using the USDI, the scores in the instrument are added, and a higher total is interpreted as an indication of a higher level of depressive disorder. Recent local studies used the USDI to determine some factors associated with depressive symptoms among Filipino college students (Lee et al., 2013; Lailo, 2018). However, this instrument alone cannot diagnose or confirm that a student actually has depression and should be used only to measure the degree of vulnerability of a student to developing severe depression (Lailo, 2018).

The validity and reliability of the USDI have already been established. Further psychometric validation studies were conducted to confirm the factor structure of the USDI using multi-cultural student populations (Sharif et al., 2011; Romaniuk & Khawaja, 2013; Khawaja et al., 2013; Habibi et al., 2014). The instrument has already been used in a number of studies estimating the prevalence of depressive symptoms among tertiary students (e.g., Mikolajczyk et al., 2008; Deb et al., 2016; Gesinde & Sanu, 2014). However, research on the application of Rasch analysis in validating the USDI was limited.

2. METHOD

2.1. Research Design

To achieve the objectives of this study, an exploratory type of research design was used. The theory on which the Rasch model is based is that the items in a scale or questionnaire were constructed with varying levels of difficulty, which can be ordered along a single continuum. The suitability of this model in analyzing questionnaire items (with unknown difficulty levels) was explored through the application of the model to the response data obtained from university-wide mental health surveys. In light of the Rasch model assumptions, a strategy for streamlining questionnaires was developed and applied to the data set to explore its soundness and replicability.

2.2. Data Sources

The data used in this study are the responses of college students to items in the University Student Depression Inventory (USDI). This questionnaire was used on two mental health surveys in a state university in Southern Luzon to detect the presence of depression in students as manifested in depressive symptoms. The first dataset was taken from the result of the university-wide survey conducted in the school year 2018-2019, referred to here as the STAT 173 survey (UPLB INSTAT, 2018), which involved 441 college students. The survey aimed to determine the level of depression among undergraduate students, specifically to describe depression incidence among students and to identify possible determinants of student depression. The second dataset was taken from the result of the survey conducted by Lailo (2018) in the school year 2017-2018 at the same university, which involved 169 college students.

2.3. Methodology

By adhering to the assumptions of the Rasch model, the following steps were done to achieve the objectives of the study: (1) select an appropriate model; (2) estimate the model parameters and identify misfitting persons; (3) after removal of person misfits, re-estimate the model parameters and identify misfitting items; (4) after removal of item misfits, re-estimate the model parameters and identify misfitting items until no further items are misfitting; (5) assess the reliability at each instance of item/person removal until reliability declines tremendously; (6) order the item severity estimates and check for consistency with established symptomatology to assess construct validity; (7) detect local dependence (LD) and differential item functioning (DIF) for possible item redundancy and bias; (8) estimate person measures and transmute with raw scores; and (9) locate the thresholds for varying severity levels. The analyses involved in the procedure were implemented in R, Microsoft Excel, and SPSS Version 28.

The same strategy described above was applied to the analysis of Lailo's (2018) data. The response data were obtained using the same instrument and from the same population of college undergraduate students, but in different semesters. The resulting streamlined versions of the instrument from the analyses of the two data were compared in terms of the number and similarity of retained items, overall instrument reliability, and constructed measures of student depression.

The main purpose of this study was to develop a strategy based on Rasch model tests for assumptions to assess and streamline a depression scale and produce a shorter version of the original scale, which is equally valid and reliable. Using two sets of survey data, the soundness of the strategy was determined as will be described in this section.

3. FINDINGS

3.1. Analysis of STAT 173 Data

3.1.1. *Model selection*

The Rasch model is a family of statistical models used in psychometrics for analyzing categorical data. The selection of the appropriate Rasch model depends on the type of data being analyzed. In general, there are two main types of Rasch models: dichotomous and polytomous. The dichotomous Rasch model is designed for binary data, where responses are either correct or incorrect. The polytomous Rasch model, on the other hand, is designed for data with more than two response options, such as Likert scales. Therefore, choosing the most appropriate model depends on the nature of the data and the research question at hand.

The most basic formulation is the dichotomous Rasch model (DRM), which is also referred to as the one-parameter logistic model (1PL). Let $X_{ni} = x \in \{0,1\}$ be a dichotomous random variable, where x = 0 and x = 1 indicate "no" and "yes" responses, respectively, to a questionnaire item. The function

$$P(X_{ni} = 1) = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \tag{1}$$

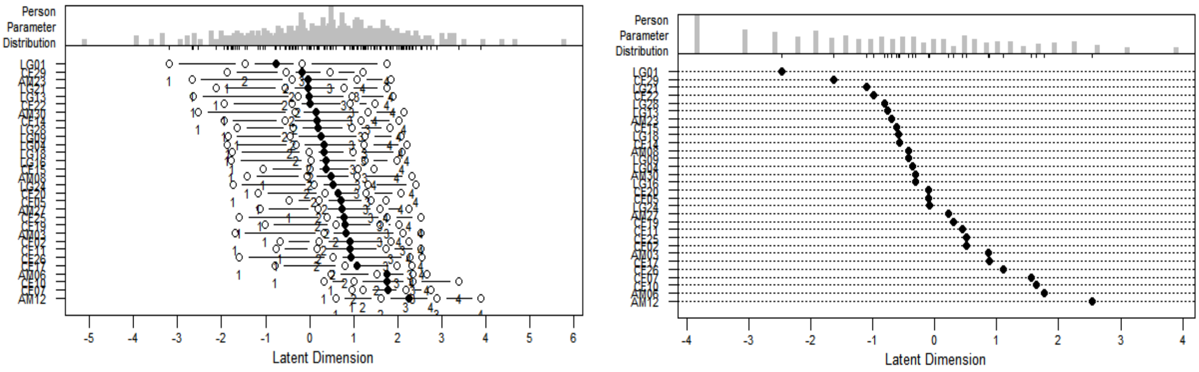
models the probability that person *n* will agree with item *i*, where δ_i is the difficulty of item *i* and β_n is the ability of person *n*. This function conjectures that the higher a person's ability relative to the difficulty of an item, the higher the probability of an affirmative response on that item, a relation that can be illustrated by a sigmoid graph with the person's ability as the abscissa and the probability of agreeing as the ordinate.

On the other hand, the Partial Credit Model (PCM; Master's 1982) is a type of Rasch model that is used to analyze polytomous data, which is data that has more than two response categories. This model assumes that a person's response to an item is a function of the person's ability, the item's difficulty, and the threshold parameters that describe the level of difficulty at which the person is able to transition from one response category to the next.

Using the two models DRM and PCM, two sets of parameter estimates were derived using the eRm package (Mair & Hatzinger, 2007) in R. Results showed that the least severe and the most severe items for both models were LG01 ("I am more tired than I used to be") and CE20 ("Going to university is pointless"), respectively. Although the item severity orders were found to be similar (with minor changes in ordering for items with moderate severity measures that are close to one another), the infit and outfit indices changed remarkably after the response was dichotomized.

The person-item map (PIM) also provides rich information about the relationships between item and person estimates and their distributions. Figure 1a shows the nearly symmetric person distribution after fitting PCM, with central tendencies located between 0.0 and +2.0 logits. The item severity range (-0.77 to +2.25) along the latent dimension also spanned the width of the person distribution, which means that the USDI was well-targeted for the given population of college students. When data were fitted to DRM, the person distribution became more dispersed and the item severity range increased in width (-2.47 to +2.54 logits) as shown in Figure 1b, although the item severity ordering did not change significantly.

Figure 1. *STAT 173 data PIMs under (a) PCM and (b) DRM.*



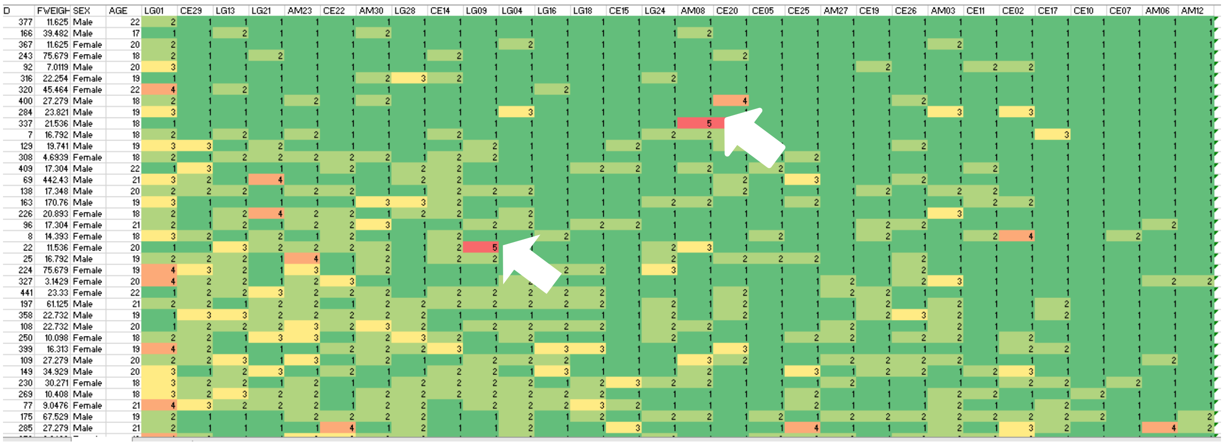
(a) (b)

The alteration of the person measure distribution above suggests the inappropriateness of dichotomizing USDI response data. Hence, it is logical to discard the use of DRM as an option to modelling originally polytomous data. Hence, the polytomous PCM was used to analyze the STAT 173 data. PCM was also used in the analysis of STAT 173 data to warrant comparison of results with Lailo's data since it was found out that a 5-point Likert scale did not apply for some items in the latter.

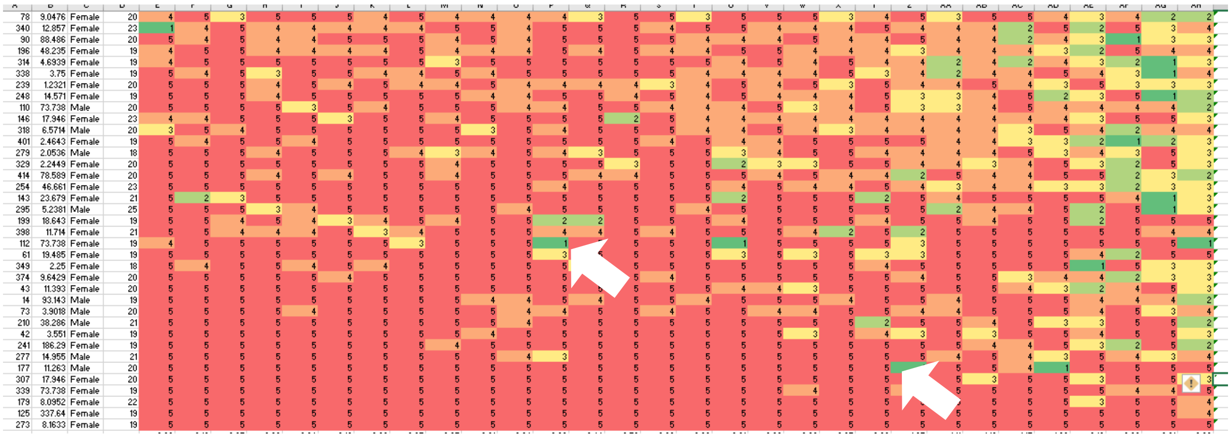
3.1.2. *Person misfits*

After fitting the PCM, person estimates were obtained with corresponding person fit analysis. The eRm package provides a summary of fit analysis with chi-square, outfit/infit mean square, and outfit/infit t indices. From the result of fit analysis, there were 69 persons with very high values for both infit and outfit exceeding the threshold of 1.3 and with chi-square p-values less than 0.05; hence, they were labelled as misfits (also referred to as underfit persons). These persons have so highly unpredictable responses that they distort the measurement system (Linacre, 2002), or they are known to cause measurement error. Another helpful approach to detecting misfits is by arranging the items and persons according to their estimated locations in the order from less to more severe/depressed and then by examining their patterns of responses to items. To illustrate this approach, a Guttman (1950) scalogram heatmap showing the noticeable patterns of responses by misfit persons was constructed as shown in Figure 2. Unusual patterns of responses, as represented by uncommon cell colors, also represent a substantial deviation from the expectations of the Rasch model (i.e., large value of residual) as was detected by infit and outfit statistics. For example, the two responses in Figure 2a (shown by the arrows) represent the high-category responses of less depressed Person #337 to items AM08 and CE20 and the high-category response of less depressed Person #22 to Item LG09. Corresponding outfit(infit) for these persons show very high mean square values, 5.62(5.05) and 1.67(2.12), respectively. For more depressed persons #112 and #177 in Figure 2b, these unusual response patterns were also detected by high outfit(infit) statistics, 3.23(3.11) and 2.68(3.63), respectively.

Figure 2. *An Excel®-generated Guttman scalogram heatmap showing the patterns of item responses of (a) low-ability (less depressed) and (b) high-ability (more depressed) persons. A row containing totally different or contrasting cell colors (e.g., a red surrounded by a dominant green or a green surrounded by a dominant red) represents the response of a misfitting person.*



(a)



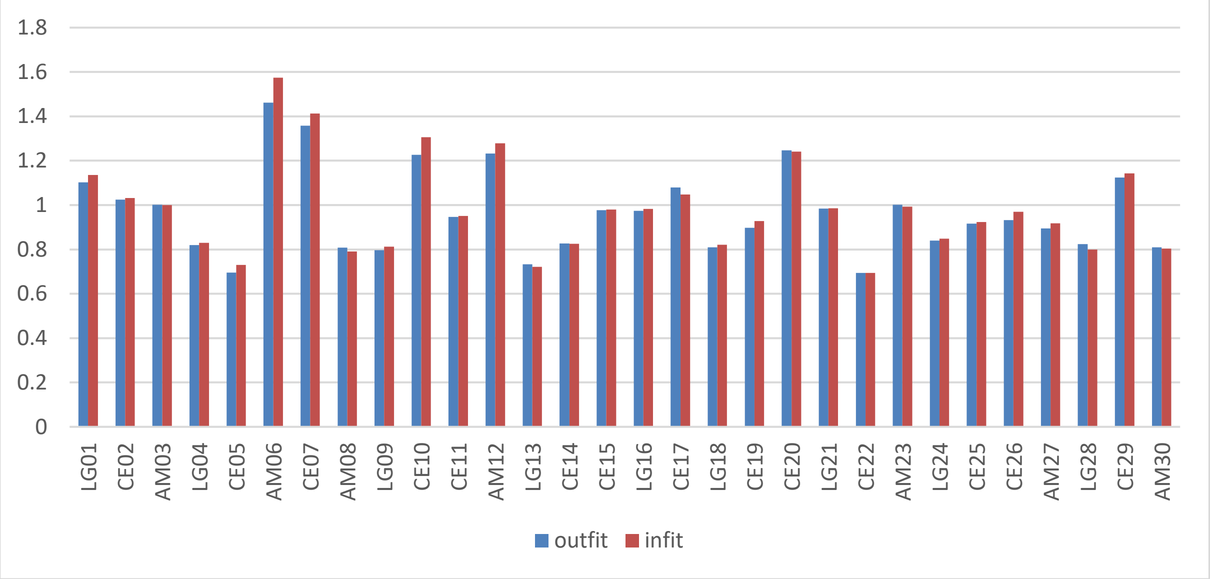
(b)

These misfitting persons were removed from the analysis, as the information they provided did not contribute to useful measurement of student depression. Following this weeding out, a 16% decrease in the number of samples was noted.

3.1.3. *Item misfits*

Using the new dataset with 372 fit persons, the item and person parameters were estimated using the PCM. The new results were obtained, which include the estimates when the complete sample was used. Comparing the two results, slight changes in the item severity estimates and ordering and noticeable changes in outfit/infit for some items were observed. Item AM06 ("I don't attend lectures as much as I used to") and CE07 ("I have thought about killing myself") consistently had very high values for both infit and outfit; hence, they were labelled as misfits. Item CE10 ("No one cares about me") had an infit value exceeding the threshold, but this is considered trivial. Hence, only two items (AM06 and CE07) were considered for removal, as these items were believed to contribute substantial error variance to analysis.

Figure 3. *Clustered column bar charts for the infit and outfit of USDI items in STAT 173 data after removal of 69 misfitting persons.*



Another approach in assessing fit of the items is by examining infit-outfit relationship through graphical method using clustered column bar (CCB) charts. In Figure 3, the unusual patterns of infit and outfit values can be observed, and problematic items can be identified. At a glance, one can note that items AM06 and CE07 have remarkably high outfit-infit values compared to the rest of the items. This observation corroborates the previous decision made to discard the two items, which are believed to be unproductive for the construction of measures for student depression.

After removal of items AM06 and CE07, three more items were found to show very high outfit-infit values: AM12 ("Going to university is pointless"), CE10 ("No one cares about me"), and CE20 ("I spend more time alone than I used to"). This is to be expected, since these items previously had tall outfit-infit bars in Figure 3 secondary to those of the two already discarded, meaning they contribute substantial measurement errors. Following item removal, both outfit and infit values of these items escalated and exceeded the cutoff, hence they were labelled as misfit items. Further removing these three items from the data showed no further items misfitting the PCM. In Table 1, there are no items with remarkably high fit statistics. This means that the STAT 173 data with the remaining 25 items conformed to the unidimensionality assumption (Wright and Panchapakesan, 1969). Furthermore, Martin-Löf test showed a nonsignificant result ($LR = 857.936$, $df = 24$, $p > 0.05$), indicating that the data appears to be unidimensional.

Table 1. *STAT 173 data item estimates and fit statistics after removal of five misfitting items, showing satisfactory fit statistics.*

Item	Severity	Outfit	Infit
LG01	-0.61	1.12	1.15
CE02	1.56	1.21	1.21
AM03	1.53	1.12	1.14
LG04	0.69	0.87	0.88
CE05	1.30	0.79	0.81
AM08	1.04	0.87	0.85
LG09	0.65	0.83	0.85
CE11	1.56	1.11	1.09
LG13	0.27	0.75	0.74
CE14	0.60	0.93	0.90
CE15	0.86	1.06	1.03
LG16	0.74	1.06	1.07
CE17	1.70	1.21	1.17
LG18	0.78	0.86	0.86
CE19	1.37	0.99	1.03
LG21	0.33	1.05	1.04
CE22	0.47	0.70	0.69
AM23	0.37	1.03	1.03
LG24	0.99	0.89	0.90
CE25	1.35	1.01	1.02
CE26	1.48	1.00	1.05
AM27	1.35	0.95	0.98
LG28	0.58	0.82	0.80
CE29	0.09	1.16	1.16
AM30	0.57	0.82	0.81

3.1.4. *Reliability assessment*

Since the first PCM estimation and throughout the instrument assessment process, reliability analysis has been carried out, especially when a group of misfit respondents or a group of misfit items were excluded from the analysis. The reliability analysis summary for each instance of excluded persons or items is shown in Table 2. The reliability indices in Stage 1 were found to be very close to Cronbach's alpha. Note that even after the elimination of 69 people, the PSI remained constant. The reliability was unaffected significantly by subsequent item reductions either. Additionally, the person separation is 5.69, which indicates that the sample of college students may be divided into around six distinct depression severity categories. This suggests that the new 25-item USDI is as equally precise and accurate in measuring student depression as the longer USDI.

Table 2. *Summary of person separation reliability assessment at each stage of fit analysis and item/person reduction of STAT 173 data.*

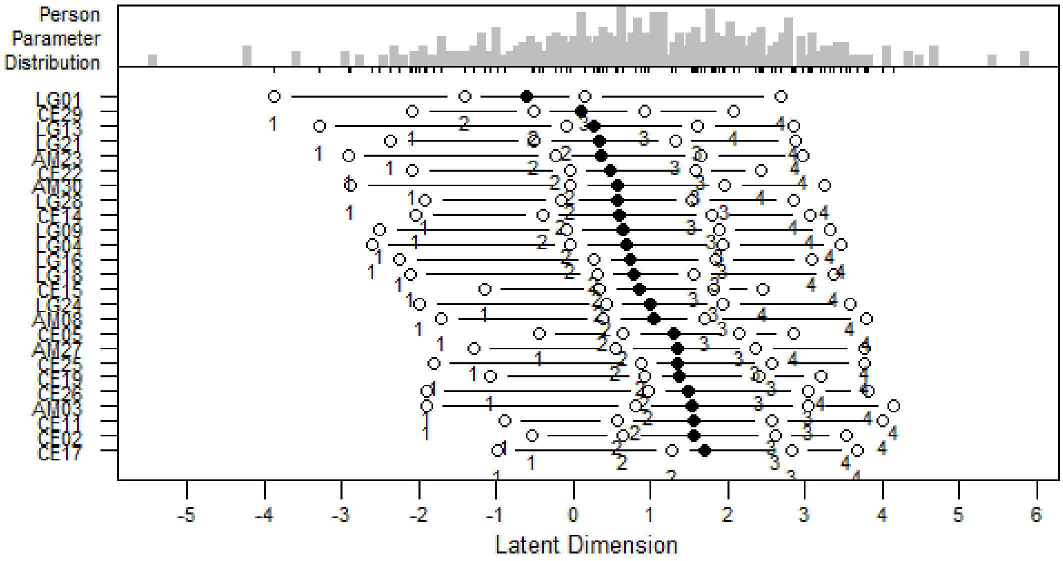
Stage	Instance	Observed variance $\sigma^2_{\hat{\beta}}$	Model error variance $\sigma^2_{\hat{e}}$	Reliability (PSI)
1	Used 30 items and 441 persons*	2.0097	0.0636	0.97
2	Used 30 items and 372 fit persons	2.6687	0.0737	0.97
3	Used 25 items (AM06, CE07, AM12, CE10, and CE20 discarded) and 372 persons**	3.1641	0.0894	0.97

*Cronbach's alpha = 0.969; **Person separation = 5.69

3.1.5. *Item severity ordering*

The item severity estimates are regarded as reliable indicators of location when a person falls on the severity spectrum of depression provided that all the items fit the PCM. The relative ordering of the 25 remaining items based on the values of item estimations represented in logits may be determined by creating a PIM. Based on the item location (solid circle) and threshold (hollow circle) estimates, the PIM in Figure 4 shows how the items are arranged from less severe to more severe manifestations of student depression. It is clear that the majority of the items in the below-median group fall under the LG subscale, whereas the majority of the items in the above-median group belong to the CE subscale, if the items are divided into two groups (i.e., less severe and more severe) based on their location above or below the median item measure. The AM subscale items do not display any specific severity classification. Based on extreme locations, anhedonia, or the loss of interest in previously enjoyed activities (cognitive-emotional), as manifested in Item CE17 ("The activities I used to enjoy"), is the most severe symptom of depression. Anhedonia is actually one of the primary signs of major depression (American Psychiatric Association, 2013).

Figure 4. *The PIM for the remaining 25 USDI items in STAT 173 data showing the locations of the items along the latent dimension (depression severity) with corresponding response category threshold estimates.*



3.1.6. *Detection of LD*

Using the criteria of Smith (2000) and Chen and Thissen (1997) on item residual correlations, one pair of LD items was identified. With a correlation coefficient of 0.3151, Items CE25 ("I feel withdrawn when I'm around others") and item CE26 ("I do not cope well") were found to be locally dependent. The choice in this case is to examine each item's contents for potential revision rather than to discard one of the items. According to research, LD items should be revised by combining the two items into a single "super-item" (Wainer & Kiely, 1987).

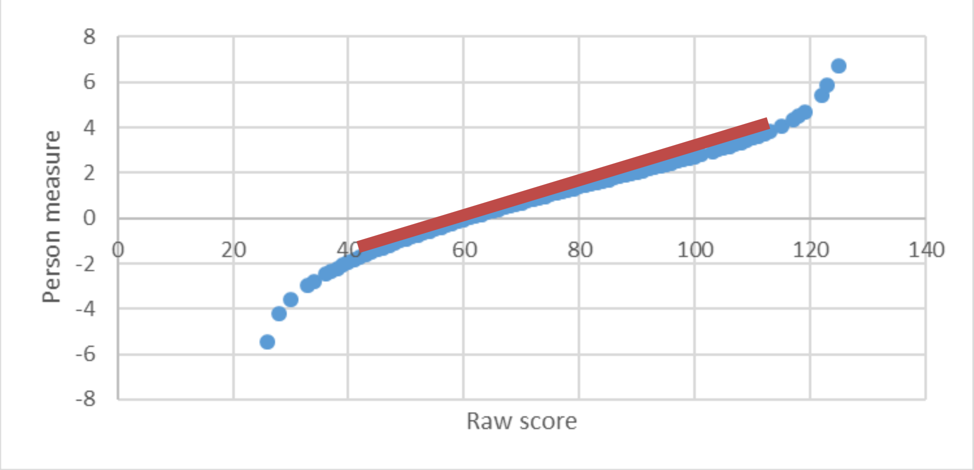
3.1.7. *Detection of item bias or DIF*

Using ordinal logistic regression in IRT with gender as the only reference group and then following the chi-square criterion based on the likelihood ratio χ^2 test (Swaminathan and Rogers, 1990), three items (LG01, AM03, and AM08) were marked for DIF. It is recommended that these items be retained and check their content for any idiosyncratic meanings or getting gender-specific item attributes for these items, which may be utilized to establish different norms for male and female students.

3.1.8. *Construction of measures*

By adding the ordinal values assigned to each student's categorical responses across all items for the remaining 25 items in the streamlined USDI, the level of depression is calculated for each student. The result is a number called the raw score, which has a range of 25 to 125. Equivalent interval-level measures for the raw scores were obtained because all the items fit the PCM. For some middle scores, an almost linear relationship can be seen (see Figure 5) if the scatterplot of raw scores and Rasch person measures is constructed. The scatter plot resembles a straight line between raw scores 40 and 110. Within this range, valid transformations from discrete raw score to continuous person measure are offered by interpolation using a linear function.

Figure 5. *Raw score to Rasch measure transformation scatterplot in STAT 173 data, showing the almost linear relationship between raw scores 40 and 110.*



3.1.9. *Scaling and classifications of depression level*

To determine provisional thresholds for classifying students into groups with varying levels of depression (i.e., very high, high, moderate, and low), the averages of item thresholds of all the items were considered. This approach was considered valid since the USDI instrument was well-targeted for the given population of college students, because it was discovered that the distribution of item measures and person measures were similar. Additionally, no item was found to have disordered thresholds, indicating that the scale structure used was effective and that student responses increased monotonically as depression levels increased. Hence, the thresholds estimated from these scale structures and probable responses to items representing depressive symptoms could be used to demarcate various levels of depression experienced by students.

Table 3. *Current classification thresholds for each category in STAT 173 based on continuous person measures expressed in logits with corresponding discrete raw score thresholds.*

Category	Person Measure		Raw Score	
	Lower limit	Upper limit	Lower limit	Upper limit
Low	—∞	-1.95	25	40
Moderate	-1.95	0.64	41	69
High	0.64	3.24	70	106
Very high	3.24	+∞	107	125

While the person separation in Table 2 suggested six categories of depression severity, this study used only four categories to warrant comparison with the results of the original STAT 173 data analysis using four levels of depression severity (i.e., 30-59 low, 60-89 moderate, 90-119 high, 120-150 very high), referred to here as the previous scale. To determine the cutoff for each category, the average of threshold estimates for categories 1 and 4 were computed and found to be at -1.95 and +3.24 logits, respectively. These two cutoff points represent the upper limit for the low depression category and the lower limit for the very high depression category. The middle threshold (or cutoff for classifying between moderate and high levels) was determined by finding the midpoint between the average threshold estimates for categories 1 and 4, which is +0.645 logits. Hence, the common distance between thresholds 1 and 2 and between thresholds 2 and 3 is 2.595 logits. The classification cutoffs (referred to as the current scale) are shown in Table 3 in which the equivalent raw score category limits are also indicated.

Using the data without misfit persons and items, the number of persons with inconsistent classifications between the previous and current scales was determined. Results showed that 90 out of 372 persons had inconsistent classifications, majority of which were transferred from a lower category to the next higher category. The inconsistencies occurred as a result of the change in raw score intervals when the logit-based thresholds were used; that is, some category interval widths were widened or narrowed when the equivalent raw score category limits were used as cutoffs. But the current classification based on interval-level logit measures is more valid, since the previous one was based on intervals of discrete raw scores derived from the sum of ordinal response data for which equally spaced intervals cannot be constructed (Yu, 2011).

3.2. Analysis of Lailo's Data

After applying the same sequential strategy on the analysis of Lailo's data, the following results were obtained. Using the outfit and infit statistics based on PCM, 17 out of 135 persons were found to be misfits and then removed. Five items (AM06, CE07, LG18, CE20, and AM23) were identified as misfits and then removed. Following this removal of sources of measurement error, the remaining 25 items achieved a good fit to the PCM while preserving the internal consistency of the original USDI at 0.96 PSI. Furthermore, no items were detected for gender DIF, while some adjacent questionnaire items were flagged for LD. Locally dependent items were not considered for removal as the result of item residual correlations might have been caused by respondents' acquiescence to the redundancy of questionnaire items or simply a false positive detection.

The distributions of the estimated person and item locations showed that the USDI was well-targeted to the given population of college students. Measures of depression at various severity levels were also constructed based on the estimated thresholds as shown in Table 4.

Table 4. *Current classification thresholds for each category in Lailo's data based on continuous person measures expressed in logits with corresponding discrete raw score thresholds.*

Category	Person Measure		Raw Score	
	Lower limit	Upper limit	Lower limit	Upper limit
Low	$-\infty$	-1.13	25	43
Moderate	-1.13	+0.39	44	68
High	+0.39	+1.91	69	100
Very high	+1.91	$+\infty$	101	125

Using the previous classification thresholds for Lailo's data (the same ordinal scale used in STAT 173 data) and the new classification thresholds, 11 persons were found to be inconsistently classified, all of whom were transferred to a next higher category.

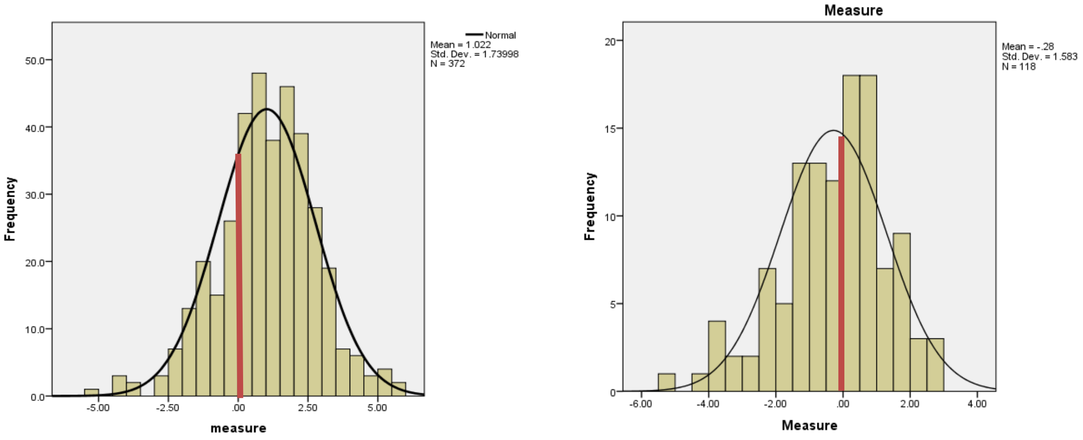
3.3. Results for STAT 173 and Lailo's Data: A Comparison

From the above analyses, evidences on the properties of the USDI items were gathered, which served as basis for item removal and further item analysis. Removal of items was based on misfit values only. Items flagged for LD and DIF were retained for further review of item contents. There were five items removed, but only three of these items were common to both data contexts (AM06, CE07, and CE20). Separate orderings of the item location estimates of the retained items common to both data contexts showed high Spearman rank-correlation coefficient ($\rho = 0.87$, $p < 0.05$), which indicates identical orderings of item severities. Finally, the reliability of the instrument was preserved between 0.96 and 0.97.

Following the estimation of the PCM using the information in retained items, final person estimates were also derived. When the two estimates of person measures obtained in STAT 173 data and Lailo's data were compared, a low score, say, 26 had fairly close person estimates (-5.45 logits for STAT 173 and -5.23 logits for Lailo, with a distance of 0.22 logits), whereas a high score, say, 113 had quite far-off transmuted person measures (+3.82 logits for STAT 173 and +2.85 logits for Lailo, with a distance of 0.97 logits). It is evident that the range of student depression measures may have been impacted by sample size or by different temporal characteristics of data. However, it was noted that the relationship between raw score and person measure for both data contexts is approximately linear between raw scores 40 and 110.

To have a glimpse of the prevalence of depression in the given population of college students, separate histograms of person measures for STAT 173 data and Lailo's data were constructed. Figure 6 shows the distribution of students along a continuum of depression severity measures. Although the mean person measure 1.02 (SD=1.74) in STAT 173 data is greater than the mean person measure -0.28 (SD=1.58) in Lailo's data, the shapes of the distributions are very similar, with slight negative skewness, -0.29 (SE=0.13) and -0.56 (SE=0.22), respectively. Despite the different time contexts of populations from which the samples were taken, the consistent prevalence of depression among college students was captured in the analyses of survey data. The STAT 173 data provided a more precise estimate of depression measures since it used a larger number of samples, which reduced the variance observed among person estimates as may be explained by the peaked distribution in Figure 6a.

Figure 6. *Histograms for the student depression measures obtained in the analysis of (a) STAT 173 data and (b) Lailo's data, showing the theoretical normal distribution derived from the mean and standard deviation of person estimates.*



(a)

(b)

Distinctive properties of some USDI items were observed in the analysis of Lailo's data. First, three of the 25 remaining items had disordered thresholds, meaning that some response categories did not function as intended. Second, two items did not have 1 responses, meaning that students did not respond "not at all" to these items. These observations were not present in the analysis of STAT 173 data.

Figure 7. *ICCs for 3 USDI items with disordered thresholds in Lailo's Data.*

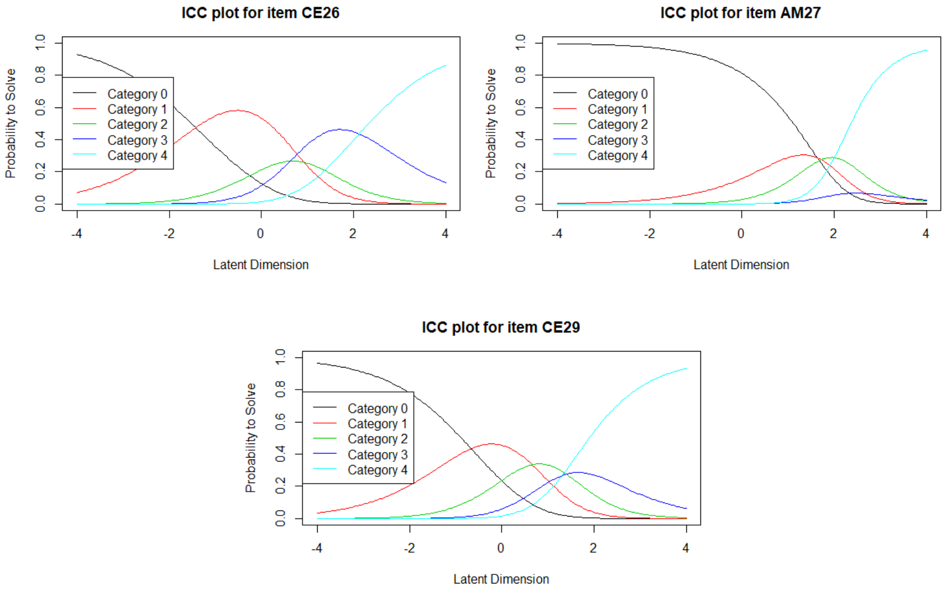


Figure 7 shows the item characteristic curves (ICC; also referred to as item response functions or category characteristics curves) of the three items with disordered thresholds. The threshold estimates for category 2 and 3 in Item CE26 were found to be 0.93 and 0.63, respectively. Also, for items CE29 and AM27, threshold estimates for categories 3 and 4 were also disordered. Items with disordered thresholds disrupt the measurement of the underlying trait (Tennant, 2004).

For these items with response options that did not work as intended, it was suggested that the rating scale structure be revised. For the two items lacking responses in one category, a 4-point response structure (i.e., 1234) be used in future data collection. For the items with disordered thresholds, the adjacent disordering categories be collapsed into one category (i.e., for CE26 a rating scale structure of 12334; for CE29 and AM27 a rating structure of 12344), hence a 4-point response structure. However, the 5-point scale structure of the rest of the items should be retained.

4. DISCUSSION and CONCLUSION

Based on the results, the procedure applying the extended Rasch model PCM in assessing and streamlining the USDI questionnaire based on two sets of response data yielded comparable results. The scale's items were reduced to the same number for both data contexts, while maintaining instrument reliability. The ordering of the items from various domains (subscales) based on item measures along the continuum of depression severity was found to be consistent with the symptomatology of clinical depression, confirming the construct validity of the streamlined version. Items flagged for LD, DIF and high outfit mean square values were recommended for further investigation of contents, problems in data collection, and possible person subgroup-specific meanings, after which decisions as to revise or entirely discard the items may be made for future use.

The application of the Rasch model in the study was found to be suitable and productive. In addition to the usefulness of outfit and infit statistics in detecting problematic items and persons, other meaningful information about the items and of the entire scale was obtained. The ordering of item severities and distributions of item and person measures provided a basis for the assessment of the instrument's targeting, which is helpful in locating provisional thresholds for various depression severity cases. The detection of item redundancy and bias was also present unlike in traditional item analysis methods. Reliability analysis through person separation provided an evidence of the scale's internal consistency, which is analogous to the Cronbach's alpha. The construction of interval-level measures of student depression would satisfy the conditions set by parametric statistical methods, which makes the computation of effect sizes after implementing interventions and comparison among group means possible.

Overall, the methodology used successfully streamlined the USDI questionnaire, from which person measures were successfully derived. The construction of measures for student depression in both data contexts was also comparable in terms of item threshold estimates. These estimates were used to set provisional thresholds for classifying students of various depression level categories based on combined estimates and, consequently, to help determine the optimal cutoff points when enough data become available. The distributions of student depression measures for both data contexts were found to be consistent despite the difference in time contexts of populations from which these samples were taken. However, estimates for person measures in the two datasets provided varied transmutations for raw scores. The anomalies observed might have been caused by varying sample sizes, survey designs and data collection procedures, items retained in the scale, and temporal characteristics of sample data.

The methodology illustrated in this study explicitly provided a sequence of steps to follow, which is applicable to assessment of other instruments used to measure the prevalence of a latent population characteristic. Generally, there are three steps: (1) fitting the data to the model by eliminating misfits; (2) analyzing retained items; and (3) constructing measures. The sequence was done in the decreasing order of importance. Since the Rasch model requires that data fit the model, decisions on discarding items or persons based on fit statistics have more weights than decisions guided by results of other analyses. In reality, there are solutions for LD and DIF items aside from removal, whereas no solution can be offered to misfitting items/persons but to discard them. It is the strong point of this procedure, espousing the Rasch requirement for invariant measurement, which was empirically demonstrated to be replicable.

In Rasch analysis, the solution for misfits includes the removal of persons with unpredictable responses. However, caution must be taken when ignoring responses from individuals due to model misfit. Another risk to the sample's representativeness is when people responses are eliminated from the survey data. While eliminating outliers improves model fit, a much smaller sample size might result in considerably more serious issues, such as inaccurate estimates of the prevalence of depression and false conclusions. Therefore, in addition to excluding person misfits from analysis of survey data, alternate methods to handle person misfits may be investigated. For example, misfits may be included in the analysis after imputing their health status; this involves replacing the aberrant item response of a person with a given location on the latent continuum by taking into account the responses of good-fit persons with the same location.

It is advised that routine data cleaning be used when analyzing survey data, and that Rasch analysis be used to identify individuals who would unintentionally introduce random noise into subsequent analyses. Rasch analysis cannot pinpoint the specific response bias that may have taken place (e.g., acquiescence, social desirability, guessing, and malingering), but it can at least identify potential sources of measurement noise that could interfere with the identification and estimation of population characteristics, particularly latent traits.

Further studies on the effect of varying proportions of misfits on precision of estimates are encouraged. These studies should include a simulation of Rasch model parameter estimation when person misfits are not included in the data and when they are included to some extent. The estimates obtained in various scenarios can be compared to determine and predict possible impacts of using misfits on the power and validity of survey data. Furthermore, in fitting the Rasch model, these studies should consider the sampling design used in data collection to reduce unwanted bias in the estimation of model parameters. A new statistical package for this purpose can be programmed to facilitate the Rasch analysis of instruments administered in surveys.

Declaration of Conflicting Interests and Ethics

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

Orcid

Sherwin Balbuena https://orcid.org/0000-0003-0183-4931

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, D.C: American Psychiatric Association.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581-594.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95-104.

Avery, L.M., Russell, D.J., Raina, P.S., Walter, S.D., & Rosenbaum, P.L. (2003). Rasch analysis of the Gross Motor Function Measure: validating the assumptions of the Rasch model to create an interval-level measure. *Archives of Physical Medicine and Rehabilitation*, *84*(5), 697-705.

Balsamo, M., Giampaglia, G., & Saggino, A. (2014). Building a new Rasch-based self-report inventory of depression. *Neuropsychiatric Disease and Treatment*, *10*, 153.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*(6), 561-571.

Bond, T.G., & Fox, C.M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Boyle, G.J. (1985). Self-report measures of depression: some psychometric considerations. *British Journal of Clinical Psychology, 24*, 45–59.

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Deb, S., Banu, P.R., Thomas, S., Vardhan, R.V., Rao, P.T., & Khawaja, N. (2016). Depression among Indian university students and its association with perceived university academic environment, living arrangements and personal issues. *Asian Journal of Psychiatry, 23*, 108-117.

Forkmann, T., Gauggel, S., Spangenberg, L., Brähler, E., & Glaesmer, H. (2013). Dimensional assessment of depressive severity in the elderly general population: Psychometric evaluation of the PHQ-9 using Rasch Analysis. *Journal of Affective Disorders, 148*(2-3), 323-330.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Gesinde, A.M., & Sanu, O.J. (2014). Prevalence and gender difference in self-reported depressive symptomatology among Nigerian university students: Implication for depression counselling. *The Counsellor, 33*(2), 129-140.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stoufer, L. Guttman, E.A. Suchman, P.L. Lazarsfeld, S.A. Star, and J.A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. IV. Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Habibi, M., Khawaja, N.G., Moradi, S., Dehghani, M., & Fadaei, Z. (2014). University student depression inventory: Measurement model and psychometric properties. *Australian Journal of Psychology, 66*(3), 149-157.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23*(1), 56-62.

Hankin, B.L. (2006). Adolescent depression: Description, causes, and interventions. *Epilepsy and Behavior, 8*(1), 102-114.

Hyde, J.S., Mezulis, A.H., & Abramson, L.Y. (2008). The ABCs of depression: Integrating affective, biological, and cognitive models to explain the emergence of the gender difference in depression. *Psychological Review, 115*(2), 291-313.

Jeong, H.J., & Lee, W.C. (2016). The level of collapse we are allowed: Comparison of different response scales in Safety Attitudes Questionnaire. *Biometrics and Biostatistics International Journal, 4*(4), 1-7.

Khawaja, N.G., & Bryden, K.J. (2006). The development and psychometric investigation of the University Student Depression Inventory. *Journal of Affective Disorders, 96*(1-2), 21-29.

Khawaja, N.G., Santos, M.L.R., Habibi, M., & Smith, R. (2013). University students' depression: A cross-cultural investigation. *Higher Education Research and Development, 32*(3), 392-406.

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement, 75*(3): 389-405.

Lailo, J.M.A. (2018). *Determinants of depressive symptoms in undergraduate UPLB students: A joint correspondence analysis*. Institute of Statistics, UPLB.

Lee, R.B., Maria, M.S., Estanislao, S., & Rodriguez, C. (2013). Factors associated with depressive symptoms among Filipino university students. *PloS One, 8*(11): e79825.

Linacre, J.M. (1997). *Guidelines for rating scales MESA Research Note #2*. Available at http://www.rasch.org/rn2.htm.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J.M., & Wright, B.D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*(2), 350.

Lim, G.Y., Tam, W.W., Lu, Y., Ho, C.S., Zhang, M.W., & Ho, R.C. (2018). Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific Reports*, *8*(1), 2861.

Lovibond, P.F., & Lovibond, S.H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335-343.

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. Available at http://www.jstatsoft.org/v20/i09 .

Maloney, P., Grawitch, M.J., & Barber, L.K. (2011). Strategic item selection to reduce survey length: Reduction in validity? *Consulting Psychology Journal: Practice and Research*, *63*, 162-175.

Marcus, M., Yasamy, M.T., Van Ommeren, M., Chisholm, D., & Saxena, S. (2012). *Depression: A Global Public Health Concern*. Geneva: World Health Organization. Available at http://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Mikolajczyk, R.T., Maxwell, A.E., El Ansari, W., Naydenova, V., Stock, C., Ilieva, S., ..., & Nagyova, I. (2008). Prevalence of depressive symptoms in university students from Germany, Denmark, Poland and Bulgaria. *Social Psychiatry and Psychiatric Epidemiology*, *43*(2), 105-112.

Nord, M. (2014). *Introduction to Item Response Theory Applied to Food Security Measurement: Basic Concepts, Parameters, and Statistics*. Technical Paper. Rome: FAO. Available at http://www.fao.org/economic/ess/ess-fs/voices/en

O'Connell, M.E., Boat, T., & Warner, K.E. (Eds.). (2009). *Committee on the prevention of mental disorders and substance abuse among children, youth, and young adults: Research advances and promising interventions. Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities.* National Academies Press.

Olsen, L.R., Jensen, D.V., Noerholm, V., Martiny, K., & Bech, P. (2003). The internal and external validity of the Major Depression Inventory in measuring severity of depressive states. *Psychological Medicine*, *33*(2), 351-356.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385-401.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. Chapters V-VII, X.

Romaniuk, M., & Khawaja, N.G. (2013). University Student Depression Inventory (USDI): Confirmatory factor analysis and review of psychometric properties. *Journal of Affective Disorders*, *150*(3), 766-775.

Sharif, A.R., Ghazi-Tabatabaei, M., Hejazi, E., Askarabad, M.H., & Dehshiri, G.R. (2011). Confirmatory factor analysis of the University Student Depression Inventory (USDI). *Procedia-Social and Behavioral Sciences*, *30*, 4-9.

Shea, T.L., Tennant, A., & Pallant, J.F. (2009). Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). *BMC Psychiatry, 9*(1), 1-10.

Smith, R.M. (2000). Fit analysis in latent trait measurement models. *Journal of applied Measurement, 1*(2), 199-218.

Spitzer, R.L., Kroenke, K., Williams, J.B., & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA, 282*(18), 1737-1744.

Stansbury, J.P., Ried, L.D., & Velozo, C.A. (2006). Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES–D) scale. *Journal of Personality Assessment, 86*(1), 10-22.

Stanton, J.M., Sinar, E.F., Balzer, W.K., & Smith, P.C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*, 167-193.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Tennant, A. (2004). Disordered thresholds: An example from the functional independence measure. *Rasch Measurement Transactions, 17*(4), 945-948

Tennant, A., & Conaghan, P.G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper?. *Arthritis Care and Research, 57*(8), 1358-1362.

UPLB INSTAT. (2018). *Utak at Puso: A Survey on the Mental Health Status of UPLB Students (STAT 173 Survey)*. University of the Philippines Los Baños, Laguna.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wongpakaran, N., Wongpakaran, T., & Kuntawong, P. (2019). Evaluating hierarchical items of the geriatric depression scale through factor analysis and item response theory. *Heliyon, 5*(8), e02300.

World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. Geneva: Author. http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1

Wright B.D., & Linacre, J.M. (1987). Dichotomous Rasch model derived from specific objectivity. *Rasch Measurement Transactions, 1*(1), 5-6

Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis*. Chicago, IL: University of Chicago, MESA Press.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*(1), 23-48.

Yu, C.H. (2011). *A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling*. Available at www.creative-wisdom.com/computer/sas/IRT.pdf.

Zigmond, A.S., & Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67*(6), 361-370.

Zung, W.W. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12*(1), 63-70.

APPENDIX A: UNIVERSITY STUDENT DEPRESSION INVENTORY (USDI) ITEMS WITH FILIPINO TRANSLATION (*in italics*)

In this inventory, the student is asked to indicate how often he/she has experienced each item over the past two weeks by responding in the following 5-point Likert scale: not at all (*hindi kailanman*), rarely (*bihira*), sometimes (*minsan*), most of the time (*madalas*), and all the time (*palagi*).

The prefix of the code indicates the subscale to which the item belongs. Hence, LG, AM, and CE refer to the Lethargy, Academic Motivation, and Cognitive/Emotional subscales, respectively.

Item (*Italics in Filipino*)	Code
1. I am more tired than I used to be. *Ako ay mas pagod ngayon kung ikukumpara sa dati.*	LG01
2. I wonder whether life is worth living. *Napapaisip ako kung may halaga pa bang mabuhay.*	CE02
3. I do not have any desire to go to lectures. *Wala na akong pagnanais na pumasok sa klase.*	AM03
4. I do not have the energy to study at my usual level. *Wala na akong ganang mag-aral gaya ng dati.*	LG04
5. I feel worthless. *Nararamdaman ko na ako ay walang halaga.*	CE05
6. I don't attend lectures as much as I used to. *Mas madalang na ako pumasok sa klase kaysa dati*.	AM06
7. I have thought about killing myself. *Sumagi sa aking isipan na magpakamatay.*	CE07
8. I don't feel motivated to study. *Wala akong motibasyon na mag-aral.*	AM08
9. My energy is low. *Wala akong gana.*	LG09
10. No one cares about me. *Walang nagmamalasakit sa akin.*	CE10
11. I feel emotionally empty. *Wala na akong nararamdamang kahit anong emosyon.*	CE11
12. Going to university is pointless. *Hindi ko nakikita ang kahalagahan ng pagpasok sa unibersidad*	AM12
13. I find it hard to concentrate. *Nahihirapan akong magpokus.*	LG13
14. I feel sad. *Nalulungkot ako*	CE14
15. I worry I will not amount to anything. *Nangangamba akong wala akong mararating sa buhay.*	CE15
16. I don't feel rested even after sleeping.	LG16

Hindi ko ramdam na ako ay nakapagpahinga kahit ako ay nakatulog na.	
17. The activities I used to enjoy no longer interest me. *Nawawalan na ako ng gana sa mga bagay na dating interesado ako.*	CE17
18. Challenges I encounter in my studies overwhelm me. *Nilalamon ako ng mga kinakaharap kong pagsubok sa aking pag-aaral*	LG18
19. I feel like I cannot control my emotions. *Pakiramdam ko, hindi ko na kontrolado ang aking emosyon*	CE19
20. I spend more time alone than I used to. *Napapadalas ang aking pag-iisa.*	CE20
21. My mood affects my ability to carry out assigned tasks. *Nakakaapekto ang aking mga emosyon sa aking abilidad na isagawa ang mga gawaing naiatas sa akin.*	LG21
22. I feel disappointed in myself. *Nakakaramdam ako ng pagkabigo sa aking sarili.*	CE22
23. I have trouble starting assignments. *Nahihirapan akong simulan ang aking mga takdang-aralin.*	AM23
24. Daily tasks take me longer than they used to. *Mas matagal kong maisagawa ang mga pangaraw-araw na gawain kaysa sa nakasanayan.*	LG24
25. I feel withdrawn when I'm around others. *Nakakaramdam ako ng hindi pagkabilang kapag napapaligiran ako ng mga tao.*	CE25
26. I do not cope well. *Hindi na ako makasabay nang maayos.*	CE26
27. I do not find study as interesting as I used to. *Hindi na ako interesadong mag-aral kaysa sa nakasanayan.*	AM27
28. My study is disrupted by distracting thoughts. *Ang aking pag-aaral ay naaantala ng mga nakakaabalang mga saloobin.*	LG28
29. I think most people are better than me. *Sa tingin ko, karamihan sa mga tao ay mas magaling kaysa sa akin.*	CE29
30. I have trouble completing study tasks. *Nahihirapan akong tapusin ang mga gawain ukol sa pag-aaral.*	AM30

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 279–302

https://doi.org/10.21449/ijate.1238248

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

Using ACER ConQuest program to examine multidimensional and many-facet models

Mahmut Sami Koyuncu[1,*], **Mehmet Şata**[2]

[1]Afyon Kocatepe University, Faculty of Education, Department of Educational Sciences, Afyonkarahisar, Türkiye
[2]Ağrı İbrahim Çeçen University, Faculty of Education, Department of Educational Sciences, Ağrı, Türkiye

ARTICLE HISTORY

Received: Jan. 17, 2023
Revised: May 25, 2023
Accepted: Jan. 12, 2023

Keywords:
ConQuest,
Multidimensionality,
Many-facet Rasch,
Validity.

Abstract: The main aim of this study was to introduce the ConQuest program, which is used in the analysis of multivariate and multidimensional data structures, and to show its applications on example data structures. To achieve this goal, a basic research approach was applied. Thus, how to use the ConQuest program and how to prepare the data set for analysis were explained step by step. Then, two example applications were made considering the multidimensional structures. Finally, different sources of variability (e.g., item, student, rater, gender), which are both multidimensional and independent of each other, were performed by considering different sources of variability together. According to the analyses, the dimensionality of the data structures must be examined in the analysis process. If the data structure is multidimensional, appropriate multidimensional IRT analyses should be performed.

1. INTRODUCTION

1.1. What is ACER ConQuest?

The ACER ConQuest (Adams et al., 2020) program is developed at the Australian Council for Educational Research (ACER) and the University of California, Berkeley. It is a paid statistical package program that can examine the fit of item response and latent regression models, including multidimensional item response models, in a single program. It also provides the integration of item responses and regression analysis (Adams et al., 2022).

The ConQuest can run the analyses of the following models (Adams et al., 2022): Rasch Simple Logistic Model (Rasch, 1980), Rating Scale Model (Andrich, 1978), Partial Credit Model (Masters, 1982), Ordered Partition Model (Wilson, 1992), Linear Logistic Test Model (Fischer, 1983), Many-facet Models (Linacre, 1994), Generalized Unidimensional Models, Multidimensional Item Response Models (Adams et al., 1997; Wang, 1995), and Latent Regression Models (Adams et al., 1997).

Joint maximum likelihood (JML) or marginal maximum likelihood (MML) estimates can be generated by ACER ConQuest generates for the parameters of the specified models. The MML estimation algorithms used are;

*CONTACT: Mahmut Sami Koyuncu ✉ koyuncumsami@gmail.com Afyon Kocatepe University, Faculty of Education, Department of Educational Sciences, Afyonkarahisar, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

- Gauss-Hermite quadrature (Volodin & Adams, 1995).
- Bock/Aitken quadrature (Bock & Aitkin, 1981)
- Monte Carlo (Volodin & Adams, 1995).
- Markov chain Monte Carlo (Patz & Junker, 1999).

The Gauss method is generally used for three or less-dimensional problems, while the Monte Carlo method is preferred for more than three-dimensional problems. Moreover, the Gauss method cannot be used when the distribution is discrete or there is no independent variable to estimate the dependent variable in the regression model. Thus, when there is a regression variable in the Conquest program, the Quadrature method is used as the default method. Otherwise, the Gauss method is used. If there is missing data in all items in a dimension, the JML method cannot be used. In addition, estimating item parameters is not possible when the JML method is used (Adams et al., 2022).

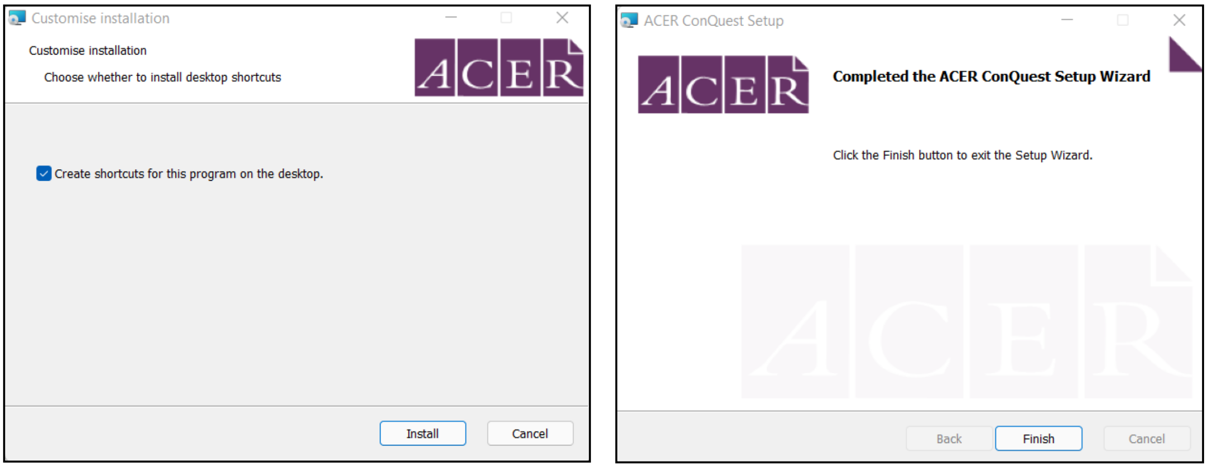
Through the ConQuest program, the following applications can be performed: item analysis (IRT and Traditional), DIF, Exploring Rater Effects, Latent correlation estimation and Estimating Latent Correlations and Testing Dimensionality, and Drawing Plausible Values (Adams et al., 2022). ACER ConQuest can model up to 50 different facets and analyze item clusters designed to produce measurements of up to 30 latent dimensions.

1.2. Installation and User Interfaces of ACER ConQuest Program

The program can be used in both Windows and Mac OS operating systems. For the Windows operating system, the program has both GUI (graphical user interface) and CMD (console interface) console versions. However, for the Mac OS system, only the console version is available to use. The GUI version is more user-friendly and has drawing functions that the console version does not have. However, especially for larger and more complex analyses, the console version, which works faster than the GUI version, may be preferred.

The ConQuest program has a 1-month free DEMO trial for users to experience. However, in the demo version, the sample size is limited to 3000 and the number of items is limited to 100. The installation of the program is simple. Figure 1 shows the program setup screens for the Windows operating system.

Figure 1. *ConQuest Windows setup screen.*



Customise installation

Customise installation

Choose whether to install desktop shortcuts

☑ Create shortcuts for this program on the desktop.

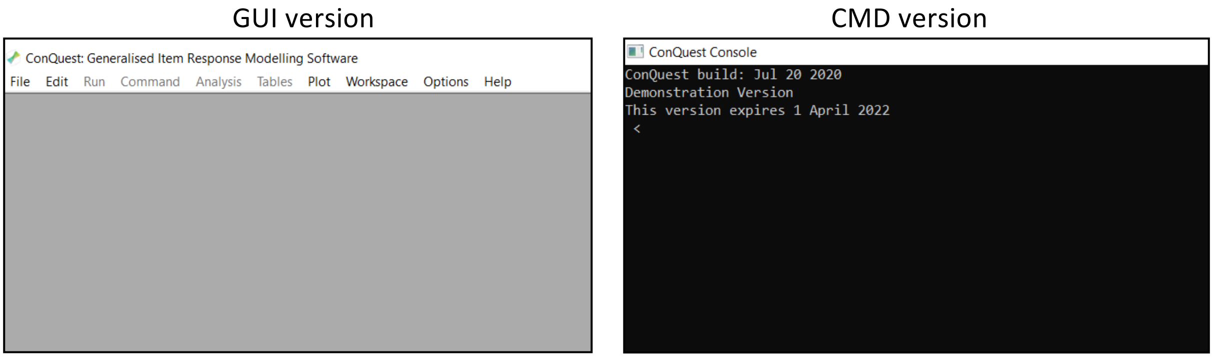
ACER ConQuest Setup

Completed the ACER ConQuest Setup Wizard

Click the Finish button to exit the Setup Wizard.

Once the program is installed on the PC, both GUI and CMD versions become ready to use. Figure 2 shows the start screens of the GUI and CMD versions of the ConQuest program.

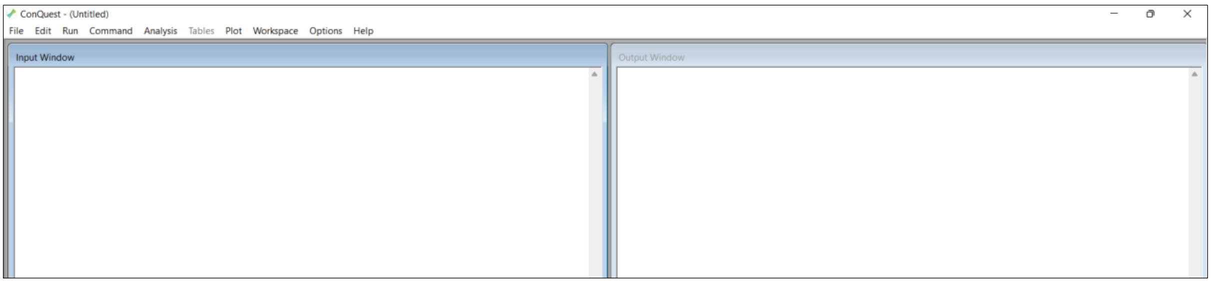
Figure 2. *Start screens of GUI and CMD versions.*


GUI version
CMD version
ConQuest: Generalised Item Response Modelling Software
File Edit Run Command Analysis Tables Plot Workspace Options Help
ConQuest Console

```
ConQuest build: Jul 20 2020
Demonstration Version
This version expires 1 April 2022
 <
```

The GUI version includes menus such as *File*, *Edit*, *Run*, *Command*, *Analysis*, *Tables*, *Plot*, *Workspace*, *Options,* and *Help*. With the New command in the File menu, a new working screen including both input and output windows is opened (Figure 3). Analysis using a command line (i.e., CMD version) can be performed in one step. when running from a command-line interpreter is to provide the command file as a command-line argument. In this demonstration, the GUI version was used.

Figure 3. *ConQuest Input and Output screen.*


ConQuest - (Untitled)
File Edit Run Command Analysis Tables Plot Workspace Options Help
Input Window
Output Window

The codes (i.e., syntax) required for analysis are entered in the *Input Window*. The analysis is performed by running the codes (i.e., syntax) with the help of the Run menu. The analyses performed are displayed in the *Output Window* section. Besides, the analysis results specially requested by the researcher can be saved as *.txt* files with the help of the syntaxes entered in the code file. For more detailed information, the ACER ConQuest Manual can be applied (Adams et al., 2022).

In educational and psychological research, measurement tools are the main data collection sources. In educational sciences, the measurements are done indirectly, so it is important to provide evidence regarding the validity and reliability of the measures (Köse, 2012). The selection of the analysis method and package program appropriate for the nature of the data are the factors that contribute to the reliability and validity of the scores obtained from the measurement tools.

Regarding the nature of the data used for educational purposes, the concept of dimensionality is important (Finch & Habing, 2003; Mroch & Bolt, 2006; Özbek-Baştuğ, 2012; Özer-Özkan & Acar-Güvendir, 2014). While providing evidence for the reliability and validity of the scores obtained from the measurement tools, determining the dimensionality of the data or the number of dimensions/factors will contribute to the reliability and validity. According to Messick (1995), there are two threats to validity: construct underrepresentation and construct-irrelevant variance. An accurate definition of the studied data set in terms of dimensionality will directly contribute to the validity of the measurements. This is due to the concern of underrepresenting the structure intended to be measured is eliminated (Messick, 1995).

Regarding the historical development of measurement theories, the classical test theory, which was founded on the assumption of total score or unidimensionality, was first put forward. Then, the one-dimensional item response theory emerged. Therefore, these models investigate one-dimensional constructs and variables. This means that the unidimensionality of the structures planned to be measured must be tested (Özbek-Baştuğ, 2012).

Various methods determine the dimensionality of the constructs or the number of dimensions. These methods are either parametric or non-parametric. Research has compared these methods with each other (Mroch & Bolt, 2006) to identify the most effective method (Stout et al., 2001). In addition to, some studies have focused only on dimensionality analysis (Jang & Roussos, 2007).

In studies on dimensionality, item (individual) and ability parameters were negatively affected because multidimensional structures were analyzed as one-dimensional structures (Özer-Özkan, 2012). Regarding the structures of the measurement tools used in the measurement of cognitive skills, it is difficult to provide the unidimensionality assumption. Considering that many skills are used together in the measurement of high-level cognitive skills, it confronts us with the fact that the unidimensionality assumption will not be met. This situation requires multidimensional analysis or modeling of the measurement tool (Ackerman, 1994).

Multidimensional modeling and analysis methods attract more and more attention day by day, as they eliminate the limitations of one-dimensional measurement models and offer models that are more suitable for real-life situations. Due to the increasing need for measuring multidimensional structures, many statistical package programs have been developed recently (Köse, 2012). These programs include IRTPRO, MULTILOG, BILOG, MIRTE, TESTFACT, PARSCALE, Xcalibre, and R package programs (i.e., eRm, pl. rasch), flexMIRT, BMIRT, and NOHARM. Almost all these programs can analyze both one-dimensional and multidimensional measurement models. Besides, the ConQuest package program can analyze both multidimensional measurement models and multivariate measurements at the same time and also allows the interactions between variables to be examined at the same time.

ACER Conquest program is frequently used in studies in many different fields recently. Its more widespread use is preferred especially in IRT model analysis, mostly in Rasch model estimations (Brnic & Greefrath, 2021; Hahn & Kähler, 2022; Jolin & Wilson, 2022; Jüttler & Schumann, 2022; Krell et al., 2022; Koch et al., 2022; Lou et al., 2022; Mischo et al., 2022; Oko, 2022; Osterhaus et al., 2022, Spink et al., 2022; Unfried et al., 2022; Wall et al., 2022). Besides, the use of the Conquest program is preferred in studies where multidimensional structures are examined or many-facet models are used (Bartolomé & Garaizar, 2022; Mendoza et al., 2022; Wang et al., 2022; Zhai, 2022).

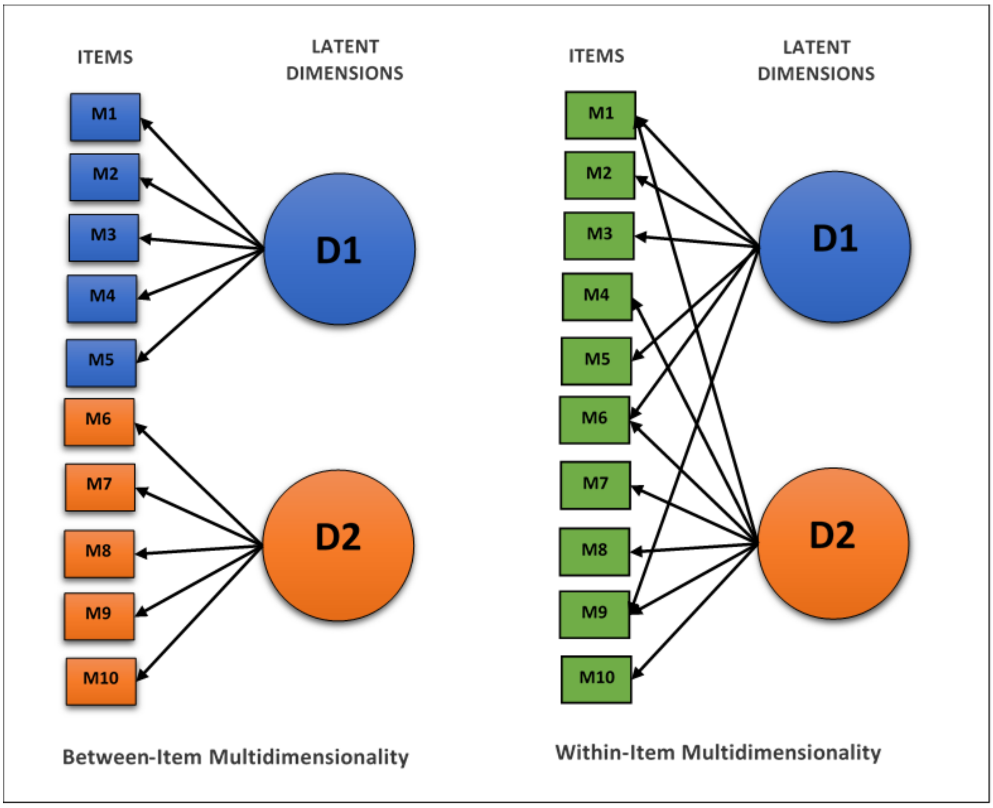
2. METHOD

2.1. Application of The Conquest Program

This research aimed to introduce the ConQuest program, applying multidimensional models on example data sets. Therefore, firstly, the general features of the ConQuest program, its installation, and the analysis process were explained. Then, multidimensional model applications were carried out on example data sets. In addition, example syntaxes appropriate for multidimensional model analysis were created to benefit the researchers. Especially considering that many structures are multidimensional by nature, this research is important in terms of eliminating the lack of multidimensional models in the literature.

By creating synthetic data on multidimensional models, three different examples scenarios were presented within the scope of the study in order to guide the researchers. Analyses were performed via the ConQuest GUI Demo version (5.12.3). The first example application belongs to between-item multidimensional models, and the second example application belongs to within-item multidimensional models. If a test consists of several one-dimensional subscales, it is Between-Item Multidimensionality. If any of the items are related to more than one latent dimension, this test is considered as Within-Item Multidimensionality (Adams et al., 1997; Wang, 1995). The structure of the synthetic data of the Between-Item and Within-Item multidimensional model created within the scope of the study was presented in Figure 4.

Figure 4. *Within-Item and Between-Item Multidimensionality.*



Also, an example of many-facet multidimensional models was presented as a third example application within the scope of the study.

2.2. Example 1: Between-Item Multidimensional Model

It is assumed that the data structure created for the Between-Item Multidimensional Model example consists of 10 Likert-type items scored from 1 to 5. As shown in Figure 4, items from 1 to 5 represent the first dimension of the scale and items from 6 to 10 represent the second dimension. The data used in the research consists of hypothetical data. The main reason for this is that the program is intended to be implemented and to guide researchers. This situation was taken into consideration as the limitation of the research. The data structure of 50 individuals was generated and the data file with the .txt extension required for analysis was prepared. Then, the command with the *.cqc* extension was created for analysis. It is important to prepare the data for analysis, and the structure in the data file must be defined in the script with the necessary syntaxes. In addition, the labels of the variables in the data file can be created in a separate file with a *.txt* extension to make the analysis outputs more understandable. Table 1 contains the command, data, and tag file examples created by the researchers for the Between-Item Multidimensional Model.

Table 1. *Example script, data, and tag files of the Between-Item Multidimensional Model.*

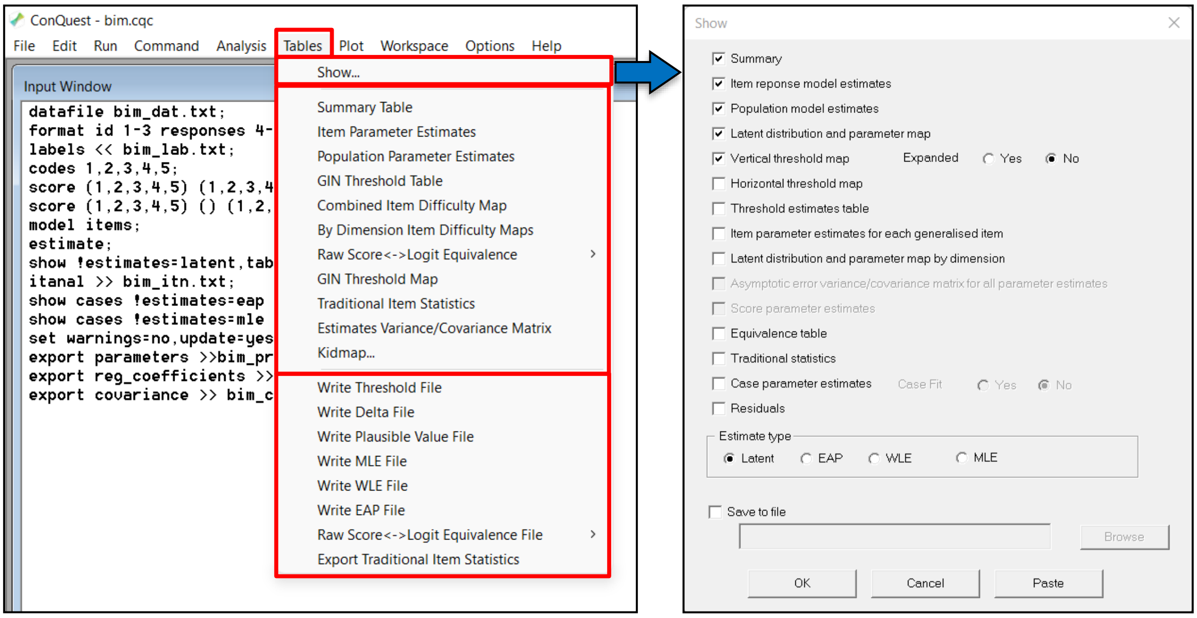
bim.cqc	bim_dat.txt	bim_lab.txt
Command statements (required)	Dataset (required)	Dataset label (optional)

`datafile bim_dat.txt;`
`format id 1-3 responses 4-13;`
`labels << bim_lab.txt;`
`codes 1,2,3,4,5;`
`score (1,2,3,4,5) (1,2,3,4,5) () ! items(1-5);`
`score (1,2,3,4,5) () (1,2,3,4,5) ! items(6-10);`
`model items;`
`estimate;`
`show !estimates=latent,tables=1:2:3:9>> bim_shw.txt;`
`itanal >> bim_itn.txt;`
`show cases !estimates=eap >> bim_eap.txt;`
`show cases !estimates=mle >> bim_mle.txt;`
	`0014555555551`
	`0025555515555`
	`0032553141311`
	`0045555555555`
	`0055555455455`
	`0065455511451`
	`0075555445555`
	`0085555553551`
	`0095555555555`
	`0105555455555`
	`0115555354551`
		`===> item`
		`1      M1`
		`2      M2`
		`3      M3`
		`4      M4`
		`5      M5`
		`6      M6`
		`7      M7`
		`8      M8`
		`9      M9`
		`10     M10`

Note. The first set of parentheses contains a set of codes (the codes list). The second set of parentheses contains a set of scores on dimension one for each of those codes (a score list). The third set contains a set of scores on dimension two (a second score list) and so on. The number of separate codes in the codes list indicates the number of response categories that will be modeled for each item. The number of score lists indicate the number of dimensions in the model. The codes and scores in the lists can be comma-delimited or space-delimited.

To perform the analysis, the command file is opened and *run* in the ConQuest program. The necessary analysis results can be added to the command file as *.txt* in the ConQuest program or created with the help of the *Tables* menus in the program after the command is run. Figure 5 displays the statistics and output files that can be created with the help of the *Show…* tab in the *Tables* menu.

Figure 5. *ConQuest Tables menu and Show tab.*



ConQuest - bim.cqc
File
Edit
Run
Command
Analysis
Tables
Plot
Workspace
Options
Help
Input Window
Show...
Summary Table
Item Parameter Estimates
Population Parameter Estimates
GIN Threshold Table
Combined Item Difficulty Map
By Dimension Item Difficulty Maps
Raw Score<->Logit Equivalence
GIN Threshold Map
Traditional Item Statistics
Estimates Variance/Covariance Matrix
Kidmap...
Write Threshold File
Write Delta File
Write Plausible Value File
Write MLE File
Write WLE File
Write EAP File
Raw Score<->Logit Equivalence File
Export Traditional Item Statistics
Show

```
datafile bim_dat.txt;
format id 1-3 responses 4-
labels << bim_lab.txt;
codes 1,2,3,4,5;
score (1,2,3,4,5) (1,2,3,4
score (1,2,3,4,5) () (1,2,
model items;
estimate;
show !estimates=latent,tab
itanal >> bim_itn.txt;
show cases !estimates=eap
show cases !estimates=mle
set warnings=no,update=yes
export parameters >>bim_pr
export reg_coefficients >>
export covariance >> bim_c
```

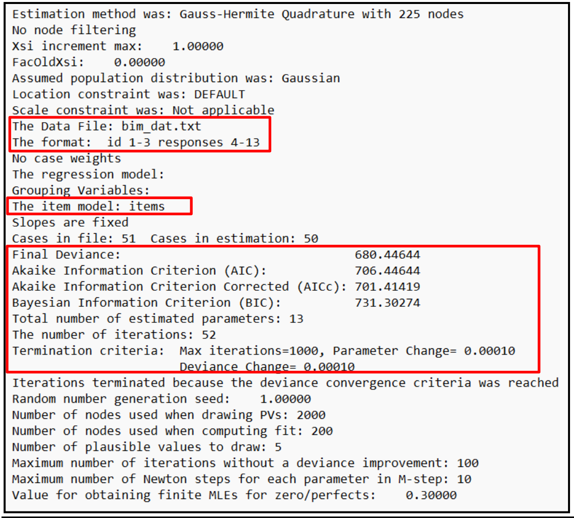
☑ Summary
☑ Item reponse model estimates
☑ Population model estimates
☑ Latent distribution and parameter map
☑ Vertical threshold map Expanded ○ Yes ● No
□ Horizontal threshold map
□ Threshold estimates table
□ Item parameter estimates for each generalised item
□ Latent distribution and parameter map by dimension
□ Asymptotic error variance/covariance matrix for all parameter estimates
□ Score parameter estimates
□ Equivalence table
□ Traditional statistics
□ Case parameter estimates Case Fit ○ Yes ● No
□ Residuals
Estimate type: ● Latent ○ EAP ○ WLE ○ MLE
□ Save to file

Rasch Analysis results for application example-1 (*bim_shw.txt*) and traditional item analysis results (*bim_itn.txt*) are presented below.

Rasch Analysis results (*bim_shw.txt*) file includes a summary of the estimation, item parameter estimates, regression coefficients, item parameter estimates for each term in the model (in this example there is only one term: item), covariance/correlation matrix, reliability coefficients, map of latent distributions, and response model parameter estimates, respectively. Table 2 presents the summary of the estimation and item parameter estimates output of the example application. Table 3 contains the universe model parameter estimations including the output of regression coefficients, covariance/correlation matrix, reliability coefficients, and brief explanations.

Table 2. *Summary of the estimation and item parameter estimates output.*

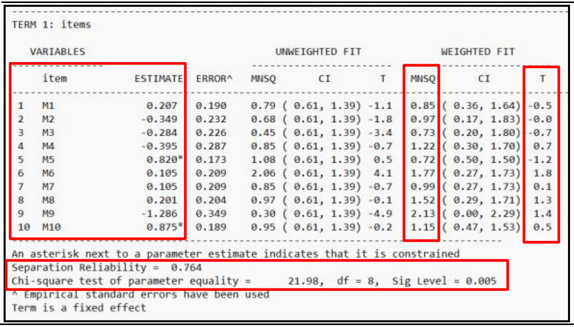
Summary of the estimation | Explanation



```
Estimation method was: Gauss-Hermite Quadrature with 225 nodes
No node filtering
Xsi increment max:    1.00000
FacOldXsi:    0.00000
Assumed population distribution was: Gaussian
Location constraint was: DEFAULT
Scale constraint was: Not applicable
The Data File: bim_dat.txt
The format:  id 1-3 responses 4-13
No case weights
The regression model:
Grouping Variables:
The item model: items
Slopes are fixed
Cases in file: 51  Cases in estimation: 50
Final Deviance:                                 680.44644
Akaike Information Criterion (AIC):             706.44644
Akaike Information Criterion Corrected (AICc): 701.41419
Bayesian Information Criterion (BIC):           731.30274
Total number of estimated parameters: 13
The number of iterations: 52
Termination criteria:  Max iterations=1000, Parameter Change= 0.00010
                       Deviance Change= 0.00010
Iterations terminated because the deviance convergence criteria was reached
Random number generation seed:    1.00000
Number of nodes used when drawing PVs: 2000
Number of nodes used when computing fit: 200
Number of plausible values to draw: 5
Maximum number of iterations without a deviance improvement: 100
Maximum number of Newton steps for each parameter in M-step: 10
Value for obtaining finite MLEs for zero/perfects:    0.30000
```

In the summary of the estimation results, there is information such as the analysed data file, the format in the data file, the desired model, and example size. Besides, Deviance, AIC, AICc and BIC values used to evaluate model fit are also included. Regarding the relative fit, (for example, which model fits better compared to more than one model), the smaller value fits the data better (Chen et al., 2013; De Ayala, 2009, p. 41).

Rasch Model Item parameter estimates | Explanation



```
--------------------------------------------------------------------------------
TERM 1: items

  VARIABLES                              UNWEIGHTED FIT            WEIGHTED FIT
---------------                    ----------------------------  --------------------
    item            ESTIMATE  ERROR^  MNSQ      CI          T    MNSQ      CI        T
--------------------------------------------------------------------------------------
1   M1                0.207   0.190   0.79 ( 0.61, 1.39) -1.1   0.85 ( 0.36, 1.64) -0.5
2   M2               -0.349   0.232   0.68 ( 0.61, 1.39) -1.8   0.97 ( 0.17, 1.83) -0.0
3   M3               -0.284   0.226   0.45 ( 0.61, 1.39) -3.4   0.73 ( 0.20, 1.80) -0.7
4   M4               -0.395   0.287   0.85 ( 0.61, 1.39) -0.7   1.22 ( 0.30, 1.70)  0.7
5   M5                0.820*  0.173   1.08 ( 0.61, 1.39)  0.5   0.72 ( 0.50, 1.50) -1.2
6   M6                0.105   0.209   2.06 ( 0.61, 1.39)  4.1   1.77 ( 0.27, 1.73)  1.8
7   M7                0.105   0.209   0.85 ( 0.61, 1.39) -0.7   0.99 ( 0.27, 1.73)  0.1
8   M8                0.201   0.204   0.97 ( 0.61, 1.39) -0.1   1.52 ( 0.29, 1.71)  1.3
9   M9               -1.286   0.349   0.30 ( 0.61, 1.39) -4.9   2.13 ( 0.00, 2.29)  1.4
10  M10               0.875*  0.189   0.95 ( 0.61, 1.39) -0.2   1.15 ( 0.47, 1.53)  0.5
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.764
Chi-square test of parameter equality =      21.98,  df = 8,  Sig Level = 0.005
^ Empirical standard errors have been used
Term is a fixed effect
```

There are item parameter estimations for each term in the model. The example presents only one term (*item*). Since the example has 10 items in total, there are some concordance statistics including the estimations and standard errors for each item.

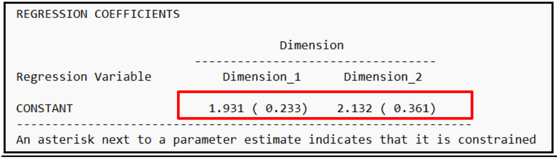
According to the item parameter estimations of the Rasch model in Table 2, the Mean-Square (MNSQ) value and the confidence interval of this value are included as the fit index. If the calculated MNSQ fit value is outside the expected confidence interval, the T statistic corresponding to the MNSQ value will exceed the |2.0|, meaning that the item does not fit the model well (Adams et al., 2022).

The parameter estimates in this table are for the difficulties of each of the items. For the purposes of model identification, ACER ConQuest constrains the difficulty estimate for the last item to ensure an average difficulty of zero. This constraint has been achieved by setting the difficulty of the last item to be the negative sum of the previous items. The fact that this item is constrained is indicated by the asterisk (*) placed next to the parameter estimate.

At the bottom of the item parameter output of the Rasch model, there are Separation Reliability and Chi-square values. Separation reliability (Wright & Stone, 1979) takes a value between 0 and 1 and is an index of equality of parameters. It provides information on how well its parameters are separated. However, it may not be useful to examine the chi-square value in all cases; it will be more useful to examine the significance of parameter equality, especially in cases such as rater severity (Adams et al., 2022).

Table 3. *Population Model Parameter Estimates.*

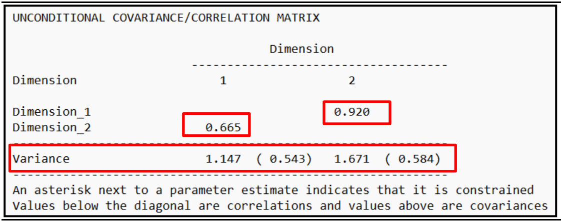
Regression coefficients | Explanation



```
REGRESSION COEFFICIENTS

                                   Dimension
                         -------------------------------
Regression Variable          Dimension_1      Dimension_2

CONSTANT                    1.931 ( 0.233)    2.132 ( 0.361)
-------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
```

These are average ability estimates in each dimension.

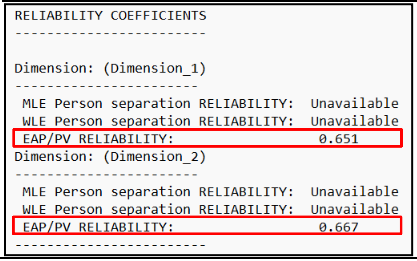
Covariance/correlation matrix | Explanation



```
UNCONDITIONAL COVARIANCE/CORRELATION MATRIX

                                   Dimension
                         ---------------------------------
Dimension                    1                   2

Dimension_1                                    0.920
Dimension_2                0.665

Variance                   1.147   ( 0.543)   1.671   ( 0.584)
-----------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Values below the diagonal are correlations and values above are covariances
```

Correlation values (0.665) and covariance values (0.920) between dimensions are included. Also, there are the estimated variance values for the two dimensions

Reliability coefficients | Explanation



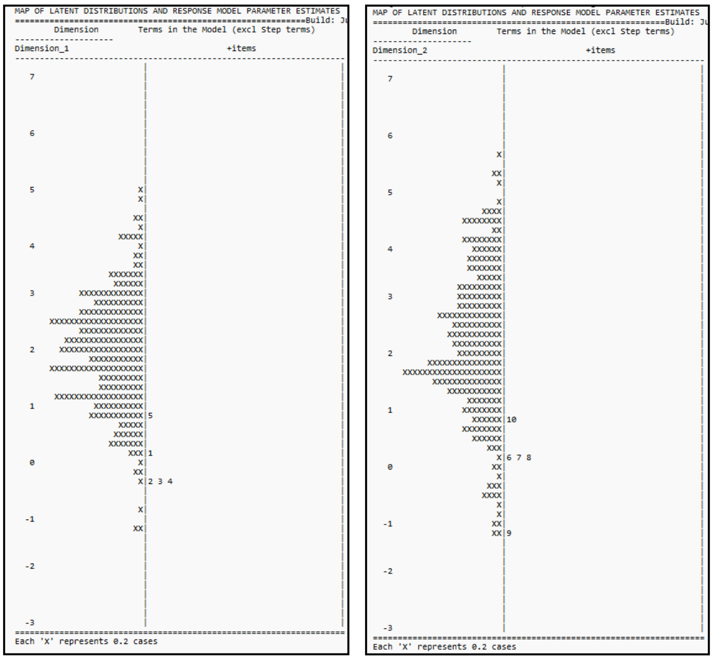
```
RELIABILITY COEFFICIENTS
------------------------

Dimension: (Dimension_1)
------------------------
 MLE Person separation RELIABILITY:  Unavailable
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:                   0.651
Dimension: (Dimension_2)
------------------------
 MLE Person separation RELIABILITY:  Unavailable
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:                   0.667
------------------------
```

There are reliability values for each dimension. There are three different reliability values. Since the calculation is not made according to the maximum likelihood estimation, only the EAP/PV value is displayed, while the other values are not.

Besides, item difficulty map output, which provides the opportunity to examine latent ability estimation and item difficulty on the same scale, and which can be obtained separately or in combination according to the dimensions, can also be created. Figure 6 displays the output of the latent distribution map and response model parameter estimations, and Figure 7 shows the output of generalized item threshold tables and maps.

Figure 6. *Map of latent distributions and response model parameter estimates output.*

Item difficulty map by dimension



Combined item difficulty map

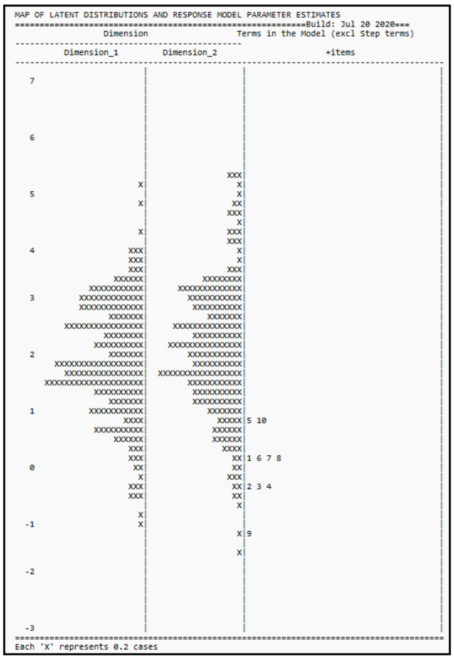
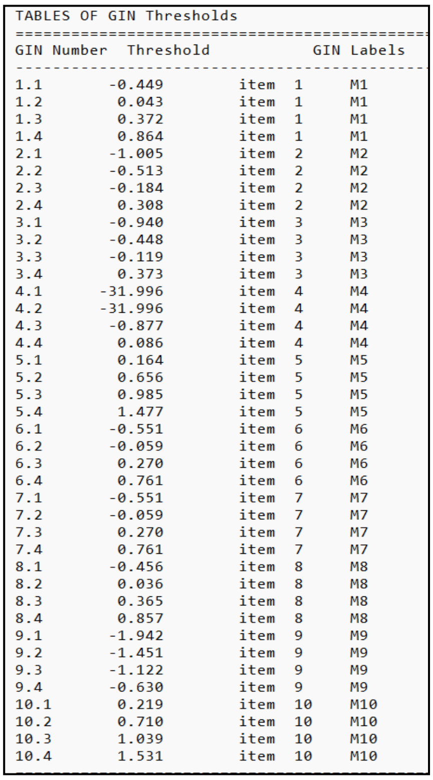


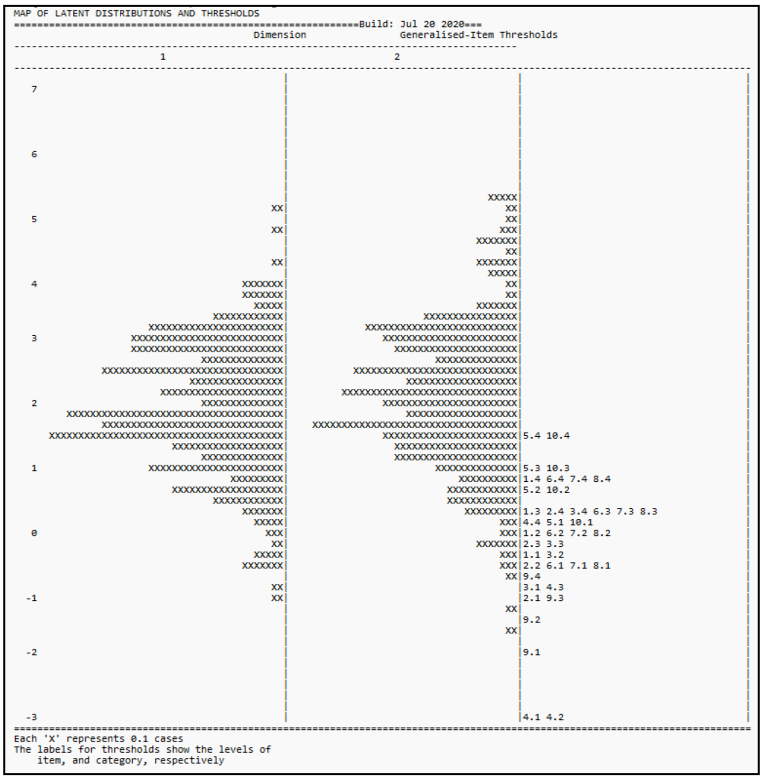
Figure 6 shows that item difficulty maps can be created separately or combined as a single output according to dimensions. In the first column, the latent ability estimation distribution of the individuals, and in the second column, the difficulty estimations of the items are located. This distribution map allows for each independent variable to be interpreted as a dependent variable by placing many sources of variability such as student, item, size, rater, and time on the same scale (logit scale) (Esfandiari, 2015). The ability estimations of the individuals in the example data set vary between -1 and +5 logit in the first dimension, they vary between -1 and +6 logit in the second dimension, and the group generally has a high ability in both dimensions. Regarding the figure in which both dimensions are combined, individuals exhibit a slightly higher ability in the second dimension. On the other hand, the difficulty levels of the items indicate that the most difficult items are item number five in the first sub-dimension and item number 10 in the second sub-dimension. According to the variable map, in which both dimensions were combined, was examined, item 9 was found the easiest. In addition, items 2, 3, and 4 had similar difficulty levels, and items 1, 6, 7, and 8 were similar to each other.

The combined item difficulty map of the first and second dimensions does not mean that they are on the same scale and comparable. The researcher should pay attention to this issue while reporting his results.

Figure 7. *Generalized item thresholds table and map.*



```
TABLES OF GIN Thresholds
================================================
GIN Number   Threshold            GIN Labels
------------------------------------------------
1.1          -0.449        item  1      M1
1.2           0.043        item  1      M1
1.3           0.372        item  1      M1
1.4           0.864        item  1      M1
2.1          -1.005        item  2      M2
2.2          -0.513        item  2      M2
2.3          -0.184        item  2      M2
2.4           0.308        item  2      M2
3.1          -0.940        item  3      M3
3.2          -0.448        item  3      M3
3.3          -0.119        item  3      M3
3.4           0.373        item  3      M3
4.1         -31.996        item  4      M4
4.2         -31.996        item  4      M4
4.3          -0.877        item  4      M4
4.4           0.086        item  4      M4
5.1           0.164        item  5      M5
5.2           0.656        item  5      M5
5.3           0.985        item  5      M5
5.4           1.477        item  5      M5
6.1          -0.551        item  6      M6
6.2          -0.059        item  6      M6
6.3           0.270        item  6      M6
6.4           0.761        item  6      M6
7.1          -0.551        item  7      M7
7.2          -0.059        item  7      M7
7.3           0.270        item  7      M7
7.4           0.761        item  7      M7
8.1          -0.456        item  8      M8
8.2           0.036        item  8      M8
8.3           0.365        item  8      M8
8.4           0.857        item  8      M8
9.1          -1.942        item  9      M9
9.2          -1.451        item  9      M9
9.3          -1.122        item  9      M9
9.4          -0.630        item  9      M9
10.1          0.219        item  10     M10
10.2          0.710        item  10     M10
10.3          1.039        item  10     M10
10.4          1.531        item  10     M10
------------------------------------------------
```



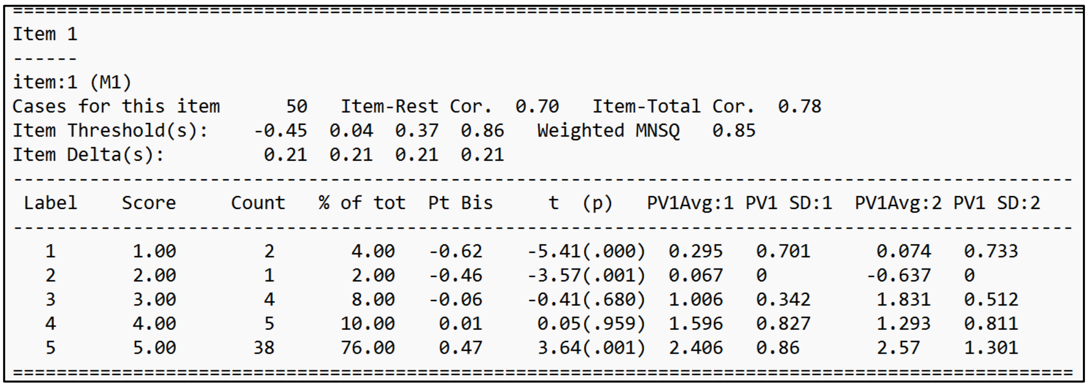
```
MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS
=========================================================Build: Jul 20 2020===
                                     Dimension              Generalised-Item Thresholds
------------------------------------------------------------------------------
                      1                                  2
```

```
Each 'X' represents 0.1 cases
The labels for thresholds show the levels of
    item, and category, respectively
```

In Figure 7, there is a map that includes the thresholds (Thurstonian) estimation values for each item and the latent talent estimations of these values along with the dimensions. The example application has 4 thresholds since the Likert-type items scored from 1 to 5. Among the values in the rightmost column, the first term indicates the item number, and the second term indicates the threshold value. For example, 5.4 represents the 4th threshold of the 5th item.

The traditional item analysis result (*bim_itn.txt*) file involves statistics and test statistics for each item. In Figure 8, there are only results for item 1, and Figure 9 shows the test statistics results.

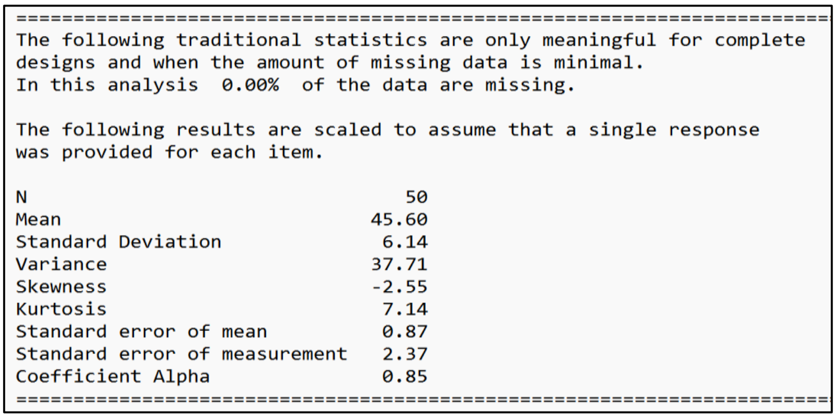
Figure 8. *Statistics related to item 1.*



```
======================================================================================================
Item 1
------
item:1 (M1)
Cases for this item      50   Item-Rest Cor.  0.70   Item-Total Cor.  0.78
Item Threshold(s):    -0.45  0.04  0.37  0.86   Weighted MNSQ   0.85
Item Delta(s):         0.21  0.21  0.21  0.21
------------------------------------------------------------------------------------------------------
 Label     Score     Count   % of tot  Pt Bis     t  (p)    PV1Avg:1 PV1 SD:1  PV1Avg:2 PV1 SD:2
------------------------------------------------------------------------------------------------------
   1        1.00       2       4.00   -0.62    -5.41(.000)  0.295    0.701      0.074    0.733
   2        2.00       1       2.00   -0.46    -3.57(.001)  0.067    0         -0.637    0
   3        3.00       4       8.00   -0.06    -0.41(.680)  1.006    0.342      1.831    0.512
   4        4.00       5      10.00    0.01     0.05(.959)  1.596    0.827      1.293    0.811
   5        5.00      38      76.00    0.47     3.64(.001)  2.406    0.86       2.57     1.301
======================================================================================================
```

Figure 8 shows the distribution of the score categories of item 1, the point biserial correlation coefficient and t value, and the MNSQ fit value. In addition, since it is a multi-category item, item threshold and delta values are also included.

Figure 9. *Conventional test statistics.*

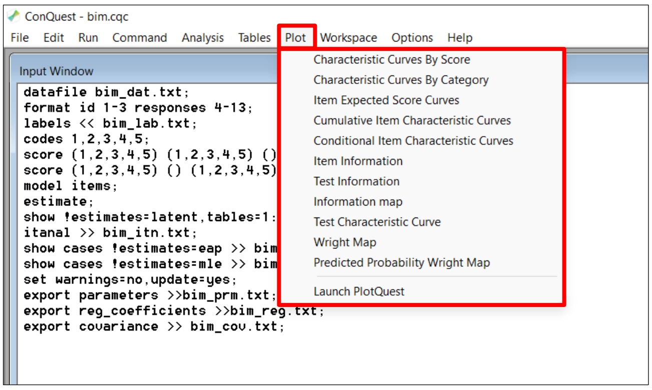


```
=======================================================================
The following traditional statistics are only meaningful for complete
designs and when the amount of missing data is minimal.
In this analysis  0.00%  of the data are missing.

The following results are scaled to assume that a single response
was provided for each item.

N                                   50
Mean                             45.60
Standard Deviation                6.14
Variance                         37.71
Skewness                         -2.55
Kurtosis                          7.14
Standard error of mean            0.87
Standard error of measurement     2.37
Coefficient Alpha                 0.85
=======================================================================
```

Figure 9 presents mean, standard deviation, variance, skewness, kurtosis, standard error of measurement, and reliability values. Since the items in the example are Likert-type, the reliability coefficient is the Cronbach Alpha value. However, when there is a test consisting of double-scored (0-1) items, the reliability value will express the KR-20 value.

After the analyses are performed, the Plot menu becomes active. With the help of the tabs in Figure 10, visual outputs such as characteristic curves, item expected score curves, cumulative and conditional item characteristic curves, item information function, test information function, information function, test characteristic curve, Wright map, and Predicted Probability Wright Map can be created. However, Test infographics and Test characteristic curves cannot be created in multidimensional models. Some example graphic outputs of example application-1 are presented in Appendix 1.

Figure 10. *ConQuest Plot menu and sub-tabs.*


ConQuest - bim.cqc
File
Edit
Run
Command
Analysis
Tables
Plot
Workspace
Options
Help
Input Window
Characteristic Curves By Score
Characteristic Curves By Category
Item Expected Score Curves
Cumulative Item Characteristic Curves
Conditional Item Characteristic Curves
Item Information
Test Information
Information map
Test Characteristic Curve
Wright Map
Predicted Probability Wright Map
Launch PlotQuest

```
datafile bim_dat.txt;
format id 1-3 responses 4-13;
labels << bim_lab.txt;
codes 1,2,3,4,5;
score (1,2,3,4,5) (1,2,3,4,5) ()
score (1,2,3,4,5) () (1,2,3,4,5)
model items;
estimate;
show !estimates=latent,tables=1:
itanal >> bim_itn.txt;
show cases !estimates=eap >> bim
show cases !estimates=mle >> bim
set warnings=no,update=yes;
export parameters >>bim_prm.txt;
export reg_coefficients >>bim_reg.txt;
export covariance >> bim_cov.txt;
```

2.3. Example 2: Within-Item Multidimensional Model

The data created for the Within-Item Multidimensional Model consists of 10 Likert-type items scored from 1 to 5. Assuming that items 1, 6, and 9 are related to both the first and second dimensions of the scale, the within-item multidimensional model in Figure 4 is defined. Items 2, 3, and 5 are only in the first dimension, while items 4, 7, 8, and 10 are only in the second dimension. After creating the data of 50 individuals, the data file with the *.txt* extension was prepared for analysis. Table 4 involves the command, data, and tag file created by the researchers for the Within-item multidimensional model.

Table 4. *An example command, data, and tag file in Within-item Multidimensional model.*

wim.cqc	wim_dat.txt	wim_lab.txt
Command statements (required)	Dataset (required)	Dataset label

datafile wim_dat.txt;
format id 1-3 responses 4-13;
labels << wim_lab.txt;
codes 1,2,3,4,5;
score (1,2,3,4,5) (1,2,3,4,5) (1,2,3,4,5) ! items(1);
score (1,2,3,4,5) (1,2,3,4,5) () ! items(2);
score (1,2,3,4,5) (1,2,3,4,5) () ! items(3);
score (1,2,3,4,5) () (1,2,3,4,5) ! items(4);
score (1,2,3,4,5) (1,2,3,4,5) () ! items(5);
score (1,2,3,4,5) (1,2,3,4,5) (1,2,3,4,5) ! items(6);
score (1,2,3,4,5) () (1,2,3,4,5) ! items(7);
score (1,2,3,4,5) () (1,2,3,4,5) ! items(8);
score (1,2,3,4,5) (1,2,3,4,5) (1,2,3,4,5) ! items(9);
score (1,2,3,4,5) () (1,2,3,4,5) ! items(10);
model items;
estimate;
show !estimates=latent,tables=1:2:3:9>> wim_shw.txt;
itanal >> wim_itn.txt;
show cases !estimates=eap >> wim_eap.txt;
show cases !estimates=mle >> wim_mle.txt;
	0014555555551
	0025555515555
	0032553141311
	0045555555555
	0055555455455
	0065455511451
	0075555445555
	0085555553551
	0095555555555
	0105555455555
	0115555354551
		===> item
		1 M1
		2 M2
		3 M3
		4 M4
		5 M5
		6 M6
		7 M7
		8 M8
		9 M9
		10 M10

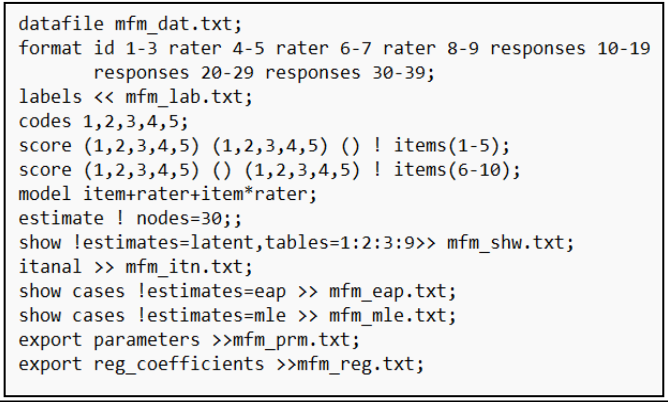
To perform the analysis, the command is opened and run. Since Rasch Analysis results (*win_shw.txt*) and traditional item analysis results (*win_itn.txt*) are similar to application example-1, they are presented in Appendix 2. However, a generalized item thresholds table and map cannot be created for Within-Item Multidimensional models. This situation will be better understood when item-1 and item-2 in the traditional item statistics part of Appendix 2 are examined. For example, since item-1 is in both dimensions, threshold values for item-1 cannot be calculated, but these values are calculated for item-2 in only one dimension. Also, some example graphic outputs of example application-2 are presented in Appendix 3.

2.4. Example 3: Many-facet Multidimensional Model

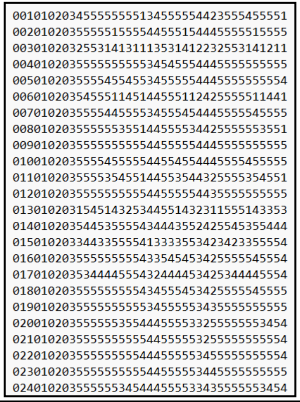
The data created for the Many-facet Multidimensional Model consists of 10 Likert-type items scored from 1 to 5. Items from 1 to 5 are assumed to represent the first dimension, while items from 6 to 10 are assumed to represent the second dimension of the scale. Data from 50 individuals in total were created. It was assumed that each individual's response to each item was scored by three different raters (MSK, FE, MS). Therefore, this example application includes two different facets (item and rater). Table 5 contains the command, data, and tag file created by the researchers for the Many-facet multidimensional model.

Table 5. *An example command, data, and tag file in Many-facet Multidimensional Model.*

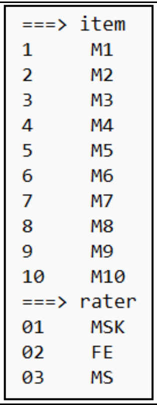
mfm.cqc	mfm_dat.txt	mfm_lab.txt
Command statements (required)	Dataset (required)	Dataset label



```
datafile mfm_dat.txt;
format id 1-3 rater 4-5 rater 6-7 rater 8-9 responses 10-19
       responses 20-29 responses 30-39;
labels << mfm_lab.txt;
codes 1,2,3,4,5;
score (1,2,3,4,5) (1,2,3,4,5) () ! items(1-5);
score (1,2,3,4,5) () (1,2,3,4,5) ! items(6-10);
model item+rater+item*rater;
estimate ! nodes=30;;
show !estimates=latent,tables=1:2:3:9>> mfm_shw.txt;
itanal >> mfm_itn.txt;
show cases !estimates=eap >> mfm_eap.txt;
show cases !estimates=mle >> mfm_mle.txt;
export parameters >>mfm_prm.txt;
export reg_coefficients >>mfm_reg.txt;
```



```
001010203455555555134555554423555455551
002010203555551555544555154445555515555
003010203255314131113531412232553141211
004010203555555555534545554445555555555
005010203555545545534555554445555555554
006010203545551145144555112425555511441
007010203555544555345554544455555545555
008010203555555355144555534425555553551
009010203555555555544555554445555555555
010010203555545555544554554445555455555
011010203555535455144553544325555354551
012010203555555555544555554435555555555
013010203154514325344551432311555143353
014010203544535555434443552425545355444
015010203344335555413333553423423355554
016010203555555555433545453425555545554
017010203534444555432444453425344445554
018010203555555555434555453425555545555
019010203555555555534555553435555555555
020010203555555355444555533255555553454
021010203555555555544555553255555555554
022010203555555555444555553455555555554
023010203555555555544555553445555555555
024010203555555345444555533435555553454
```



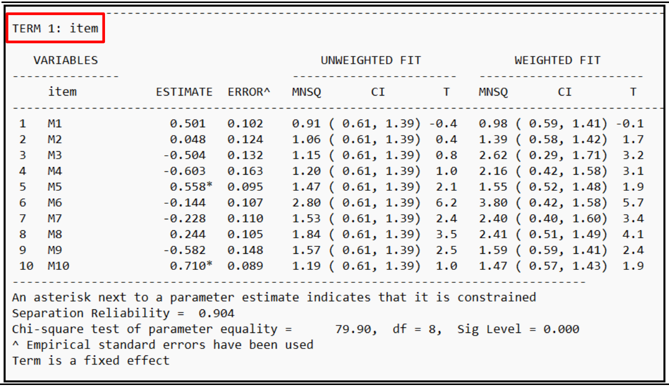
```
===> item
1     M1
2     M2
3     M3
4     M4
5     M5
6     M6
7     M7
8     M8
9     M9
10    M10
===> rater
01    MSK
02    FE
03    MS
```

To perform the analysis, the command file is opened and run. Rasch Analysis results (*mfm_shw.txt*) and traditional item analysis results (*mfm_itn.txt*) for application example-3 are presented below.

Appendix 4 displays the summary of the estimation and Population Model Parameter Estimates (regression coefficients, covariance/correlation matrix, reliability coefficients), which are similar to the previous application example. On the other hand, Appendix 5 presents example graphic printouts. However, unlike the other two applications, the number of terms in the model has changed. Since a two-facet model is created, there are three terms: *item*, *rater*, and *item*rater*. Therefore, the item parameters of each term in the model were estimated. Table 6 presents the predicted item parameters for each term in the model.

Table 6. *Item parameter estimates for each term in the model.*

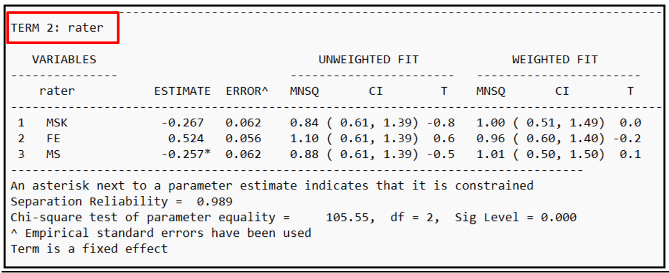
Explanation



```
------------------------------------------------------------------------------------
TERM 1: item

  VARIABLES                              UNWEIGHTED FIT            WEIGHTED FIT
---------------                     -----------------------   -----------------------
    item          ESTIMATE  ERROR^   MNSQ      CI        T     MNSQ      CI        T
------------------------------------------------------------------------------------
1   M1              0.501   0.102    0.91 ( 0.61, 1.39) -0.4   0.98 ( 0.59, 1.41) -0.1
2   M2              0.048   0.124    1.06 ( 0.61, 1.39)  0.4   1.39 ( 0.58, 1.42)  1.7
3   M3             -0.504   0.132    1.15 ( 0.61, 1.39)  0.8   2.62 ( 0.29, 1.71)  3.2
4   M4             -0.603   0.163    1.20 ( 0.61, 1.39)  1.0   2.16 ( 0.42, 1.58)  3.1
5   M5              0.558*  0.095    1.47 ( 0.61, 1.39)  2.1   1.55 ( 0.52, 1.48)  1.9
6   M6             -0.144   0.107    2.80 ( 0.61, 1.39)  6.2   3.80 ( 0.42, 1.58)  5.7
7   M7             -0.228   0.110    1.53 ( 0.61, 1.39)  2.4   2.40 ( 0.40, 1.60)  3.4
8   M8              0.244   0.105    1.84 ( 0.61, 1.39)  3.5   2.41 ( 0.51, 1.49)  4.1
9   M9             -0.582   0.148    1.57 ( 0.61, 1.39)  2.5   1.59 ( 0.59, 1.41)  2.4
10  M10             0.710*  0.089    1.19 ( 0.61, 1.39)  1.0   1.47 ( 0.57, 1.43)  1.9
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.904
Chi-square test of parameter equality =      79.90,  df = 8,  Sig Level = 0.000
^ Empirical standard errors have been used
Term is a fixed effect
```

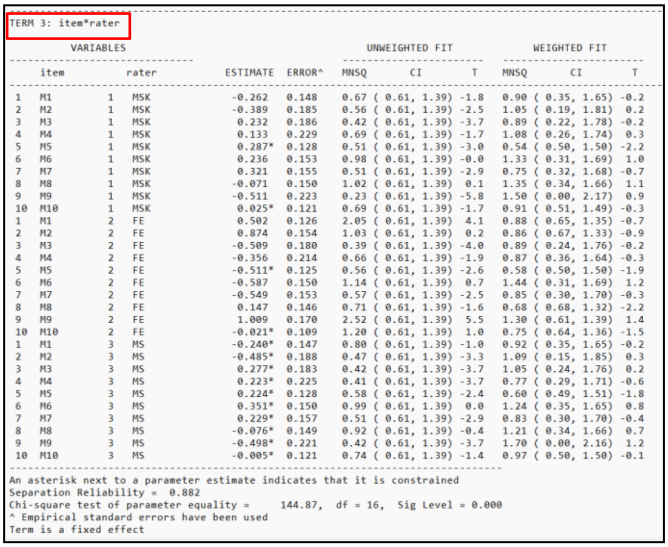
Term 1: item

It includes parameter estimations of ten items and some fit values.



```
TERM 2: rater

   VARIABLES                              UNWEIGHTED FIT             WEIGHTED FIT
---------------                       -----------------------   -----------------------
    rater             ESTIMATE  ERROR^  MNSQ       CI        T    MNSQ       CI        T
----------------------------------------------------------------------------------------
 1   MSK               -0.267   0.062    0.84 ( 0.61, 1.39) -0.8   1.00 ( 0.51, 1.49)  0.0
 2   FE                 0.524   0.056    1.10 ( 0.61, 1.39)  0.6   0.96 ( 0.60, 1.40) -0.2
 3   MS                -0.257*  0.062    0.88 ( 0.61, 1.39) -0.5   1.01 ( 0.50, 1.50)  0.1
-------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.989
Chi-square test of parameter equality =     105.55,  df = 2,  Sig Level = 0.000
^ Empirical standard errors have been used
Term is a fixed effect
```

Term 2: rater
Parameter estimates and some fit values of three different raters are included.
A negative estimate indicates that the rater is leniency, and a positive estimate indicates that the rater is severity.

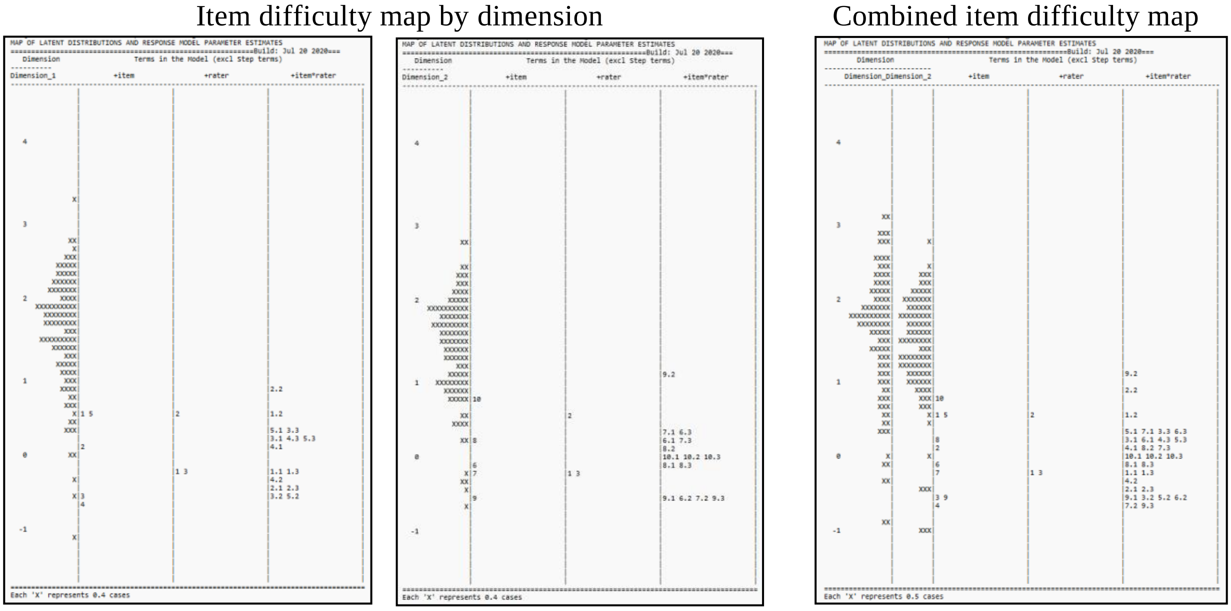


```
TERM 3: item*rater
          VARIABLES                                  UNWEIGHTED FIT            WEIGHTED FIT
------------------------------                   -----------------------   -----------------------
     item           rater          ESTIMATE  ERROR^  MNSQ       CI        T   MNSQ       CI        T
----------------------------------------------------------------------------------------------------
 1   M1         1   MSK             -0.262   0.148   0.67 ( 0.61, 1.39) -1.8  0.90 ( 0.35, 1.65) -0.2
 2   M2         1   MSK             -0.389   0.185   0.56 ( 0.61, 1.39) -2.5  1.05 ( 0.19, 1.81)  0.2
 3   M3         1   MSK              0.232   0.186   0.42 ( 0.61, 1.39) -3.7  0.89 ( 0.22, 1.78) -0.2
 4   M4         1   MSK              0.133   0.229   0.69 ( 0.61, 1.39) -1.7  1.08 ( 0.26, 1.74)  0.3
 5   M5         1   MSK              0.287*  0.128   0.51 ( 0.61, 1.39) -3.0  0.54 ( 0.50, 1.50) -2.2
 6   M6         1   MSK              0.236   0.153   0.98 ( 0.61, 1.39) -0.0  1.33 ( 0.31, 1.69)  1.0
 7   M7         1   MSK              0.321   0.155   0.51 ( 0.61, 1.39) -2.9  0.75 ( 0.32, 1.68) -0.7
 8   M8         1   MSK             -0.071   0.150   1.02 ( 0.61, 1.39)  0.1  1.35 ( 0.34, 1.66)  1.1
 9   M9         1   MSK             -0.511   0.223   0.23 ( 0.61, 1.39) -5.8  1.50 ( 0.00, 2.17)  0.9
 10  M10        1   MSK              0.025*  0.121   0.69 ( 0.61, 1.39) -1.7  0.91 ( 0.51, 1.49) -0.3
 1   M1         2   FE               0.502   0.126   2.05 ( 0.61, 1.39)  4.1  0.88 ( 0.65, 1.35) -0.7
 2   M2         2   FE               0.874   0.154   1.03 ( 0.61, 1.39)  0.2  0.86 ( 0.67, 1.33) -0.9
 3   M3         2   FE              -0.509   0.180   0.39 ( 0.61, 1.39) -4.0  0.89 ( 0.24, 1.76) -0.2
 4   M4         2   FE              -0.356   0.214   0.66 ( 0.61, 1.39) -1.9  0.87 ( 0.36, 1.64) -0.3
 5   M5         2   FE              -0.511*  0.125   0.56 ( 0.61, 1.39) -2.6  0.58 ( 0.50, 1.50) -1.9
 6   M6         2   FE              -0.587   0.150   1.14 ( 0.61, 1.39)  0.7  1.44 ( 0.31, 1.69)  1.2
 7   M7         2   FE              -0.549   0.153   0.57 ( 0.61, 1.39) -2.5  0.85 ( 0.30, 1.70) -0.3
 8   M8         2   FE               0.147   0.146   0.71 ( 0.61, 1.39) -1.6  0.68 ( 0.68, 1.32) -2.2
 9   M9         2   FE               1.009   0.170   2.52 ( 0.61, 1.39)  5.5  1.30 ( 0.61, 1.39)  1.4
 10  M10        2   FE              -0.021*  0.109   1.20 ( 0.61, 1.39)  1.0  0.75 ( 0.64, 1.36) -1.5
 1   M1         3   MS              -0.240*  0.147   0.80 ( 0.61, 1.39) -1.0  0.92 ( 0.35, 1.65) -0.2
 2   M2         3   MS              -0.485*  0.188   0.47 ( 0.61, 1.39) -3.3  1.09 ( 0.15, 1.85)  0.3
 3   M3         3   MS               0.277*  0.183   0.42 ( 0.61, 1.39) -3.7  1.05 ( 0.24, 1.76)  0.2
 4   M4         3   MS               0.223*  0.225   0.41 ( 0.61, 1.39) -3.7  0.77 ( 0.29, 1.71) -0.6
 5   M5         3   MS               0.224*  0.128   0.58 ( 0.61, 1.39) -2.4  0.60 ( 0.49, 1.51) -1.8
 6   M6         3   MS               0.351*  0.150   0.99 ( 0.61, 1.39)  0.0  1.24 ( 0.35, 1.65)  0.8
 7   M7         3   MS               0.229*  0.157   0.51 ( 0.61, 1.39) -2.9  0.83 ( 0.30, 1.70) -0.4
 8   M8         3   MS              -0.076*  0.149   0.92 ( 0.61, 1.39) -0.4  1.21 ( 0.34, 1.66)  0.7
 9   M9         3   MS              -0.498*  0.221   0.42 ( 0.61, 1.39) -3.7  1.70 ( 0.00, 2.16)  1.2
 10  M10        3   MS              -0.005*  0.121   0.74 ( 0.61, 1.39) -1.4  0.97 ( 0.50, 1.50) -0.1
-------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.882
Chi-square test of parameter equality =     144.87,  df = 16,  Sig Level = 0.000
^ Empirical standard errors have been used
Term is a fixed effect
```

*Term 3: item*rater*
Since it is the interaction of the item and rater terms, there are 3x10 (a total of 30) parameter estimations and some fit values. The term interaction is used to determine the joint effect between the variables rather than the main effects.

Similarly, due to the increase in the number of terms in the model, there were some changes in the item difficulty map outputs created according to the latent ability distribution and item difficulty. In Figure 11, there is an item difficulty map created for each dimension separately or by combining the dimensions.

Figure 11. *The item difficulty map formed separately according to the dimension and the item difficulty map formed by combining the dimensions.*



Item difficulty map by dimension

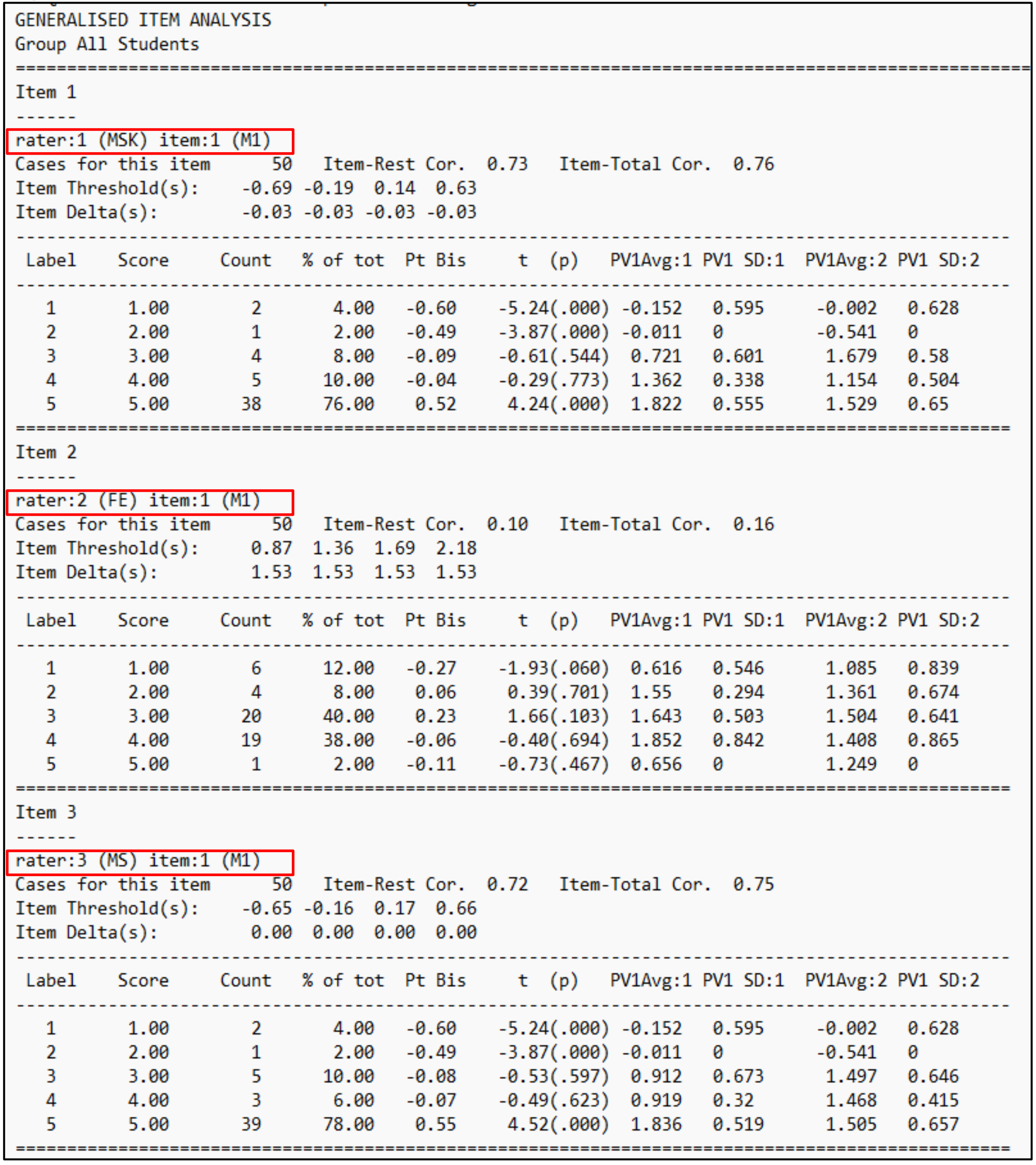
Combined item difficulty map

As can be seen in Figure 11, each dimension was calculated separately and combined, and individual, item, rater, and item*rater interactions were given together. Results for each dimension can be interpreted separately or in combination. The most difficult item is the number 10 followed by 1 and 5. In the example, the most difficult items were items 5 and 10. Regarding the logit values of the raters, raters numbered one and three were more generous, while rater numbered two was stricter. For item*rater interactions, rater number two exhibited more rigid behaviors in items numbered one, two, and nine.

The Item difficulty map obtained for the multidimensional and many-facet model provides that both the items, the dimensions, the raters, and the item and rater interactions are given on the same scale. In addition, it transforms all sources of variability into logit scales and makes them dependent variables. For example, when the combined item difficulty map is examined, it is seen that the first dimension has a wider range than the second dimension. In addition, it provides the opportunity to see information such as which is the most difficult question or who is the severity rater.

Considering the traditional item analysis results, since there are two terms in the model, item statistics were calculated as much as the interaction number of these two terms. As it is assumed that each item is evaluated by three different raters, traditional item statistics for each item were calculated for each rater. Figure 12 presents the traditional item statistical outputs of three different raters for item-1

Figure 12. *Statistics of three different raters for item-1.*



```
GENERALISED ITEM ANALYSIS
Group All Students
=================================================================================================
Item 1
------
rater:1 (MSK) item:1 (M1)
Cases for this item      50   Item-Rest Cor.  0.73   Item-Total Cor.  0.76
Item Threshold(s):     -0.69 -0.19  0.14  0.63
Item Delta(s):         -0.03 -0.03 -0.03 -0.03
-----------------------------------------------------------------------------------------------
 Label     Score     Count  % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1  PV1Avg:2 PV1 SD:2
-----------------------------------------------------------------------------------------------
   1       1.00        2      4.00    -0.60    -5.24(.000) -0.152   0.595     -0.002   0.628
   2       2.00        1      2.00    -0.49    -3.87(.000) -0.011   0         -0.541   0
   3       3.00        4      8.00    -0.09    -0.61(.544)  0.721   0.601      1.679   0.58
   4       4.00        5     10.00    -0.04    -0.29(.773)  1.362   0.338      1.154   0.504
   5       5.00       38     76.00     0.52     4.24(.000)  1.822   0.555      1.529   0.65
===============================================================================================
Item 2
------
rater:2 (FE) item:1 (M1)
Cases for this item      50   Item-Rest Cor.  0.10   Item-Total Cor.  0.16
Item Threshold(s):      0.87  1.36  1.69  2.18
Item Delta(s):          1.53  1.53  1.53  1.53
-----------------------------------------------------------------------------------------------
 Label     Score     Count  % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1  PV1Avg:2 PV1 SD:2
-----------------------------------------------------------------------------------------------
   1       1.00        6     12.00    -0.27    -1.93(.060)  0.616   0.546      1.085   0.839
   2       2.00        4      8.00     0.06     0.39(.701)  1.55    0.294      1.361   0.674
   3       3.00       20     40.00     0.23     1.66(.103)  1.643   0.503      1.504   0.641
   4       4.00       19     38.00    -0.06    -0.40(.694)  1.852   0.842      1.408   0.865
   5       5.00        1      2.00    -0.11    -0.73(.467)  0.656   0          1.249   0
===============================================================================================
Item 3
------
rater:3 (MS) item:1 (M1)
Cases for this item      50   Item-Rest Cor.  0.72   Item-Total Cor.  0.75
Item Threshold(s):     -0.65 -0.16  0.17  0.66
Item Delta(s):          0.00  0.00  0.00  0.00
-----------------------------------------------------------------------------------------------
 Label     Score     Count  % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1  PV1Avg:2 PV1 SD:2
-----------------------------------------------------------------------------------------------
   1       1.00        2      4.00    -0.60    -5.24(.000) -0.152   0.595     -0.002   0.628
   2       2.00        1      2.00    -0.49    -3.87(.000) -0.011   0         -0.541   0
   3       3.00        5     10.00    -0.08    -0.53(.597)  0.912   0.673      1.497   0.646
   4       4.00        3      6.00    -0.07    -0.49(.623)  0.919   0.32       1.468   0.415
   5       5.00       39     78.00     0.55     4.52(.000)  1.836   0.519      1.505   0.657
===============================================================================================
```

Figure 12 shows that the statistical values of the items change according to the rater's attitude. While the item-rest correlation values for the item are high for raters 1 and 3 (r=0.72-0.73), the value calculated for rater 2 is low (r=0.10).

In addition to this Many-facet Multidimensional Model, hypothetical, some research examples in the field of education can also be given. For instance, research such as scoring language skills with a multidimensional structure, which is frequently used in classroom measurement and assessment practices, can be designed by peer assessment. As another example, research can be conducted on scoring students' presentation skills by more than one rater. As a different example, the socio-economic levels and genders of the students can be included in the model as a variable, and analyses can be carried out through the Conquest program by using the Many-facet Multidimensional Model in evaluating academic achievements with a multidimensional structure.

3. DISCUSSION, CONCLUSION and SUGGESTIONS

Please This research aimed to introduce ConQuest, the statistical package program used in the analysis of multidimensional and many-facet data structures, and to show its applications using an example data set. Thus, the installation of the program and the steps of the analysis process were explained.

Conquest is a user-friendly program because of its simple interface. When real-life situations are examined, it is often observed that data sets are complex and multidimensional. Thus, the meaning of performing analyses with only one-dimensional data sets actually means that many data sets cannot be analyzed in real terms. This situation (construct under-representation) creates a negative situation on the validity of direct measurements (Messick, 1995). In this context, analyzing data sets representing real-life situations (e.g., ConQuest) will contribute to the validity of the measurements. Besides, an important feature that distinguishes ConQuest from other programs that perform multidimensional IRT analyses is that it can simultaneously include many different sources of variability in the analysis and show all variables on a single scale. All sources of variability can be interpreted in an interdependent manner. Also, the ConQuest program can provide outputs for the main effects of the variables as well as their common interactions. Thus, rater biases and item biases, which are frequently used in validity studies, can also be analyzed. The third example is an application that takes this situation into account and has not been tested in previous studies. It can be stated that Many-Facet can be applied in many cases that require multidimensionality analysis.

In a nutshell, the ConQuest program is a suitable and user-friendly package program for many-facet and multidimensional data analysis. Many-facet multidimensional analyses can be easily run via the ConQuest package program in situations where there is more than one construct such as higher-order mental skills and in cases where decisions are made by jury evaluations. In addition to its advantages such as easy use, simple interface, and fast analysis, it also has disadvantages such as being a paid program and limited demo version.

Acknowledgments

The authors would like to thank the blind reviewers for their useful comments and insightful suggestions.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Mahmut Sami Koyuncu: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Mehmet Şata**: Methodology, Supervision, and Validation.

Orcid

Mahmut Sami Koyuncu https://orcid.org/0000-0002-6651-4851
Mehmet Şata https://orcid.org/0000-0003-2683-4997

REFERENCES

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement In Education*, *7*(4), 255–278. https://doi.org/10.1207/s15324818ame0704_1

Adams, R.J., Wilson, M.R., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*, 1–24. https://doi.org/10.1177%2F0146621697211001

Adams, R.J., Wilson, M.R., & Wu, M.L. (1997). Multilevel Item Response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioural Statistics*, *22*, 46–75. https://doi.org/10.2307/1165238

Adams, R., Cloney, D., Wu, M., Osses, A., Schwantner, V., & Vista, A. (2022). ACER ConQuest Manual. *https://conquestmanual.acer.org/*

Adams, R.J, Wu, M.L, Cloney, D., and Wilson, M.R. (2020). *ACER ConQuest: Generalised Item Response Modelling Software* [Computer software]. Version 5. Camberwell, Victoria: Australian Council for Educational Research.

Andrich, D. (1978). A Rating Formulation for Ordered Response Categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Bartolomé, J., & Garaizar, P. (2022). Design and Validation of a Novel Tool to Assess Citizens' Netiquette and Information and Data Literacy Using Interactive Simulations. *Sustainability*, *14*(6), 3392. https://doi.org/10.3390/su14063392

Bock, D.R., & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: An Application of the EM Algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Brnic, M., & Greefrath, G. (2021, September 13–16). *Does The Gender Matter? The Use of A Dıgıtal Textbook Compared To Prınted Materıals.* 15th International Conference on Technology In Mathematics Teaching (ICTMT 15), Copenhagen, Denmark.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

De Ayala, R.J. (2009). *The theory and practice of ıtem response theory. Methodology in the Social Sciences.* New York: Guildford.

Finch, H., & Habing, B. (2003, April). *Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items.* Paper presented at the annual meeting of the National Council on Measurement, Chicago.

Fischer, G.H. (1983). Logistic Latent Trait Models with Linear Constraints. *Psychometrika, 48*, 3–26. https://doi.org/10.1007/BF02314674

Hahn, I. & Kähler, J. (2022). *NEPS Technical Report for Science: Scaling Results of Starting Cohort 3 for Grade 11* (NEPS Survey Paper No. 93). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP93:1.0

Jang, E.E., & Roussos, L.A. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based non-parametric approach. *Journal of Educational Measurement*, *44*(1), 1-21. https://doi.org/10.1111/j.1745-3984.2007.00024.x

Jolin, J., & Wilson, M. (2022). Developing a Theory of Two Latent Soft Skills Progress Variables using the BEAR Assessment System: Validity Evidence for the Internal Structure of the Social Evaluative in the Workplace Instrument. *Journal of Psychoeducational Assessment*, *40*(3), 381-399. https://doi.org/10.1177/07342829211057641

Jüttler, M., & Schumann, S. (2022). The long-term effects of students' economic competencies on the transition from school to university in the international context. *Research in Comparative and International Education*, *17*(2), 196-224. https://doi.org/10.1177/17454999221086191

Krell, M., Khan, S., Vergara, C., Cofré, H., Mathesius, S., & Krüger, D. (2022). Pre-Service Science Teachers' Scientific Reasoning Competencies: Analysing the Impact of Contributing Factors. *Research in Science Education*, 1-21. https://doi.org/10.1007/s11165-022-10045-x

Koch, A., Wißhak, S., Spener, C., Naumann, A., & Hochholdinger, S. (2022). Transfer knowledge of trainers in continuing vocational education and training: Construction and piloting of a test instrument. *Journal for Research on Adult Education*, 1-17. https://doi.org/10.1007/s40955-022-00210-0

Köse, İ.A. (2012). Çok boyutlu madde tepki kuramı [Multidimensional Item Response Theory]. *Journal of Measurement and Evaluation in Education and Psychology*, *3*(1), 221-229.

Linacre, J.M. (1994). *Many-Facet Rasch Measurement.* MESA Press.

Lou, J., Chen, H., & Li, R. (2022). Emotional Intelligence Scale for Male Nursing Students and Its Latent Regression on Gender and Background Variables. *Healthcare*, *10*(5), 814. https://doi.org/10.3390/healthcare10050814

Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Mendoza, N.B., Cheng, E.C., & Yan, Z. (2022). Assessing teachers' collaborative lesson planning practices: Instrument development and validation using the SECI knowledge-creation model. *Studies in Educational Evaluation*, *73*, 101139. https://doi.org/10.1016/j.stueduc.2022.101139

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Mischo, C., Wolstein, K., & Peters, S. (2022). Professional vision of early childhood teachers: relations to knowledge, work experience and teacher child-interaction. *Early Years*, 1-17. https://doi.org/10.1080/09575146.2022.2028129

Mroch, A.A., & Bolt, D.M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education*, *19*(1), 67-91. https://doi.org/10.1207/s15324818ame1901_4

Oko, J. (2022). Creating a motivation scale for secondary school students in Papua New Guinea. *Journal of Applied Learning and Teaching*, *5*(1), 1-10. https://doi.org/10.37074/jalt.2022.5.1.4

Osterhaus, C., Kristen-Antonow, S., Kloo, D., & Sodian, B. (2022). Advanced scaling and modeling of children's theory of mind competencies: Longitudinal findings in 4-to 6-year-olds. *International Journal of Behavioral Development*, *46*(3), 251-259. https://doi.org/10.1177/01650254221077334

Özbek-Baştuğ, O.Y. (2012). Assessment of dimensionality in social science subtest. *Educational Sciences: Theory and Practice*, *12*(1), 375-385.

Özer-Özkan, Y. (2012). Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması. [A Comparison of Estimated Achivement Scores Obtained From Student Achievement Assessment Test Utilizing Classical Test Theory, Unidimensional And Multidimensional Item Response Theory Models] [Doctoral dissertation, Ankara University]. National Thesis Center of Higher Education Board. https://tez.yok.gov.tr/UlusalTezMerkezi/

Özer-Özkan, Y., & Acar-Güvendir, M. (2014). The analysis of large-scale tests applied in Turkey in terms of their multidimensionality. *Mehmet Akif Ersoy University Journal of Education Faculty, 1*(29), 31-47.

Patz, R.J., & Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178. https://doi.org/10.2307/1165199

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Test.* University of Chicago Press.

Spink, J., Cloney, D., & Berry, A. (2022, January 01). Beyond letters and numbers: the COVID-19 pandemic and foundational literacy and numeracy in Indonesia. *International Education Research.* https://research.acer.edu.au/int_research/7

Stout, W., Froelich, A.G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M.A.J. van Duijn, & T.A.B. Snijders (Eds.), *Essay on item response theory* (pp. 357-375). Springer. https://doi.org/10.1007/978-1-4613-0169-1_19

Unfried, A., Rachmatullah, A., Alexander, A., & Wiebe, E. (2022). An alternative to STEBI-A: validation of the T-STEM science scale. *International Journal of STEM Education, 9*(1), 1-14. https://doi.org/10.1186/s40594-022-00339-x

Volodin, N., & Adams, R.J. (1995). Identifying and estimating a d-dimensional item response model. *International Objective Measurement Workshop,* University of California.

Wall, S.P., Castillo, P., Shuchat-Shaw, F., Norman, E., Brown, D., Martinez-López, N., & Ravenell, J. E. (2022). Targeting versus Tailoring Educational Videos for Encouraging Deceased Organ Donor Registration in Black-Owned Barbershops. *Journal of Health Communication*, *27*(1), 37-48. https://doi.org/10.1080/10810730.2022.2035021

Wang, W. (1995). *Implementation and application of the multidimensional random coefficients multinomial logit*. [Unpublished Doctoral dissertation]. University of California.

Wang, X., Yan, Z., Huang, Y., Tang, A., & Chen, J. (2022). Re-Developing the Adversity Response Profile for Chinese University Students. *International Journal of Environmental Research and Public Health, 19*, 6389. https://doi.org/10.3390/ijerph19116389.

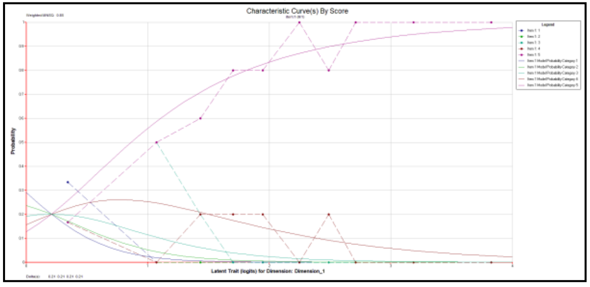
Wilson, M.R. (1992). The ordered partition model: an extension of the partial credit model. *Applied Psychological Measurement, 16*, 309-325. https://doi.org/10.1177/014662169201600401

Wright, B.D., & Stone, M.H. (1979). *Best test design: Rasch measurement*. MESA Press.

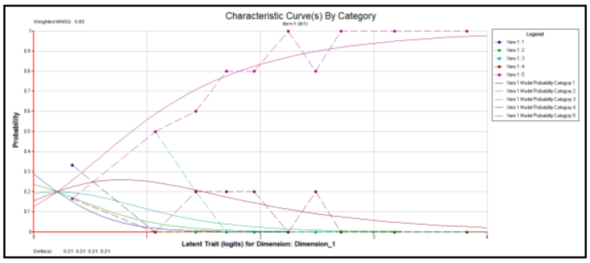
APPENDIX

Appendix 1. *Between-Item Multidimensional Model Sample PlotQuest Outputs.*

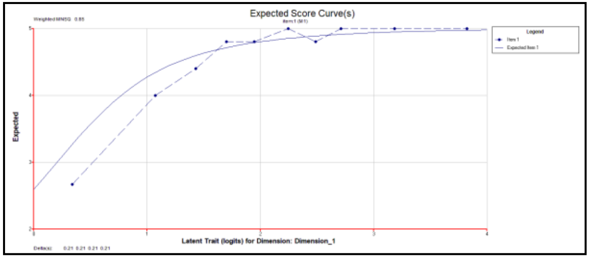
Characteristic Curves By Score



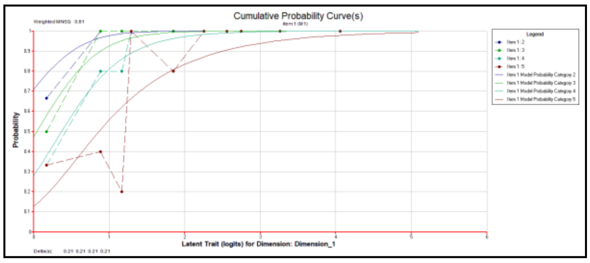
Characteristic Curves By Category



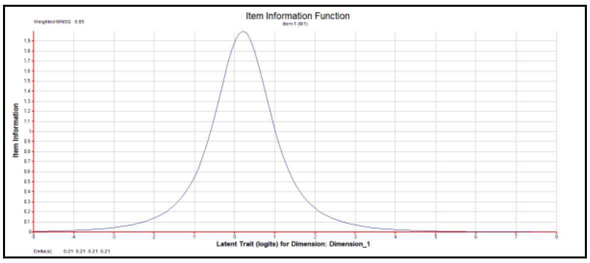
Item Expected Score Curves



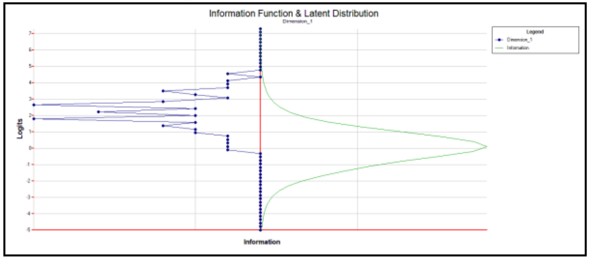
Cumulative Item Characteristic Curves



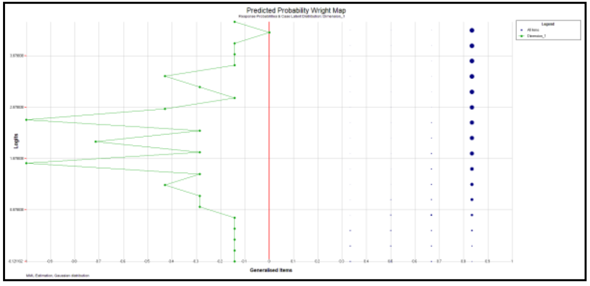
Item Information



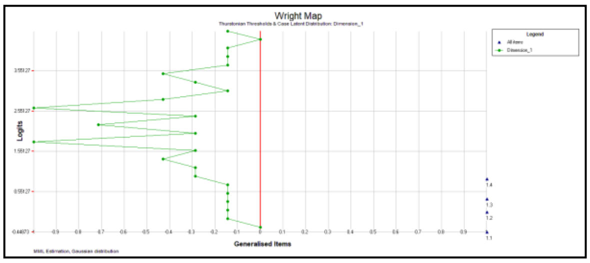
Information map



Predicted Probability Wright Map

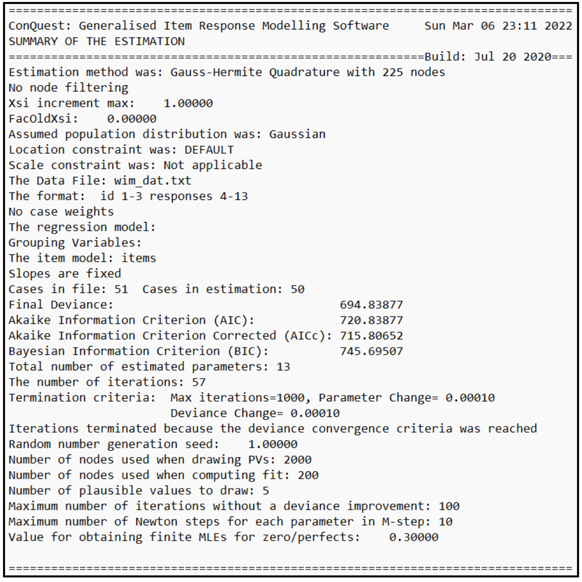


Wright Map



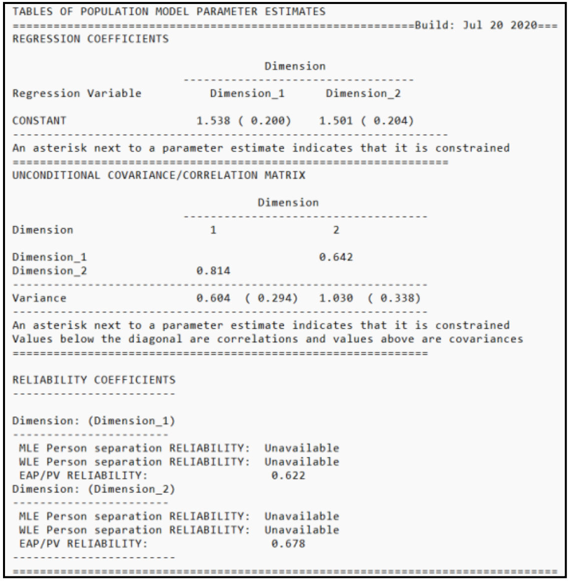
Appendix 2. *Within - Item Multidimensionality Rasch Analysis results (wim_shw.txt), traditional item analysis results (wim_itn.txt).*

Summary of the estimation



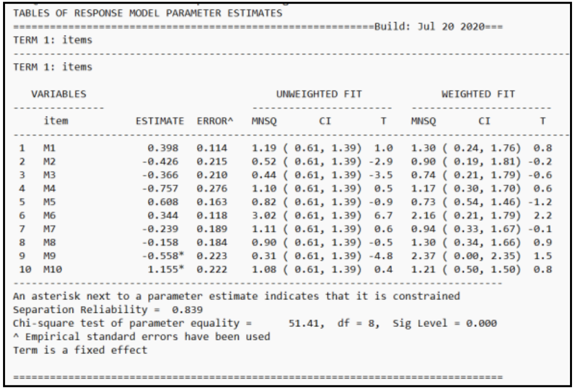
```
================================================================================
ConQuest: Generalised Item Response Modelling Software     Sun Mar 06 23:11 2022
SUMMARY OF THE ESTIMATION
=========================================================Build: Jul 20 2020===
Estimation method was: Gauss-Hermite Quadrature with 225 nodes
No node filtering
Xsi increment max:    1.00000
FacOldXsi:    0.00000
Assumed population distribution was: Gaussian
Location constraint was: DEFAULT
Scale constraint was: Not applicable
The Data File: wim_dat.txt
The format:  id 1-3 responses 4-13
No case weights
The regression model:
Grouping Variables:
The item model: items
Slopes are fixed
Cases in file: 51  Cases in estimation: 50
Final Deviance:                              694.83877
Akaike Information Criterion (AIC):          720.83877
Akaike Information Criterion Corrected (AICc): 715.80652
Bayesian Information Criterion (BIC):        745.69507
Total number of estimated parameters: 13
The number of iterations: 57
Termination criteria:  Max iterations=1000, Parameter Change= 0.00010
                       Deviance Change= 0.00010
Iterations terminated because the deviance convergence criteria was reached
Random number generation seed:    1.00000
Number of nodes used when drawing PVs: 2000
Number of nodes used when computing fit: 200
Number of plausible values to draw: 5
Maximum number of iterations without a deviance improvement: 100
Maximum number of Newton steps for each parameter in M-step: 10
Value for obtaining finite MLEs for zero/perfects:    0.30000


================================================================================
```

Population Model Parametre Estimates



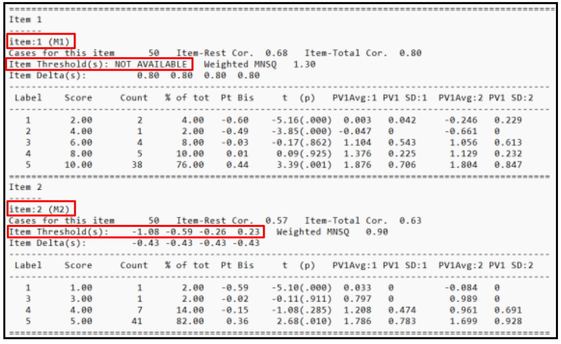
```
TABLES OF POPULATION MODEL PARAMETER ESTIMATES
===========================================================Build: Jul 20 2020===
REGRESSION COEFFICIENTS

                                    Dimension
                          ----------------------------------
Regression Variable         Dimension_1      Dimension_2

CONSTANT                  1.538 ( 0.200)    1.501 ( 0.204)
-----------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
=================================================================
UNCONDITIONAL COVARIANCE/CORRELATION MATRIX

                                   Dimension
                          ------------------------------------
Dimension                     1                  2

Dimension_1                                    0.642
Dimension_2               0.814
--------------------------------------------------------------
Variance                  0.604  ( 0.294)   1.030  ( 0.338)
--------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Values below the diagonal are correlations and values above are covariances
==============================================================

RELIABILITY COEFFICIENTS
------------------------

Dimension: (Dimension_1)
-----------------------
 MLE Person separation RELIABILITY:  Unavailable
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:                  0.622
Dimension: (Dimension_2)
-----------------------
 MLE Person separation RELIABILITY:  Unavailable
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:                  0.678
------------------------
=================================================================================
```

Rasch Model Item parameter estimates

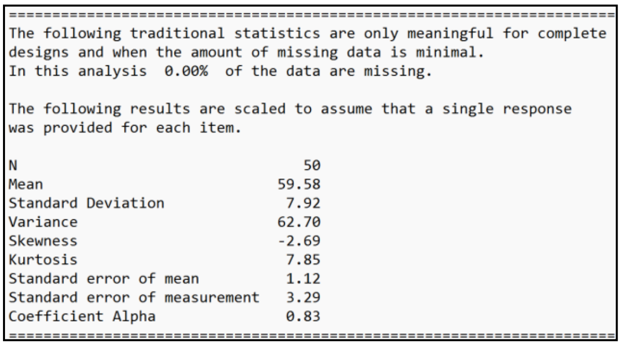


```
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
===========================================================Build: Jul 20 2020===
TERM 1: items
-------------------------------------------------------------------------------------
TERM 1: items

  VARIABLES                                  UNWEIGHTED FIT            WEIGHTED FIT
--------------                          ----------------------   ----------------------
    item           ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
-------------------------------------------------------------------------------------
 1   M1               0.398   0.114    1.19 ( 0.61, 1.39)  1.0   1.30 ( 0.24, 1.76)  0.8
 2   M2              -0.426   0.215    0.52 ( 0.61, 1.39) -2.9   0.90 ( 0.19, 1.81) -0.2
 3   M3              -0.366   0.210    0.44 ( 0.61, 1.39) -3.5   0.74 ( 0.21, 1.79) -0.6
 4   M4              -0.757   0.276    1.10 ( 0.61, 1.39)  0.5   1.17 ( 0.30, 1.70)  0.6
 5   M5               0.608   0.163    0.82 ( 0.61, 1.39) -0.9   0.73 ( 0.54, 1.46) -1.2
 6   M6               0.344   0.118    3.02 ( 0.61, 1.39)  6.7   2.16 ( 0.21, 1.79)  2.2
 7   M7              -0.239   0.189    1.11 ( 0.61, 1.39)  0.6   0.94 ( 0.33, 1.67) -0.1
 8   M8              -0.158   0.184    0.90 ( 0.61, 1.39) -0.5   1.30 ( 0.34, 1.66)  0.9
 9   M9              -0.558*  0.223    0.31 ( 0.61, 1.39) -4.8   2.37 ( 0.00, 2.35)  1.5
10   M10              1.155*  0.222    1.08 ( 0.61, 1.39)  0.4   1.21 ( 0.50, 1.50)  0.8
-------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.839
Chi-square test of parameter equality =      51.41,  df = 8,  Sig Level = 0.000
^ Empirical standard errors have been used
Term is a fixed effect

=========================================================================
```

Traditional Item (item-1 and item-2) and Test statistics

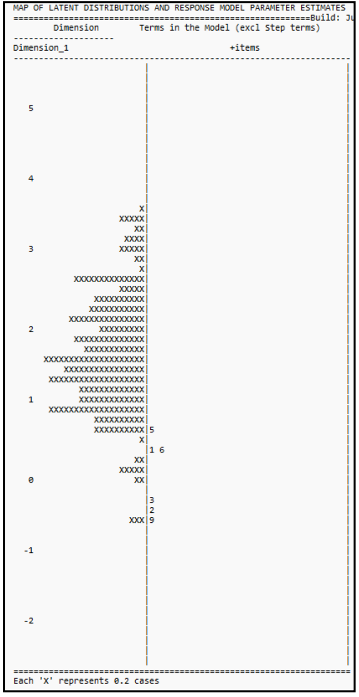


```
==========================================================================================
Item 1
------
item:1 (M1)
Cases for this item      50   Item-Rest Cor.  0.68   Item-Total Cor.  0.80
Item Threshold(s): NOT AVAILABLE   Weighted MNSQ   1.30
Item Delta(s):          0.80  0.80  0.80  0.80
-----------------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)    PV1Avg:1 PV1 SD:1  PV1Avg:2 PV1 SD:2
-----------------------------------------------------------------------------------------
   1       2.00        2       4.00   -0.60   -5.16(.000)  0.003   0.042     -0.246   0.229
   2       4.00        1       2.00   -0.49   -3.85(.000) -0.047   0         -0.661   0
   3       6.00        4       8.00   -0.03   -0.17(.862)  1.104   0.543      1.056   0.613
   4       8.00        5      10.00    0.01    0.09(.925)  1.376   0.225      1.129   0.232
   5      10.00       38      76.00    0.44    3.39(.001)  1.876   0.706      1.804   0.847
=========================================================================================
Item 2
------
item:2 (M2)
Cases for this item      50   Item-Rest Cor.  0.57   Item-Total Cor.  0.63
Item Threshold(s):     -1.08 -0.59 -0.26  0.23   Weighted MNSQ   0.90
Item Delta(s):         -0.43 -0.43 -0.43 -0.43
-----------------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)    PV1Avg:1 PV1 SD:1  PV1Avg:2 PV1 SD:2
-----------------------------------------------------------------------------------------
   1       1.00        1       2.00   -0.59   -5.10(.000)  0.033   0         -0.084   0
   3       3.00        1       2.00   -0.02   -0.11(.911)  0.797   0          0.989   0
   4       4.00        7      14.00   -0.15   -1.08(.285)  1.208   0.474      0.961   0.691
   5       5.00       41      82.00    0.36    2.68(.010)  1.786   0.783      1.699   0.928
=========================================================================================
```

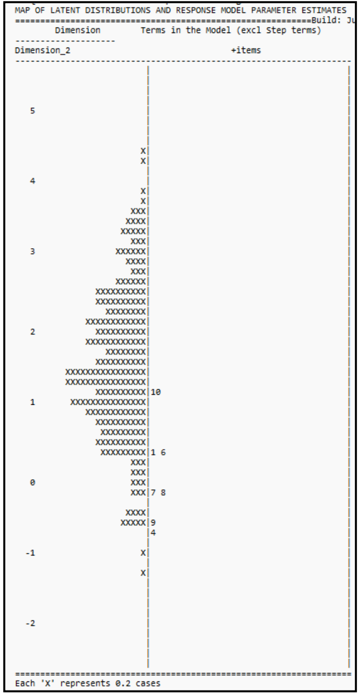


```
===================================================================
The following traditional statistics are only meaningful for complete
designs and when the amount of missing data is minimal.
In this analysis  0.00%  of the data are missing.

The following results are scaled to assume that a single response
was provided for each item.

N                                  50
Mean                            59.58
Standard Deviation               7.92
Variance                        62.70
Skewness                        -2.69
Kurtosis                         7.85
Standard error of mean           1.12
Standard error of measurement    3.29
Coefficient Alpha                0.83
===================================================================
```

Map of latent distributions and response model parameter estimates outputs

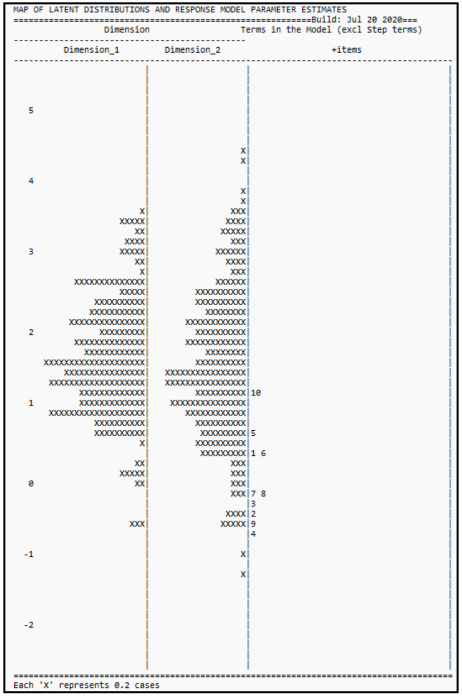


```
MAP OF LATENT DISTRIBUTIONS AND RESPONSE MODEL PARAMETER ESTIMATES
===========================================================Build: Ju
        Dimension        Terms in the Model (excl Step terms)
--------------------
Dimension_1                                +items
-------------------------------------------------------------------
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
   5                      |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
   4                      |                                       |
                          |                                       |
                          |                                       |
                         X|                                       |
                     XXXXX|                                       |
                        XX|                                       |
                      XXXX|                                       |
   3                 XXXXX|                                       |
                        XX|                                       |
                         X|                                       |
            XXXXXXXXXXXXXX|                                       |
                     XXXXX|                                       |
                XXXXXXXXXX|                                       |
               XXXXXXXXXXX|                                       |
           XXXXXXXXXXXXXXX|                                       |
   2             XXXXXXXXX|                                       |
            XXXXXXXXXXXXXX|                                       |
              XXXXXXXXXXXX|                                       |
      XXXXXXXXXXXXXXXXXXXX|                                       |
          XXXXXXXXXXXXXXXX|                                       |
       XXXXXXXXXXXXXXXXXXX|                                       |
             XXXXXXXXXXXXX|                                       |
   1         XXXXXXXXXXXXX|                                       |
       XXXXXXXXXXXXXXXXXXX|                                       |
                XXXXXXXXXX|                                       |
                XXXXXXXXXX|5                                      |
                         X|                                       |
                          |1 6                                    |
                        XX|                                       |
                     XXXXX|                                       |
   0                    XX|                                       |
                          |                                       |
                          |3                                      |
                          |2                                      |
                       XXX|9                                      |
                          |                                       |
                          |                                       |
  -1                      |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
  -2                      |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
===================================================================
Each 'X' represents 0.2 cases
```



```
MAP OF LATENT DISTRIBUTIONS AND RESPONSE MODEL PARAMETER ESTIMATES
===========================================================Build: Ju
        Dimension        Terms in the Model (excl Step terms)
--------------------
Dimension_2                                +items
-------------------------------------------------------------------
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
   5                      |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                         X|                                       |
                         X|                                       |
                          |                                       |
   4                      |                                       |
                         X|                                       |
                         X|                                       |
                       XXX|                                       |
                      XXXX|                                       |
                     XXXXX|                                       |
                       XXX|                                       |
   3                XXXXXX|                                       |
                      XXXX|                                       |
                       XXX|                                       |
                    XXXXXX|                                       |
                XXXXXXXXXX|                                       |
                XXXXXXXXXX|                                       |
                  XXXXXXXX|                                       |
              XXXXXXXXXXXX|                                       |
   2            XXXXXXXXXX|                                       |
              XXXXXXXXXXXX|                                       |
                  XXXXXXXX|                                       |
                XXXXXXXXXX|                                       |
          XXXXXXXXXXXXXXXX|                                       |
          XXXXXXXXXXXXXXXX|                                       |
                XXXXXXXXXX|10                                     |
   1       XXXXXXXXXXXXXXX|                                       |
              XXXXXXXXXXXX|                                       |
                XXXXXXXXXX|                                       |
                 XXXXXXXXX|                                       |
                XXXXXXXXXX|                                       |
                 XXXXXXXXX|1 6                                    |
                       XXX|                                       |
                       XXX|                                       |
   0                   XXX|                                       |
                       XXX|7 8                                    |
                          |                                       |
                      XXXX|                                       |
                     XXXXX|9                                      |
                          |4                                      |
                          |                                       |
  -1                     X|                                       |
                          |                                       |
                         X|                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
  -2                      |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
                          |                                       |
===================================================================
Each 'X' represents 0.2 cases
```

By dimension item difficulty maps

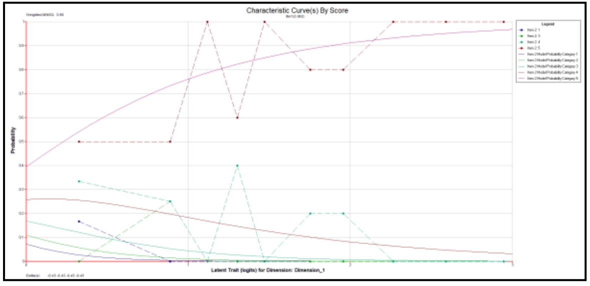


```
MAP OF LATENT DISTRIBUTIONS AND RESPONSE MODEL PARAMETER ESTIMATES
===========================================================Build: Jul 20 2020===
                  Dimension                  Terms in the Model (excl Step terms)
----------------------------------------------
          Dimension_1         Dimension_2                      +items
---------------------------------------------------------------------------------------
                          |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
   5                      |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
                          |                  X|                                       |
                          |                  X|                                       |
                          |                   |                                       |
   4                      |                   |                                       |
                          |                  X|                                       |
                          |                  X|                                       |
                         X|                XXX|                                       |
                     XXXXX|               XXXX|                                       |
                        XX|              XXXXX|                                       |
                      XXXX|                XXX|                                       |
   3                 XXXXX|             XXXXXX|                                       |
                        XX|               XXXX|                                       |
                         X|                XXX|                                       |
            XXXXXXXXXXXXXX|             XXXXXX|                                       |
                     XXXXX|         XXXXXXXXXX|                                       |
                XXXXXXXXXX|         XXXXXXXXXX|                                       |
               XXXXXXXXXXX|           XXXXXXXX|                                       |
           XXXXXXXXXXXXXXX|       XXXXXXXXXXXX|                                       |
   2             XXXXXXXXX|         XXXXXXXXXX|                                       |
            XXXXXXXXXXXXXX|       XXXXXXXXXXXX|                                       |
              XXXXXXXXXXXX|           XXXXXXXX|                                       |
      XXXXXXXXXXXXXXXXXXXX|         XXXXXXXXXX|                                       |
          XXXXXXXXXXXXXXXX|   XXXXXXXXXXXXXXXX|                                       |
       XXXXXXXXXXXXXXXXXXX|   XXXXXXXXXXXXXXXX|                                       |
             XXXXXXXXXXXXX|         XXXXXXXXXX|10                                     |
   1         XXXXXXXXXXXXX|    XXXXXXXXXXXXXXX|                                       |
       XXXXXXXXXXXXXXXXXXX|       XXXXXXXXXXXX|                                       |
                XXXXXXXXXX|         XXXXXXXXXX|                                       |
                XXXXXXXXXX|          XXXXXXXXX|5                                      |
                         X|         XXXXXXXXXX|                                       |
                          |          XXXXXXXXX|1 6                                    |
                        XX|                XXX|                                       |
                     XXXXX|                XXX|                                       |
   0                    XX|                XXX|                                       |
                          |                XXX|7 8                                    |
                          |                   |3                                      |
                          |               XXXX|2                                      |
                       XXX|              XXXXX|9                                      |
                          |                   |4                                      |
                          |                   |                                       |
  -1                      |                  X|                                       |
                          |                   |                                       |
                          |                  X|                                       |
                          |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
  -2                      |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
                          |                   |                                       |
=======================================================================================
Each 'X' represents 0.2 cases
```

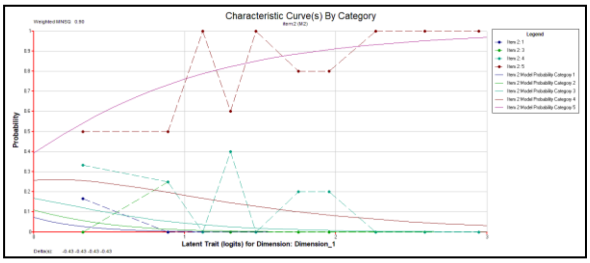
Combined item difficulty map

Appendix 3. *Within-Item Multidimensional Model Sample PlotQuest Outputs.*

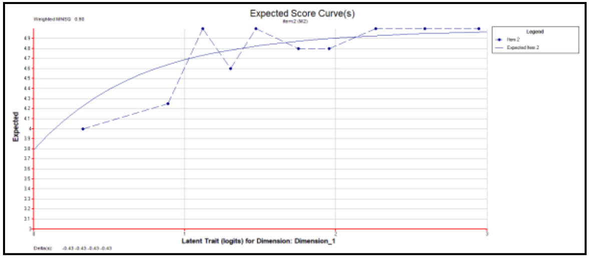
Characteristic Curves By Score



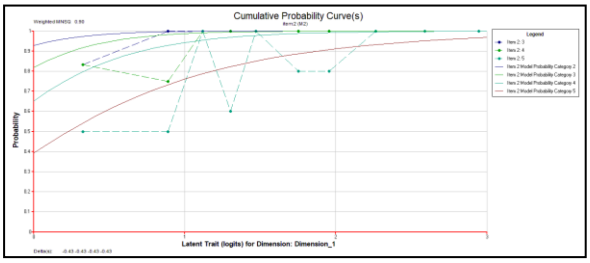
Characteristic Curves By Category



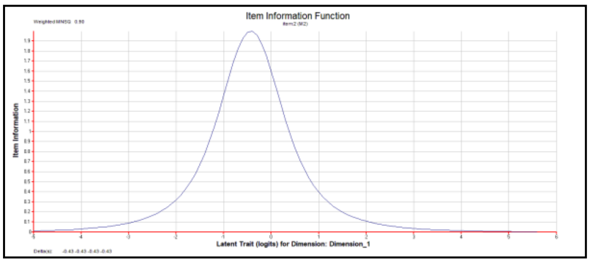
Item Expected Score Curves



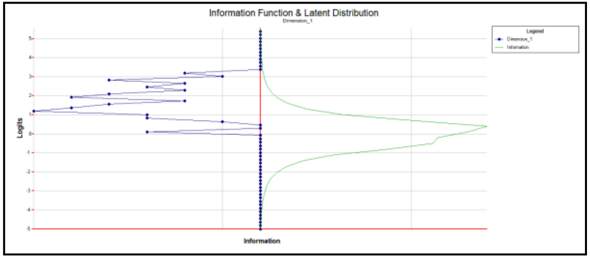
Cumulative Item Characteristic Curves



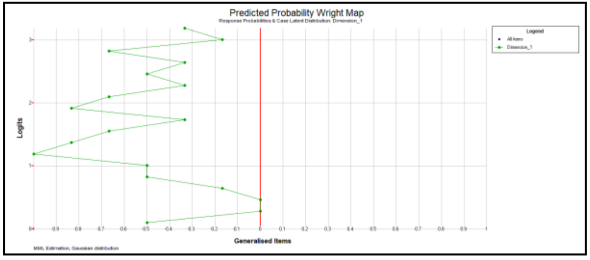
Item Information



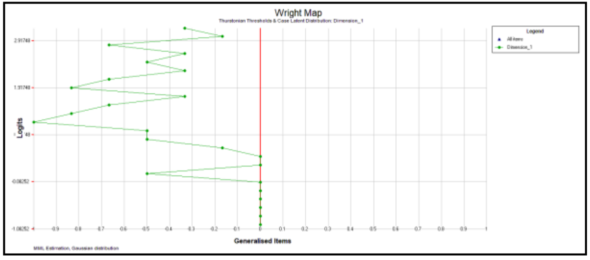
Information map



Predicted Probability Wright Map

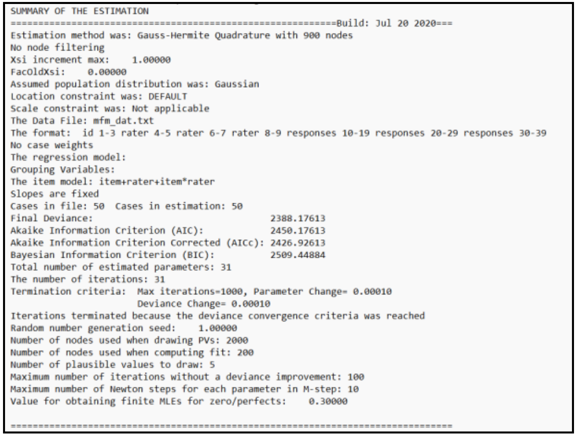


Wright Map



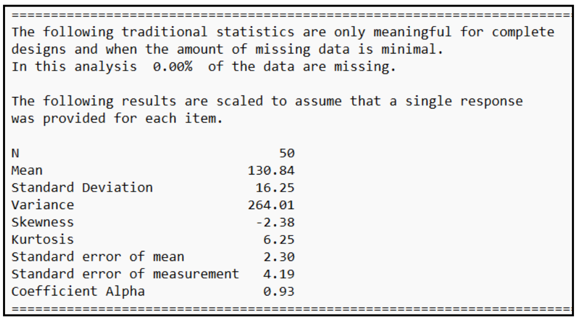
Appendix 4. *Many-facet Multidimensional Model Rasch Analysis results (wim_shw.txt), traditional item analysis results (wim_itn.txt).*

Summary of the estimation



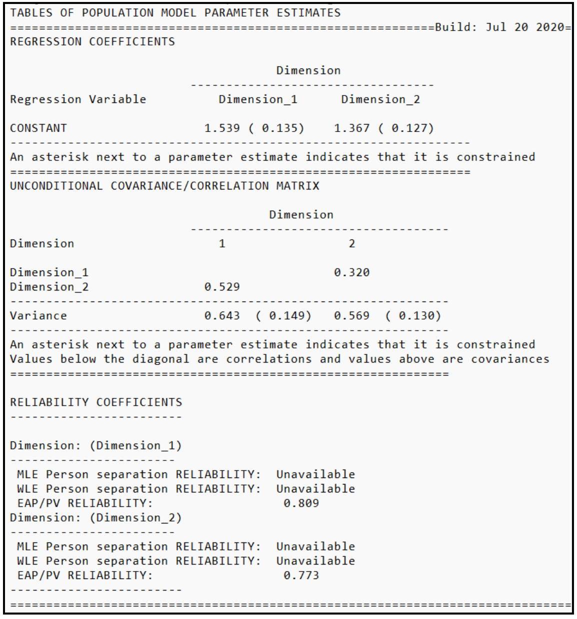
```
SUMMARY OF THE ESTIMATION
===========================================================Build: Jul 20 2020===
Estimation method was: Gauss-Hermite Quadrature with 900 nodes
No node filtering
Xsi increment max:    1.00000
FacOldXsi:    0.00000
Assumed population distribution was: Gaussian
Location constraint was: DEFAULT
Scale constraint was: Not applicable
The Data File: mfm_dat.txt
The format:  id 1-3 rater 4-5 rater 6-7 rater 8-9 responses 10-19 responses 20-29 responses 30-39
No case weights
The regression model:
Grouping Variables:
The item model: item+rater+item*rater
Slopes are fixed
Cases in file: 50  Cases in estimation: 50
Final Deviance:                               2388.17613
Akaike Information Criterion (AIC):           2450.17613
Akaike Information Criterion Corrected (AICc): 2426.92613
Bayesian Information Criterion (BIC):         2509.44884
Total number of estimated parameters: 31
The number of iterations: 31
Termination criteria:  Max iterations=1000, Parameter Change= 0.00010
                       Deviance Change= 0.00010
Iterations terminated because the deviance convergence criteria was reached
Random number generation seed:    1.00000
Number of nodes used when drawing PVs: 2000
Number of nodes used when computing fit: 200
Number of plausible values to draw: 5
Maximum number of iterations without a deviance improvement: 100
Maximum number of Newton steps for each parameter in M-step: 10
Value for obtaining finite MLEs for zero/perfects:    0.30000

================================================================================
```

Traditional test statistics



```
========================================================================
The following traditional statistics are only meaningful for complete
designs and when the amount of missing data is minimal.
In this analysis  0.00%  of the data are missing.

The following results are scaled to assume that a single response
was provided for each item.

N                                 50
Mean                          130.84
Standard Deviation             16.25
Variance                      264.01
Skewness                       -2.38
Kurtosis                        6.25
Standard error of mean          2.30
Standard error of measurement   4.19
Coefficient Alpha               0.93
========================================================================
```

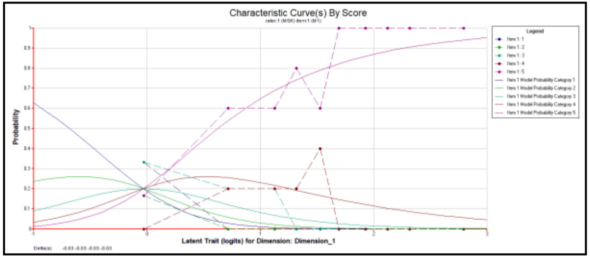
Population Model Parametre Estimates



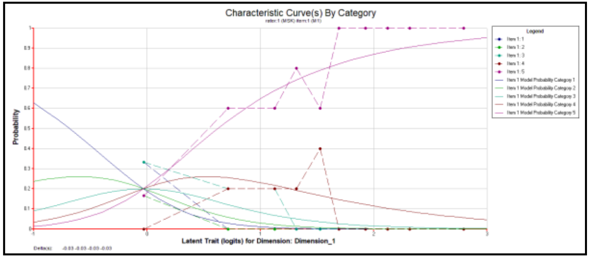
```
TABLES OF POPULATION MODEL PARAMETER ESTIMATES
==========================================================Build: Jul 20 2020=
REGRESSION COEFFICIENTS

                                     Dimension
                         ----------------------------------
Regression Variable           Dimension_1      Dimension_2

CONSTANT                     1.539 ( 0.135)    1.367 ( 0.127)
----------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
================================================================
UNCONDITIONAL COVARIANCE/CORRELATION MATRIX

                                    Dimension
                         ------------------------------------
Dimension                      1                  2

Dimension_1                                     0.320
Dimension_2                  0.529
-------------------------------------------------------------
Variance                     0.643  ( 0.149)   0.569  ( 0.130)
-------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Values below the diagonal are correlations and values above are covariances
=============================================================

RELIABILITY COEFFICIENTS
------------------------

Dimension: (Dimension_1)
------------------------
 MLE Person separation RELIABILITY:  Unavailable
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:                   0.809
Dimension: (Dimension_2)
------------------------
 MLE Person separation RELIABILITY:  Unavailable
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:                   0.773
------------------------
==========================================================================
```

Appendix 5. *Many-facet Multidimensional Model Sample PlotQuest Outputs).*

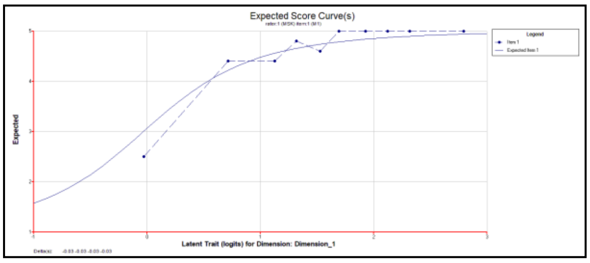
Characteristic Curves By Score



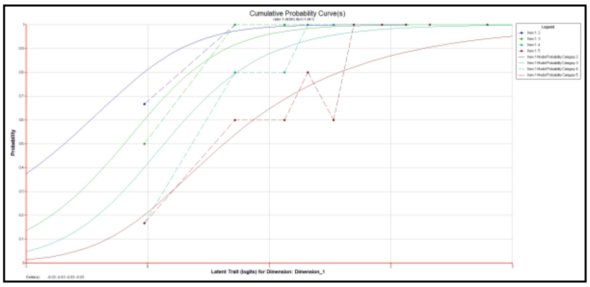
Characteristic Curves By Category



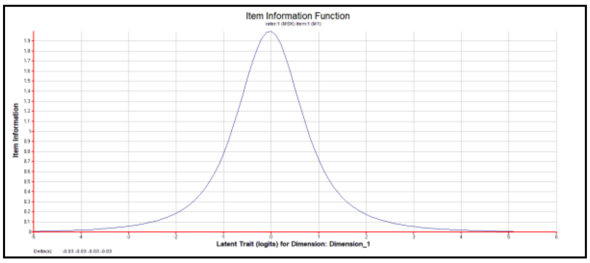
Item Expected Score Curves



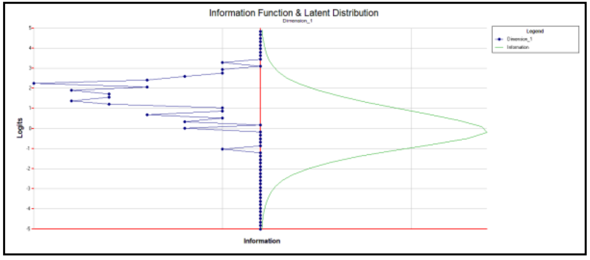
Cumulative Item Characteristic Curves



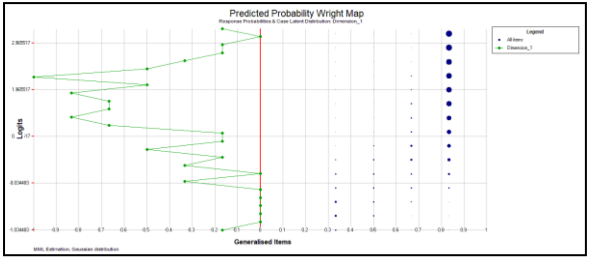
Item Information



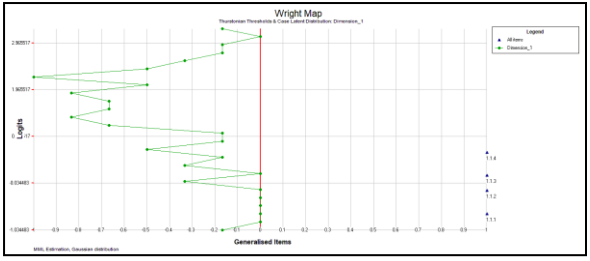
Information map



Predicted Probability Wright Map



Wright Map



International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 303–330

https://doi.org/10.21449/ijate.1217352

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

The effects of blended learning activities based on the ASSURE model in teaching on students and teachers in music lessons

Sevim Irmis Engizli [1], **Ali Korkut Uludag** [2,*]

[1]Ministry of Education, Samsun, Türkiye
[2]Ataturk University, Kazım Karabekir Faculty of Education, Department of Fine Arts Education, Erzurum, Türkiye

ARTICLE HISTORY

Received: Dec. 11, 2022
Revised: May 09, 2023
Accepted: June 08, 2023

Keywords:
ASSURE model,
Blended learning,
Music education,
Stations method,
Web 2.0 tools.

Abstract: This study, carried out during COVID-19, aimed at evaluating the effects of music lesson activities prepared with blended learning and the ASSURE instructional design model on both students and music teachers. 10th grade students (n=30) in a public school participated in the study. An action research design was prepared with a combination of learning at stations method, mobile games (Rhythm Cat, NoteWorks) and Web 2.0 tools (Thinglink, Plickers, Kahoot). "Teacher diary" and "semi-structured interview protocol" were used as data collection tools. The data were analyzed by content analysis with Maxqda software. Interrater reliability of the two experts asked to code the qualitative data was calculated to increase the reliability of the study. Following the implementation, the findings showed that the students' independent learning, motivation, collaboration, making learning fun, interaction, communication, competition, socialization and productivity improved intensively. It was understood that the emerged negativities were not related to the teaching tools used in the action research procedure but were generally related to the negative learning conditions due to the pandemic. The most notable negative aspects were difficulty in technological competence, difficulties in self-regulation, temporary anxiety, digital access and some short-term technological malfunctions. The study concludes with several recommendations and highlights the points that need further attention in such innovative research.

1. INTRODUCTION

Declared as a global pandemic by the World Health Organization in March 2020, COVID-19 has caused unprecedented devastating effects in all areas of life, including education (Cheng & Lam, 2021; Joseph & Lennox, 2021). Due to the rapid spread of the pandemic, many countries have initiated new practices and education policies in universities (Qiu et al., 2020). Mishra et al. (2020) emphasized that education should be renewed in the face of such serious conditions, and it should be redesigned accordingly. The pandemic, which appeared suddenly and resulted in unprecedented crises all around the world, has caused unexpected shifts in music education as well (Daubney & Fautley, 2020).

*CONTACT: Ali Korkut Uludag ✉ korkutuludag@atauni.edu.tr Ataturk University, Kazım Karabekir Faculty of Education, Department of Fine Arts Education, Erzurum, Türkiye

e-ISSN: 2148-7456 /© IJATE 2023

The rapid growth of information and communication technologies as an alternative to traditional learning methods has provided new opportunities for individuals on how to acquire information (Anohina, 2005). Kim (2013) suggested that today's information society has altered the education environment drifting apart from the teacher-centered, ex parte and rote learning teaching methods. Herein, Cautreels (2003) pointed out that teachers should create a student-centered learning environment instead of using traditional methods. In this sense, the perspectives and worldviews of the new-gen students brought up with a digitalized education system should be well understood (Livari et al., 2020).

Music education-related studies on COVID-19 are limited (Thorgersen & Mars, 2021). Calderón-Garrido et al. (2021) stated that music education practices should be re-evaluated in the face of the negativities caused by COVID-19, and it has become necessary to reflect the new ideas related to teaching. According to a music teacher, systematically turning technology into classroom activities during COVID-19 is an effective way to enhance students' experiences and encourage creativity that was not possible before the pandemic (Beirnes & Randles, 2022). Practical lesson activities designed through the use of music technology to stimulate students' creativity create a "fun" and "interesting" lesson atmosphere for students (Kim, 2013).

Similar problems, encountered in distance education in previous studies, have paved the way for the current research to plan the action research procedure. These problems can be listed as some students' lack of interest and attention in online lessons since they are not used to learning with smartphones and computers (Mishra et al., 2020), insufficient internet access, insufficient workspace at home, different family problems, insufficient internet infrastructure (Roman et al., 2021), inequality of opportunity in education (Adam & Metljak, 2022), inadequacy of traditional teaching methodologies (Biasutti et al., 2022), and difficulties in conveying how to do homework (Livari et al., 2020). In the face of these problems, action research based on both technology combination and innovative teaching models was designed in order to minimize the teaching and learning setbacks triggered by the pandemic.

1.2. Purpose of the Research

The current research aims at determining the meanings attributed by the students and the music teachers to the technology-supported blended learning activities designed with the ASSURE instructional design model. In studies where the ASSURE design model and blended learning activities are used together (AlNajdi, 2018; Çetinkaya, 2017; Karaduman et al., 2019; Valverde-Berrocoso & Fernández-Sánchez, 2020), it is seen that technology-supported learning tools are used. For example, AlNajdi (2018) designed a blended learning environment based on the ASSURE model in order to teach the Arabic alphabet to primary school students in the USA in an effective and fun way and concluded that students' pronunciation and reading skills improved as a result of the application. In another example, Çetinkaya (2017) organised lesson activities by combining ASSURE instructional design model, web-assisted instructional tools and a personalised blended learning model in order to help middle school students learn the "matter and heat" unit more easily and concluded that students' achievement levels increased. Kristianti et al. (2017) used the Autograph software learning approach in their pre-test-post-test control group experimental study based on the ASSURE design model, and as a result of the application, they found that the critical thinking skills of students learning mathematics improved. It is seen that technological learning tools are definitely utilised in these studies in the literature. In addition, there is no music education study in the literature in which the ASSURE design model and blended learning are used together. However, there are experimental music teaching studies prepared with blended learning methods (Adileh, 2012; Edward et al., 2018; Hietanen & Ruismäki, 2017; Ruokonen & Ruismäki, 2016). For example, Adileh (2012) compared the effects of blended and face-to-face learning methods on students' music course achievement and attitudes towards the course with experimental design research.

The results of the study showed that the blended learning group was more successful than the face-to-face (FTF) learning group in terms of both course achievement and attitudes towards music learning. The ASSURE model, blended learning, modern pedagogies and instructional technologies in the literature guided the design of the current study. All of these elements were utilised in the action research procedure of the current research. Therefore, the current research was conducted to determine the meanings attributed by the students and the music teacher to the technology-supported blended learning activities organised on the ASSURE instructional design model. For this purpose, the following research questions were sought to be answered: Given the unique features of modern instructional tools such as the ASSURE instructional design model, blended learning, Thinglink, Kahoot, Plickers, mobile games and station technique, as well as past research, this study sought to answer seven research questions. In line with the purpose of the research and the literature background, answers to the following research questions were sought.

1.3. Literature Review and Developing Research Questions

It has been determined that very few studies have been conducted on the teaching practices used in music lessons or the activities carried out during the Covid-19 pandemic period (Pozo et al., 2022). In addition, not much has been researched about online music teaching due to rapidly developing online technologies (Salvador et al., 2021). Online music teaching, which became mandatory during the Covid-19 pandemic period, has caused various difficulties for music teachers (Calderón-Garrido & Gustems-Carnicer, 2021; Gibson, 2021; Salvador et al., 2021). This painful and difficult process has forced music teachers to find new options for a qualitative change in the online education process (Adam, 2022). The importance of creative approaches and modern technologies in music education, which are thought to be a potential force for these options, has already been emphasised in innovative studies (Sastre et al., 2013). It has also been pointed out that it is necessary to find new ways to actively involve students in music education (Rosen et al., 2013). The use of modern pedagogies and new technologies in music teaching during the Covid-19 pandemic has qualities that will benefit music teachers.

1.3.1. *Effects of blended learning model on learning*

Blended learning, defined as a systematic combination of online and classroom teaching, is based on activating and supporting learning (Dziuban et al., 2004). It has become increasingly clear that blended learning can overcome various limitations and challenges associated with online learning and face-to-face teaching (Alammary et al., 2014). Blended learning has advantages such as recording materials, collaboration, problem-solving, independent learning and motivation (Higde & Aktamış, 2021). Experimental research results have shown that blended learning improves students’ learning performance at a higher level compared to traditional teaching approaches (Fazal & Bryant, 2019) and directs students to collaborative learning, web-assisted learning and independent learning (Hiğde & Aktamış, 2021). There are also some disadvantages such as difficulty in technological competence (Prasad et al., 2018) and lack of self-regulation (Rasheed et al., 2020) that may be encountered in blended learning. Based on the findings of previous research, the following first research question was formulated to evaluate the effects of blended learning on students' experiences of learning basic music theory.

RQ1: What positive or negative learning experiences did the participants gain as a result of the blended learning?

1.3.2. *Effects of Thinglink on learning*

Thinglink allows adding maps, quizzes, tables, videos, calendars and data via interactive tags on visuals in order to motivate students to the lesson (Yalçın, 2021). Thus, teachers have the opportunity to create learning methodologies that arouse students’ curiosity with Thinglink using 360/VR technology (Asatillayevna, 2022) since it is an easily accessible platform and has the potential to improve learning (Pringle et al., 2022). These features also enable Thinglink to support online and face-to-face learning efforts in hybrid learning environments (Batista et al., 2022; Edwards-Smith, 2022). It also improves students' creative thinking skills (Al Fatihah et al., 2022) by encouraging communication and critical thinking (Edwards-Smith, 2022). Based on the findings of previous research, the following second research question was formulated to evaluate the effects of the Thinglink platform on students' experiences of learning basic music theory.

RQ2: What learning experiences did the participants gain with the Thinglink app?

1.3.3. *Effects of Kahoot on learning*

Kahoot is a platform that attracts students' attention, increases their interaction, enables them to compete (Ayaz, 2019; Wang, 2015), enables them to learn while having fun (Gil et al., 2022), increases their motivation (Korkmaz & Tetik, 2018), improves teachers' technological-pedagogical content knowledge, saves time in assessment and evaluation and makes students more active (Şimşek et al., 2017). It is also effective in providing feedback (Ayaz, 2019). Korkmaz and Tetik (2018) stated that there is no negative aspect of Kahoot in their research, but they pointed out that the students experienced fear during the application as their internet quotas were running out and their internet speeds were slowing down. Based on the findings of previous research, the following third research question was formulated to evaluate the effects of the Kahoot platform on students' experiences of learning basic music theory.

RQ3: What positive or negative learning experiences did the participants gain with Kahoot?

1.3.4. *Effects of Plickers on learning*

Plickers is an effective application developed to help revitalize traditional teacher-centered classrooms (Krause et al., 2017). The most distinctive feature of Plickers is that students can participate in the application without having a personal digital device (Chng & Gurvitch, 2018) as it allows using individual student codes instead of personal mobile devices that students have to buy for the lesson (Attard & Holmes, 2020). It also allows teachers to receive instant feedback (Aguirre et al., 2019; Chng & Gurvitch, 2018; García, 2016). Teachers can quickly assess whether students understand the topics clearly or not thanks to Plickers (Aguirre et al., 2019). Göksün and Gürsoy (2019) found in their experimental research that students liked the feedback form of Plickers the most and had a lot of fun during the application. Despite these positive features of Plickers, there is also the fact that it may take some time for teachers to learn to use such technological tools correctly (Aguirre et al., 2019). In addition, Göksün and Gürsoy (2019) stated that students liked the QR code system in Plickers, but they criticized the risk of miss canning. Based on the findings of previous research, the following fourth research question was formulated to evaluate the effects of the Plickers platform on students' experiences of learning basic music theory.

RQ4: What positive or negative learning experiences did the participants gain with Plickers?

1.3.5. *Effects of mobile games on learning*

Mobile learning, whose popularity is rapidly increasing (Ronimus et al., 2020), has the potential to change the way students learn, the content of the lessons, the practices related to learning and the classroom dynamics (Gay et al., 2001). Its most distinctive feature is that it provides effective learning anywhere and anytime (Ernst et al., 2013; Ogata & Yano, 2004; OuYang et al., 2010; Palazón & Giráldez, 2018; Paule-Ruiz et al., 2017). Researchers have shown empirical evidence to highlight the benefits of mobile learning (Crompton & Burke, 2020) because mobile learning provides positive results that improve students' knowledge and skills, and increase interest and motivation with active participation (Paule-Ruiz et al., 2017). Besides the benefits of mobile learning environments, there are also some disadvantages (Torun & Dargut, 2015; Zhou et al., 2010). Examples of these disadvantages are application difficulties caused by mobile devices' small screens (Park, 2011), high cost and access problems (Cooper & Spencer, 2009). Based on the findings of previous research, the following fifth research question was formulated to assess the effects of mobile games on students' experiences of learning basic music theory.

RQ5: What positive or negative learning experiences did the participants gain with Rhythm Cat and NoteWorks mobile games?

1.3.6. *The effects of the learning at stations method on learning*

According to students' views, the learning at stations method has qualities that strengthen communication, trigger collaborative work, provide creativity and improve thinking skills (Alacapınar, 2009). As a result of research involving the teaching of socio-scientific subjects with the learning at stations method, an increase in the students' academic success and motivation was achieved, and students also found the learning at stations method practical, entertaining and remarkable (Türe et al., 2020). Eilks (2002) similarly pointed out that students' cooperative learning and motivations increased with the learning at stations method, but also stated that students might have difficulties in the first place, and they might drift away from their goals in long sessions. Continuous visiting of the stations is considered as negativity according to the students (Türe et al., 2020). In another study, some students found the activities boring and stated that their time was limited (Avcı, 2015). Students are faced with the possibility of disliking and disregarding stations since they have different learning styles (Fehrle & Schulz, 1977). Teachers may also encounter some negativities in their learning at station practices. In particular, teachers who do not know students' needs and lesson objectives may have difficulties in preparing activities regarding this method and in managing time effectively (Sears, 2007). Based on the findings of previous research, the following sixth research question was formulated to evaluate the effects of the station technique on students' basic music theory learning experiences.

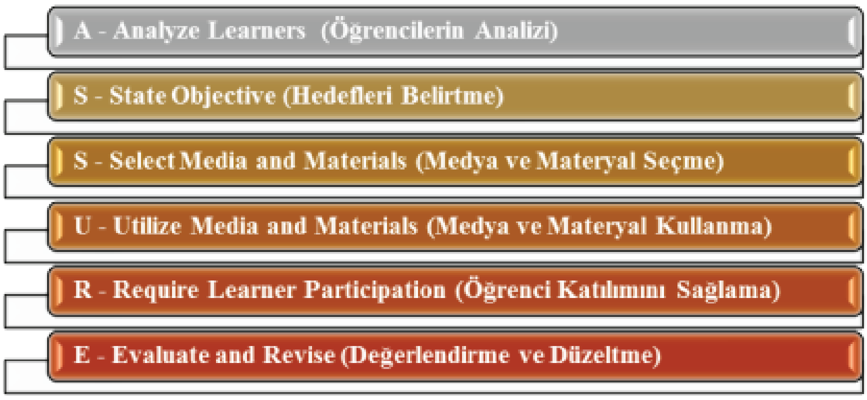
RQ6: What positive or negative learning experiences did the participants gain with the learning at stations method?

RQ7: What positive or negative teaching experiences did the music teacher gain as a result of music lesson activities?

2. METHOD

The current research was carried out as action research among qualitative research methods. Action research enables researchers to modify and alter education (Bresler, 1995). It was decided to use action research in the current research in the face of the negativities as a result of COVID-19 in music teaching since it seeks an answer to the pedagogue's question "How can I improve my application?" (Cain, 2008). It also enables the practitioner himself or a researcher to be directly involved in the process (Yıldırım & Şimşek, 2006). That is to say, the researcher in the action research can be involved in the research and can also be the data collection tool (subject) (Özerbaş et al., 2010). However, what is mentioned here is not an unmethodological process (Eroğlu, 2021). Herein, the ASSURE instructional design model was utilized to make the action research procedure more methodological. ASSURE instructional design model prioritizes education activities with intense technological content (Gündüzalp & Yıldız, 2020). This model, a six-step instructional system design, demonstrates how to select, use and assess and evaluate both the technological and instructional resources (Kim & Downey, 2016). Figure 1 below shows the steps in ASSURE instructional design model.

Figure 1. *ASSURE instructional design model* (Gündüzalp & Yıldız, 2020).


A - Analyze Learners (Öğrencilerin Analizi)
S - State Objective (Hedefleri Belirtme)
S - Select Media and Materials (Medya ve Materyal Seçme)
U - Utilize Media and Materials (Medya ve Materyal Kullanma)
R - Require Learner Participation (Öğrenci Katılımını Sağlama)
E - Evaluate and Revise (Değerlendirme ve Düzeltme)

Analyze learners: In this step, learner's skills, prior knowledge, attitudes, ages, grades and learning styles are determined (Bajracharya, 2019) to identify their characteristics (Ibrahim, 2015).

State objectives: The objectives of the lesson should be clearly stated, students should be told what to achieve at the end of the implementation (Faryadi, 2007) and the objectives should be demonstrated with behavioral terms (Batir & Sadi, 2021).

Select media and materials: The most appropriate method and lesson materials are selected to achieve the objectives (Altın, 2021).

Utilize media and materials: It is planned how the materials selected for the lesson activity will be used by the students (Batir & Sadi, 2021).

Require learner participation: Learners are encouraged to participate actively in the lesson to maintain effective teaching (Altın, 2021).

Evaluation and revision: The researcher seeks answers to the two questions in the last step: 1. Have the learners achieved the learning objectives? 2. Have the media/materials been utilized for their intended purpose? It is expected to revise the process as a result of the evaluation (Batir & Sadi, 2021). It is required to run through the whole process to revise and expected to revise if needed (Smaldino et al., 2015).

2.1. Research Sample

The research sample includes students (n=30) in 10 classrooms in a public high school in Turkey in the spring term of the 2020-2021 academic year. 16 of the participants are females and 14 are males.

2.2. Data Collection Tools

The process of collecting the research data was started by obtaining the necessary ethics committee permissions and Ministry of National Education permissions. Before the interviews were conducted, the parents of the students were contacted with the help of the school administration and teachers. An explanation text was sent to the parents via the WhatsApp application. This explanation text included comprehensive explanations about ethical permissions and the purpose of the action research procedure. All parents gave permission for their children to participate in the study and sent a parental consent form to the teacher who would conduct the study. The majority of the parents sent the parental consent form via WhatsApp application. Some parents submitted the parental consent form to the teacher themselves. The music teacher kept a diary after each study phase of the action research procedure.

2.2.1. *Semi-structured interview protocol (see Appendices)*

A semi-structured interview protocol was used to identify the students' music lesson experiences, developed by the researcher in accordance with the views of two experts having approximately 20 years of experience in music education. The interview protocol with 6 questions was sent to the experts. The experts are currently working as professors in the music education department of a state university. They suggested excluding one of the questions. The interviews were carried out individually by the music teacher in the music classroom. All the interviews were recorded with the students' and parents' consent. These recordings were transcribed and transferred to Maxqda software.

2.2.2. *Teacher diary*

The music teacher kept a diary after each activity. The activities, teaching experiences and students' learning experiences were evaluated in these diaries. The primary aim was to reveal the positive or negative teaching experiences of the teacher for the process. These diaries were also transferred to Maxqda software.

2.3. Data Analysis

All qualitative data were analyzed with content analysis. The themes, categories and codes determined according to each interview question were associated with student answers on the MAXmaps folder, classified and transformed into relationship maps. While developing the relationship maps, the coded sections called for each code on the Maxqda software were ordered according to the highest weight score. The questions in the semi-structured interview form were prepared to answer the first, second, third, fourth, fifth and sixth research questions. The data obtained from teacher diaries were used to answer the seventh research question.

2.4. Validity and Reliability

Inter-rater reliability was calculated to increase the reliability of the research. A chart called "Sample Code Definition Revisions" (as cited in Creswell, 2019) developed by Guest, Bunce and Johnson (2006) was used. The code table created for our research is shown in Table 1 below.

Table 1. *Code chart.*

Code	Fun learning
Code Definition	Whether the students turned to the concept of "learning with fun" after the action research procedure.
Full Definition	Students learn the subjects included in the action research procedure while having fun.
Sample Quotes	"Action research procedure provided me fun learning in music class"
Usage Time	When students actually mention the word "fun learning" or a synonym
Time to Use	When students' talk about fun learning cannot be reasonably interpreted

In the data analysis phase, "Cohen Kappa" values were calculated to find the inter-coder agreement value. Viera and Garret (2005) stated the agreement value ranges as "poor agreement" if .20 or less than .20, "below average agreement" if between .21-.40, "moderate agreement" if between .41-.60, "good agreement" if between .61-.80 and "very good agreement" if between .81-1.00. The result of the agreement value between the coders showed that there was a 91.48% agreement. Miles et al. (2013) suggested reaching a rate of 85% to 90% for inter-coder agreement (as cited in Creswell, 2019).

2.5. ASSURE Instructional Design Model Process Steps

In this section, all stages of the ASSURE Instructional Design Model and implementation steps based on the action research procedure are included.

2.5.1. *Analyze learners*

Students have encountered technology-related learning problems in the face of the negative effects of the COVID-19 pandemic process. These problems negatively affected students' academic achievement levels, motivation and interest in the course. There is no disabled student in the study group. It has been determined that students have different cultural characteristics and cultural groups. The socio-economic status of the students is generally moderate and they have been living in the city for a long time. Students continued their previous music lessons with traditional teacher-centered teaching methods.

2.5.2. *State objectives*

The teaching objectives of the research were planned as follows:

1. Collaborative working aspects of students will be developed with the station technique.
2. Students will develop basic music theory topics with mobile games.
3. Students will learn the instrument types and music genres used in Turkish music with the Thinglink platform.
4. The music teacher will evaluate the students' learning levels with the Plickers and Kahoot platforms.

2.5.3. *Select materials*

The teacher chose Thinglink, Kahoot, Plickers Web 2.0 tools, Rhythm Cat and NoteWorks mobile games and station techniques for the action research procedure process.

2.5.4. *Utilize materials - action research procedure*

The purpose of the action research procedure is to reveal students' basic music theory experiences. The teacher encouraged students to work both individually and in groups by using various Web 2.0 tools, mobile games and station techniques. Conducting action research mainly involved seven steps, which are discussed in detail below.

2.5.4.1. Thinglink App. This phase was carried out in the form of face-to-face training. Measure types were taught to the students. Thinglink material was prepared as follows: First, world and Türkiye maps were transferred to the Thinglink platform. Colored icons were placed on countries and cities. These icons are loaded with the URL of a musical piece from YouTube. Each piece of music represented different types of measures. Information about the measurement type is written in the description section. On the left part of the statement, a visual about the subject was placed. Students learned the subjects by touching these icons. While the students practiced individually, other students were allowed to repeat the topics.

2.5.4.2. Kahoot App. At this stage, an online exam was conducted on the Kahoot platform. The exam covered the first step topics. Before the lesson, multiple choice questions were prepared on Kahoot. These questions were supported by audio and videos. Because the teacher used Kahoot in the online lesson, he entered the "Teach" game option. In this section, the classic game part was chosen. Students were sent a PIN for the game. Students entered the first tab (Play Kahoot! – Enter game PIN here) by typing “kahoot.it” into Google. After the students entered the PIN, they wrote their names in the "Nickname" field. After these procedures, the teacher started the exam called "World music genres and instrument types" from the "My Kahoots" section.

2.5.4.3. Plickers App. In the third step, the Plickers application was utilized in the form of face-to-face education with students. The teacher prepared multiple choice questions on Plickers. The questions were prepared by considering basic music theory topics. The teacher added pictures and videos next to the questions. He then wrote the students' names in the class list section on the Plickers platform. After typing the names, he printed out the QR cards. He distributed these cards to the students according to the numbers on the student list. All transactions are stored as data within the platform. Scores and reports were shown to the students.

2.5.4.4. Kahoot App. Kahoot application was carried out in the form of face-to-face training. Exam questions were prepared by considering basic music theory topics. In this step, the processes in the distance education Kahoot application were followed.

2.5.4.5. NoteWorks Mobil Game App. The teacher first installed the NoteWorks mobile game on his phone. He mirrored his phone to the smart board with the HDMI cable. Students applied the mobile game individually and in turn. Students generally preferred to look at the phone screen during the game. Other students in the class followed the application on the smart board. This strategy allowed students to reinforce the topics. The practitioner tried to know the notes falling in sequence on the stave. Afterwards, he tried to touch the notes on the piano. Each practitioner tried to get the highest score with this game based on quick thinking**.**

2.5.4.6. Rhythm Cat Mobil Game App. In this step, the mobile game Rhythm Cat was applied to teach students basic musical rhythms. The application was made face-to-face in some classes and in the form of distance education in some classes. Students tried to progress through gradually rising levels and collect points while studying basic music rhythms. During the game, the notes corresponding to the correct touches turned green. The wrong touches turned black. Rhythm Cat has high quality soundtracks that include a variety of tempos and musical styles. During the game, the students tried to harmonize the rhythms and the game music. It is necessary to ensure this harmony at the basis of the game. Students practiced the game individually. Other students accompanied the rhythms they saw on the screen by applauding.

2.5.4.7. Station Technique App. In the final step, a station technique study was carried out covering basic music theory topics. 5 stations were created in the classroom. Students were distributed equally to these stations. Students created a discussion environment and determined a chief for each station. The stations were given the task of creating a concept map. In addition, different colored cardboard and colored felt-tip pens were distributed to the stations. Each station was given 10 minutes of working time. Station chiefs remained stationary at their stations until the end of the work cycle. After each 10-minute practice session, station members moved to the next station. The station chiefs gave information about the work done in the previous session to the students who came to their stations. The tasks assigned to the stations are as follows: First station: measure types, second station: world music genres, third station: instrumental ensembles, fourth station: rhythm information and fifth station: basic music theory.

2.5.5. *Require learner participation*

A WhatsApp group was set up involving the parents to ensure student participation. The parents were informed about the action research procedure.

2.5.6. *Evaluate and revise*

In this step, the teacher evaluated whether students achieved the objectives, the adequacy and effectiveness of the media and materials, and the degree of student participation (Heinich et al., 2002). The teacher did not need to revise the procedure as a result of these evaluations.

3. FINDINGS

"Code Sub-Model" was used in all figures in which the data were processed. All codes were organised and transformed into figures in the "Creative Coding" section. While creating the figures, label connection lines with code frequency were used. These lines moving over the main codes defined the sub-codes. The line width showing the codes reflects the code frequency. The frequencies of the codes are shown on these connection lines. Under the code labels, the ordinal numbers of the students who expressed opinions on the relevant code were shown in order according to the word frequency. However, the ordinal number and word count of the student with the most intense opinion were emphasised first. These relationship maps can be explained as a code distribution model for the participants who expressed positive or negative opinions. In this part of the study, firstly, the findings related to the first research question "What positive or negative learning experiences did the participants gain as a result of the blended learning?" are presented. Figure 1 below shows the findings of the students' blended learning environments and music lesson learning experiences in the form of a relationship map. It is seen that positive and negative codes are formed within the blended learning experiences of the students. There were 6 sub-codes under the code of positive blended learning experiences and 3 sub-codes under the code of negative blended learning experiences.

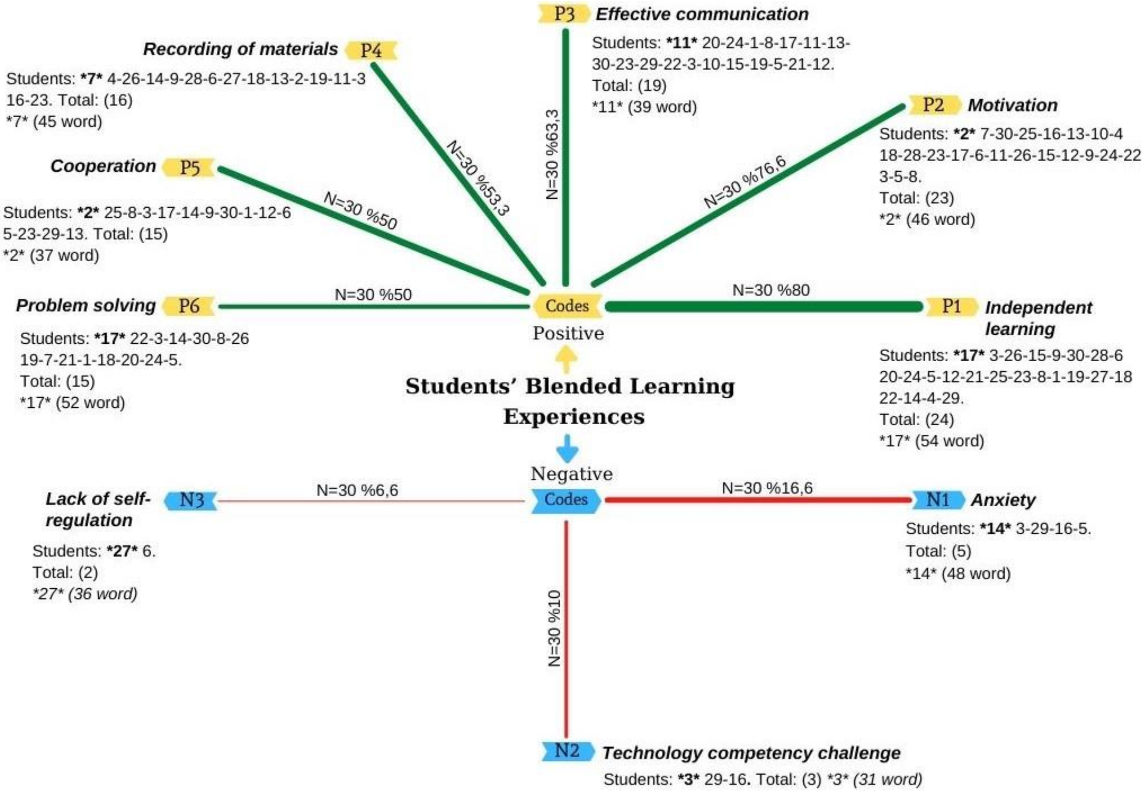
3.1. Findings About the First Research Question

As emphasized in Figure 2, positive codes are more than negative codes in terms of both student distribution and code diversity. This indicates that the students adopted blended learning environments. Students (N=30, f=24), under the main code showing positive views, were orientated towards the sub-code named P1 "Independent Learning" at the highest level with a rate of 80%. The students (f=15) showed the least distribution under this code on the sub-codes named P5 "Cooperation" with a rate of 50% and P6 "Problem Solving" with a rate of 50% (f=15). Students (f=14) exhibited the most intense distribution under the main code showing negative views under the sub-code named N1 "Anxiety" with a rate of 46.6%. When the students maintained their studies independently of the instructor, they encountered a lack of self-regulation skills in organizing and managing the process. When the students maintained their studies independently of the instructor, they encountered a lack of self-regulation skills in organizing and managing the process. The opinions of the students who turned to these codes are as follows:

“Blended learning had a positive impact on my motivation. I learned the subjects on my own. I had very good communication with my teacher. Including my friends. We also worked as a team with my friends. Our motivation has increased. Thanks to blended learning, we have overcome the problems. I loved these works (student number 17).”

“I had shortcomings in blended learning studies. I was not successful in this process. But it was an important opportunity for us to learn individually. The materials were recorded. This was an important occasion for me (student number 27).” Students' blended learning experiences were given in Figure 2 below.

Figure 2. *Relationship map of students' blended learning environments and music lesson learning experiences.*

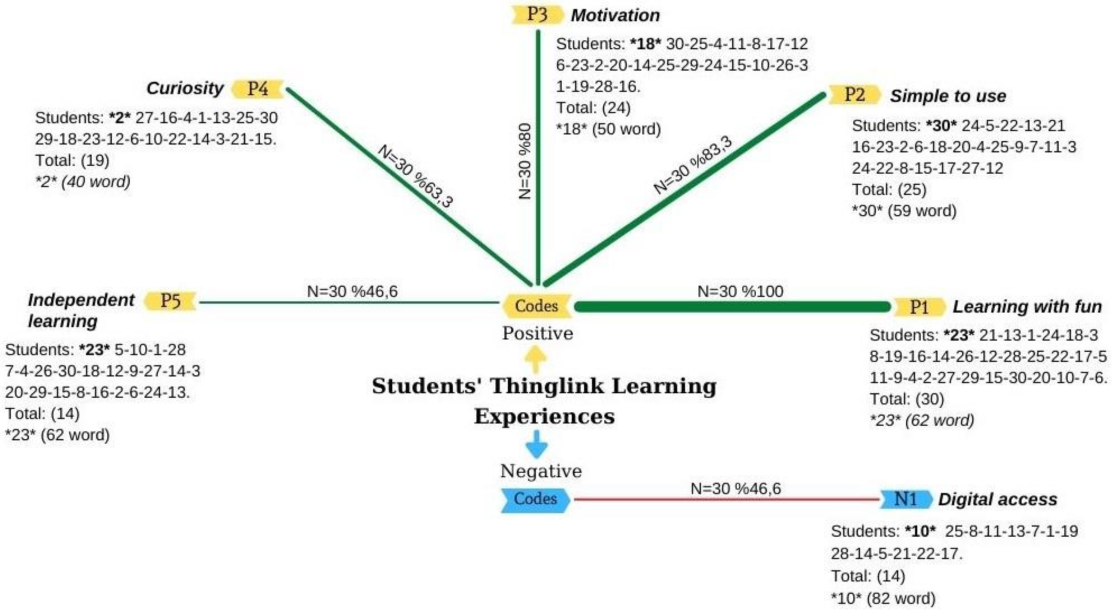

P3 Effective communication
Students: *11* 20-24-1-8-17-11-13-30-23-29-22-3-10-15-19-5-21-12.
Total: (19)
11 (39 word)
N=30 %63,3
Recording of materials P4
Students: *7* 4-26-14-9-28-6-27-18-13-2-19-11-3 16-23. Total: (16)
7 (45 word)
N=30 %53,3
P2 Motivation
Students: *2* 7-30-25-16-13-10-4 18-28-23-17-6-11-26-15-12-9-24-22 3-5-8.
Total: (23)
2 (46 word)
N=30 %76,6
Cooperation P5
Students: *2* 25-8-3-17-14-9-30-1-12-6 5-23-29-13. Total: (15)
2 (37 word)
N=30 %50
Problem solving P6
Students: *17* 22-3-14-30-8-26 19-7-21-1-18-20-24-5.
Total: (15)
17 (52 word)
N=30 %50
Codes
Positive
N=30 %80
P1 Independent learning
Students: *17* 3-26-15-9-30-28-6 20-24-5-12-21-25-23-8-1-19-27-18 22-14-4-29.
Total: (24)
17 (54 word)
Students' Blended Learning Experiences
Negative
Codes
Lack of self-regulation N3
Students: *27* 6.
Total: (2)
27 (36 word)
N=30 %6,6
N=30 %16,6
N1 Anxiety
Students: *14* 3-29-16-5.
Total: (5)
14 (48 word)
N=30 %10
N2 Technology competency challenge
Students: *3* 29-16. Total: (3) *3* (31 word)

3.2. Findings About the Second Research Question

In this section, the findings related to the second research question "What learning experiences did the participants gain with the Thinglink app?" are presented. According to Figure 3, students developed positive codes intensively. All students (N=30, f=30), under the main code showing positive views, were orientated towards the sub-code P1 "Learning with Fun" at the highest level with a rate of 100%. Students (f=14) showed the least distribution under the code showing positive views on the sub-code P5 "Independent Learning" with a rate of 46,6%. The students (f=14), under the code showing negative views, were orientated towards the sub-code named N1 "Digital Access" at the highest level with a rate of 46,6%. It is remarkable that all students adopted the "learning with fun" (P1) code. Even though the students focused heavily on positive codes, they showed an equal distribution between negative code "digital access" (N1) and positive code "curiosity" (P4). This can be interpreted that focusing heavily on positive codes does not mean eliminating the possibility of encountering the "digital access" problem, frequently experienced in distance education. Moreover, student number 10, tending towards the positive P1, P3, P4 and P5 codes, also focused on the negative code "digital access" with the highest word distribution. The fact that a student focused on both positive and negative codes with such an approach and distribution showed that his/her attitude was objective. The opinion of student number 10 is as follows:

"Before starting these studies, our teacher had mentioned that he would make a Thinglink application in the lesson. I was wondering about this program. Out of curiosity, I watched videos about Thinglink. After watching the video, my curiosity increased even more. While the applications were being made, I could not understand how the time passed with Thinglink. I also had a lot of fun during the lesson. This study increased my motivation towards the course. Thinglink strengthened my communication with my teacher (student number 10)." Students' Thinglink learning experiences were given in Figure 3 below.

Figure 3. *Relationship map of students' music lesson learning experiences with the Thinglink app.*


P3 Motivation
Students: *18* 30-25-4-11-8-17-12 6-23-2-20-14-25-29-24-15-10-26-3 1-19-28-16.
Total: (24)
18 (50 word)
N=30 %80
Curiosity P4
Students: *2* 27-16-4-1-13-25-30 29-18-23-12-6-10-22-14-3-21-15.
Total: (19)
2 (40 word)
N=30 %63,3
P2 Simple to use
Students: *30* 24-5-22-13-21 16-23-2-6-18-20-4-25-9-7-11-3 24-22-8-15-17-27-12
Total: (25)
30 (59 word)
N=30 %83,3
Independent learning P5
N=30 %46,6
Students: *23* 5-10-1-28 7-4-26-30-18-12-9-27-14-3 20-29-15-8-16-2-6-24-13.
Total: (14)
23 (62 word)
Codes
Positive
N=30 %100
P1 Learning with fun
Students: *23* 21-13-1-24-18-3 8-19-16-14-26-12-28-25-22-17-5 11-9-4-2-27-29-15-30-20-10-7-6.
Total: (30)
23 (62 word)
Students' Thinglink Learning Experiences
Negative
Codes
N=30 %46,6
N1 Digital access
Students: *10* 25-8-11-13-7-1-19 28-14-5-21-22-17.
Total: (14)
10 (82 word)

3.3. Findings About the Third Research Question

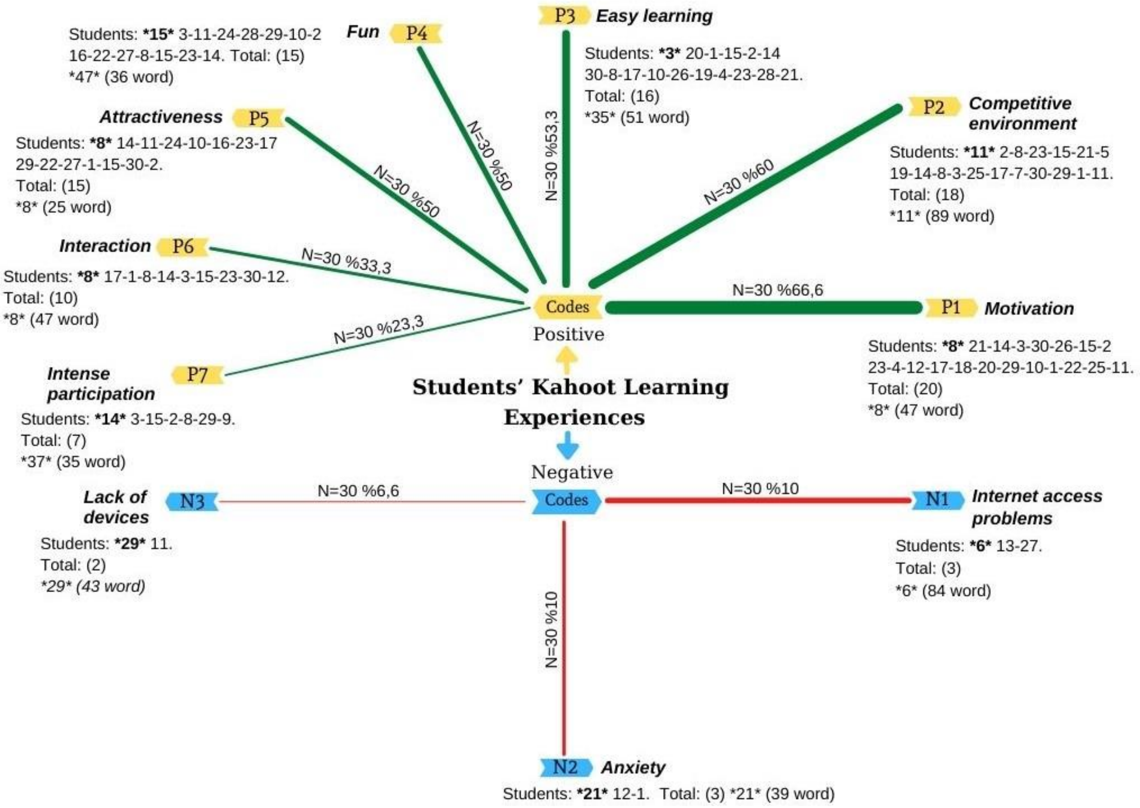
In this section, the findings related to the third research question "What positive or negative learning experiences did the participants gain with Kahoot?" are presented. When Figure 4 is examined, it is seen that the distribution of the students was on 7 positive and 3 negative codes. The students (N=30, f=20), with a rate of 66,6% under the code showing positive views, were orientated towards the sub-code named P1 "Motivation" at the highest level. Students (f=7) exhibited the least distribution under the code showing positive views on the sub-code named P7 "Intense Participation" with a rate of 23,3%. The students (f=3), under the code, indicating negative views, orientated towards the sub-code named N1 "Internet Access" at the highest level with a rate of 10%. In addition, three new types of negative behaviour emerged-which had not been seen in the studies reviewed. An intersection was observed between students' positive and negative codes. For instance, it is seen that student number 11 went for the positive code "competitive environment" (P2) with the highest word distribution, s/he also went for the negative code "device supply" (N3). Thinking that Kahoot increases the competition in the classroom, student number 11 drew attention to the problem s/he had with the supply of devices, which shows that her awareness was high. The opinion of student number 11 is as follows:

"We had a lot of fun with my friends in Kahoot app. I wish all my friends had the opportunity to participate in this application. I'm so sorry they couldn't attend. Because some of my friends could not participate in this fun application due to the problems caused by distance education. It was a very fun app. It was a fun and exciting competition between us and our friends. We had so much fun competing with each other. My motivation has increased a lot. So, I got high scores (student number 11)."

An issue that should be emphasized is the possibility that the few negative participation codes in Kahoot were related to the problems in distance education. Depending on this experience, it has become inevitable to face internet access problems for some students as the "Internet access problem" is one of the most basic problems of the distance education process. It is seen that the students numbered 6, 13 and 27 who had internet problems did not go for any positive code except for the N1 code. The opinion of student number 6 is as follows:

"I had problems with internet access from time to time in most of my classes. I could not participate in Kahoot applications due to problems caused by a poor internet connection at home. I tried to follow my music lesson and other lessons through EBA. But most of the time I had problems in terms of the internet. I had problems connecting to the system. Most of the time I could not connect to the system at all. I'm so sorry for the problems I've had (student number 6).” Students' blended learning experiences were given in Figure 4 below.

Figure 4. *Relationship map of students' online exam experiences with Kahoot.*

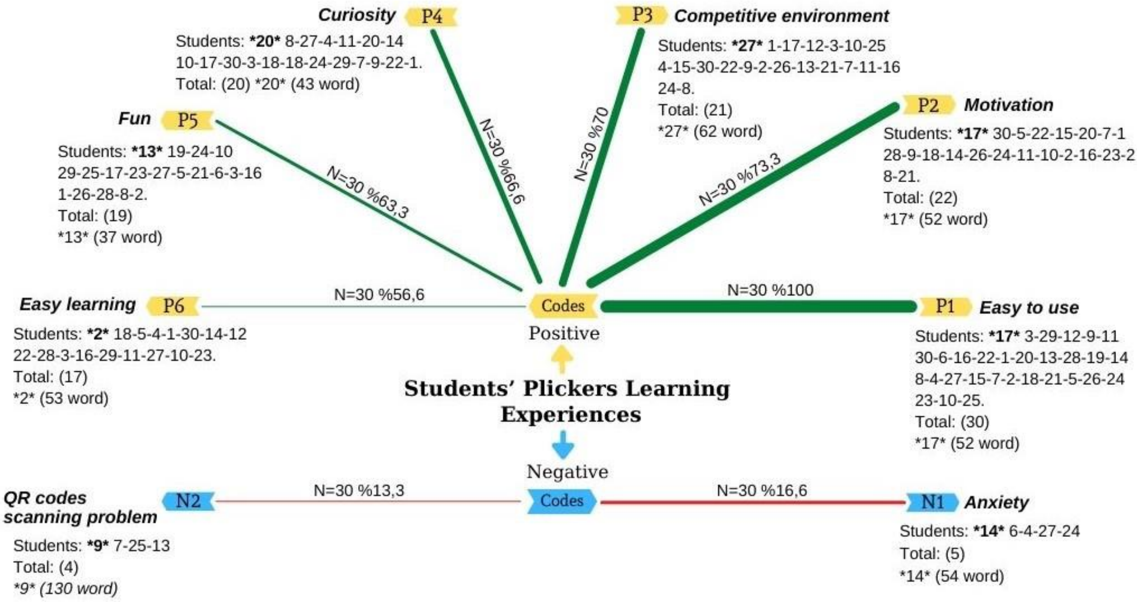

P3 Easy learning
Students: *3* 20-1-15-2-14 30-8-17-10-26-19-4-23-28-21. Total: (16) *35* (51 word)
Fun P4
Students: *15* 3-11-24-28-29-10-2 16-22-27-8-15-23-14. Total: (15) *47* (36 word)
Attractiveness P5
Students: *8* 14-11-24-10-16-23-17 29-22-27-1-15-30-2. Total: (15) *8* (25 word)
P2 Competitive environment
Students: *11* 2-8-23-15-21-5 19-14-8-3-25-17-7-30-29-1-11. Total: (18) *11* (89 word)
N=30 %50
N=30 %50
N=30 %53,3
N=30 %60
Interaction P6
Students: *8* 17-1-8-14-3-15-23-30-12. Total: (10) *8* (47 word)
N=30 %33,3
N=30 %23,3
N=30 %66,6
Codes
Positive
P1 Motivation
Students: *8* 21-14-3-30-26-15-2 23-4-12-17-18-20-29-10-1-22-25-11. Total: (20) *8* (47 word)
Intense participation P7
Students: *14* 3-15-2-8-29-9. Total: (7) *37* (35 word)
Students' Kahoot Learning Experiences
Negative
Codes
Lack of devices N3
N=30 %6,6
N=30 %10
N1 Internet access problems
Students: *29* 11. Total: (2) *29* (43 word)
Students: *6* 13-27. Total: (3) *6* (84 word)
N=30 %10
N2 Anxiety
Students: *21* 12-1. Total: (3) *21* (39 word)

3.4. Findings About the Fourth Research Question

In this section, the findings related to the fourth research question "What positive or negative learning experiences did the participants gain with Plickers?" are presented. As seen in Figure 5, all of the students went for the positive P1 code. All students (N=30, f=30), under the code showing positive views, were orientated towards the sub-code P1 "Easy to Use" at the highest level with a rate of 100%. The students (f=7) showed the least distribution under the code showing positive views on the sub-code named P6 "Easy Learning" with a rate of 23,3%. Students (f=5), under the code indicating negative opinions, were orientated towards the sub-code named N1 "Anxiety" at the highest level with a rate of 16,6%. In other learning experiences, no positive or negative code emerged in which the students participated in full numbers. It was also observed that student number 17, who had the highest word distribution in the “easy to use” code, went for all positive codes. The fact that all students used the Plickers easily showed that they continued this activity without any problems. In addition, even if a negative technical problem such as “scanning QR codes” arose, the students still pointed out that the app was easy to use. It was seen that this problem is related to the structure of the classroom and the seating arrangements of some students. As in Kahoot, the students showed great interest in the code “competitive environment” in their Plickers experience (n=21). Student number 9 went for the negative code “miss canning QR-code” and also the positive P1, P2, P3 and P4 codes. The opinion of student number 9 is as follows:

“Our teacher taught us the Plickers program in face-to-face training. In my opinion, it is a very easy and practical program. I was very curious about this program. All we had to do was to raise the cards that our teacher had given us, with the answer option facing up. But since I was sitting by the window, the answers of me and a few of my friends were not reflected on the screen. My teacher said it was due to the glare on the cards. On the next try, when our teacher changed our places, the problem disappeared completely. The reflection of the answer options on the screen created a sweet competition between us. I can easily say that the application that increases my motivation the most is Plickers (student number 9).” Students' Plickers learning experiences were given in Figure 5 below.

Figure 5. *Relationship map of students’ Plickers and online exam experiences.*

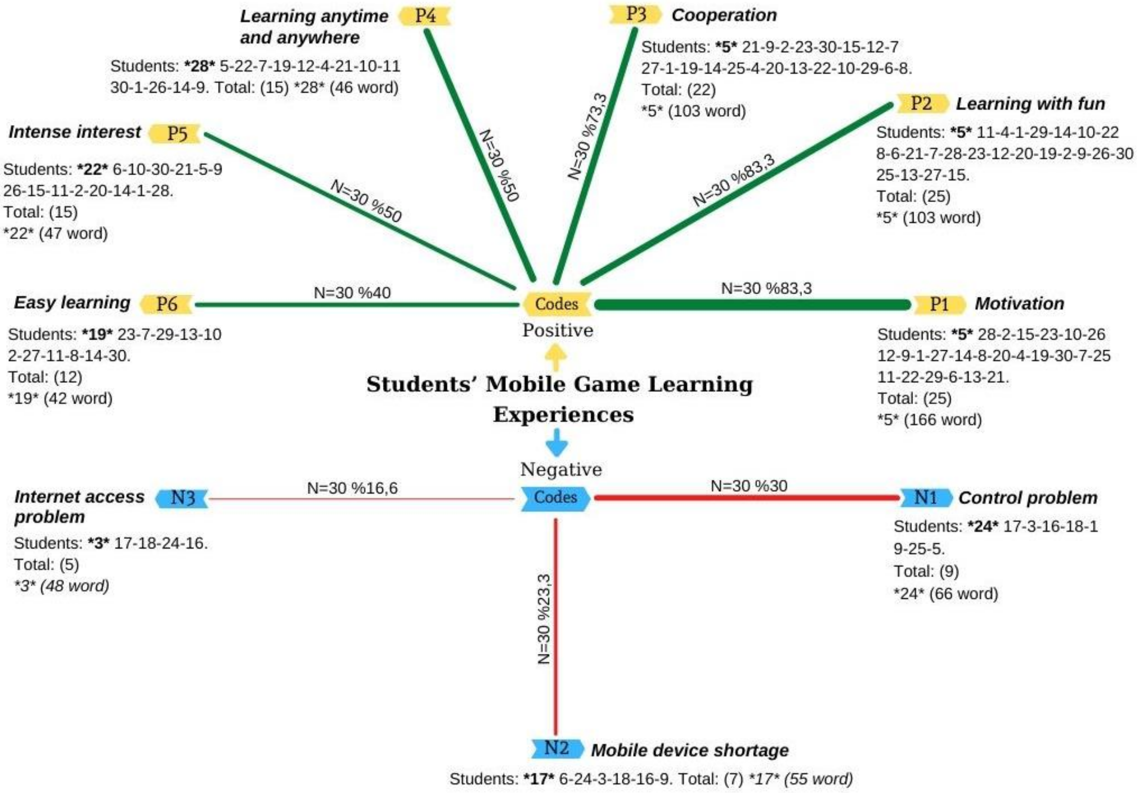


3.5. Findings About the Fifth Research Question

In this section, the findings related to the fifth research question "What positive or negative learning experiences did the participants gain with Rhythm Cat and NoteWorks mobile games?" are presented. According to Figure 6, students went for positive codes mostly and showed an equal distribution between “motivation” (P1) and “learning with fun” (P2) codes. The students also showed an equal distribution in “learning anytime anywhere” (P4) and “deep interest” (P5) codes. Students (N=30, f=25), under the code showing positive views, were orientated towards the sub-code named P1 "Motivation" at the highest level with a rate of 83.3%. Students (f=12) exhibited the lowest distribution under the code showing positive views on the sub-code named P6 "Easy Learning" with a rate of 40%. The students (f=9), under the code showing negative views, were orientated towards the sub-code named N1 "Control Problem" at the highest level with a rate of 30%. Unlike other learning experiences, the students went for the code “Collaboration”, an important function of modern pedagogy. Another remarkable finding is that the student with the highest word distribution in P1, P2 and P3 codes is student number 5. It is seen that student number 5 also went for the “quick and easy learning” (P4) code with the word distribution ratio in the second rank. This student was also the student with the highest word distribution among all learning experiences throughout the procedure. It is understood that student number 5 evaluated mobile game apps in all aspects. The students numbered “3, 16, 17, 18 and 24” went for all negative codes and did not go for any positive codes, which showed that they had significant and permanent problems with mobile game applications. The opinion of student number 5 is as follows:

“Thanks to Rhythm Cat and NoteWorks games, I was able to learn some subjects that I did not understand in previous music lessons very quickly and easily. Thanks to these games, I understood these topics better. Also, lots of fun games. I believe that playing such games in class increases my motivation. Not only me, but I think all my friends had a lot of fun in the lesson. These games add a different atmosphere to the lesson. I would love to have these games implemented in other lessons. Music lesson has become a very different lesson for me now. Frankly, I used to not care much about music lessons. I had the opportunity to play these games at home. I would also like to point out that; In my first mobile game application, I had a minor control problem with the touchscreen as my own phone was a bit small. But when I got used to the games in a short time, this problem disappeared (student number 5).” Students' Mobil games learning experiences were given in Figure 6 below.

Figure 6. *Relationship map of students’ mobile games and basic music theory learning experiences.*

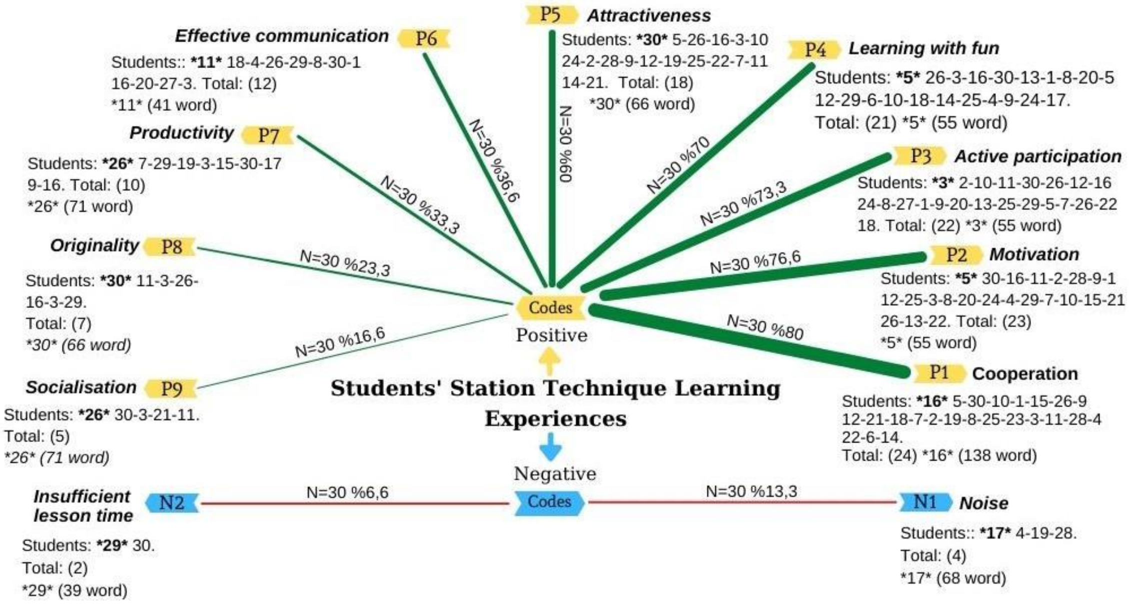

Learning anytime and anywhere
P4
Students: *28* 5-22-7-19-12-4-21-10-11 30-1-26-14-9. Total: (15) *28* (46 word)
P3
Cooperation
Students: *5* 21-9-2-23-30-15-12-7 27-1-19-14-25-4-20-13-22-10-29-6-8.
Total: (22)
5 (103 word)
P2
Learning with fun
Students: *5* 11-4-1-29-14-10-22 8-6-21-7-28-23-12-20-19-2-9-26-30 25-13-27-15.
Total: (25)
5 (103 word)
Intense interest
P5
Students: *22* 6-10-30-21-5-9 26-15-11-2-20-14-1-28.
Total: (15)
22 (47 word)
N=30 %50
N=30 %50
N=30 %73,3
N=30 %83,3
Easy learning
P6
N=30 %40
Codes
Positive
N=30 %83,3
P1
Motivation
Students: *19* 23-7-29-13-10 2-27-11-8-14-30.
Total: (12)
19 (42 word)
Students: *5* 28-2-15-23-10-26 12-9-1-27-14-8-20-4-19-30-7-25 11-22-29-6-13-21.
Total: (25)
5 (166 word)
Students’ Mobile Game Learning Experiences
Negative
Internet access problem
N3
N=30 %16,6
Codes
N=30 %30
N1
Control problem
Students: *3* 17-18-24-16.
Total: (5)
3 (48 word)
Students: *24* 17-3-16-18-1 9-25-5.
Total: (9)
24 (66 word)
N=30 %23,3
N2
Mobile device shortage
Students: *17* 6-24-3-18-16-9. Total: (7) *17* (55 word)

3.6. Findings About the Sixth Research Question

In this section, the findings related to the sixth research question "What positive or negative learning experiences did the participants gain with the learning at stations method?" are presented. As seen in Figure 7, students went for positive codes the most in the learning at stations method practice among all learning experiences. “Collaboration” (P1) was the code for which students have the most positive distribution in this experience. Students (N=30, f=24), under the positively oriented main code, oriented towards the sub-code named P1 "Cooperation" at the highest level with a rate of 80%. Students (f=5) exhibited the lowest distribution under this main code with a rate of 16,6% on the sub-code named P9 "Socialisation". Students (f=4), with a rate of 13,3% under the code indicating negative views, were orientated towards the sub-code named N1 "Noise" at the highest level. The choice of the collaboration code mostly by the students showed that they achieved the learning outcome aimed by modern pedagogy. In addition, the students showed a close distribution in the “collaboration” (P1), “motivation” (P2), “active participation” (P3) and “learning with fun” (P4) codes. It was understood that the negative codes were not related to the content of learning at stations method practice since the negative codes “noise” (N1) and “inadequacy of course periods" (N2) negatively affected students' other tasks during the application. It was observed that these students who expressed negative opinions obviously went for positive codes as well. It was determined that students, unlike other applications, went for the codes "active participation" (P3), "productivity" (P7), "authenticity" (P8) and "socialization" (P9) in their learning at stations method learning experiences. These codes were not found in the learning experiences of other applications. The opinions of student number 16, who has the most positive word distribution, on the "collaboration" (P1) code are as follows:

"With this interesting station technique work that our teacher prepared for us in the music lesson, we had a fun time in the lesson. As a result of the station work, I was able to help most of my friends. I believe that we are doing original work by supporting each other with our friends. This has greatly increased my motivation and energy. We exchanged information with our friends during the activities. As a result, we worked intensely as a team. I believe that we are very active and very productive in station technique work. The exercises were so much fun. It was also very instructive. Our communication with our teacher has increased. At the same time, our social relations increased. I enjoyed this lesson very much. I don't think the station technique has any negative aspects (student number 5)." Students' station technique learning experiences were given in Figure 7 below.

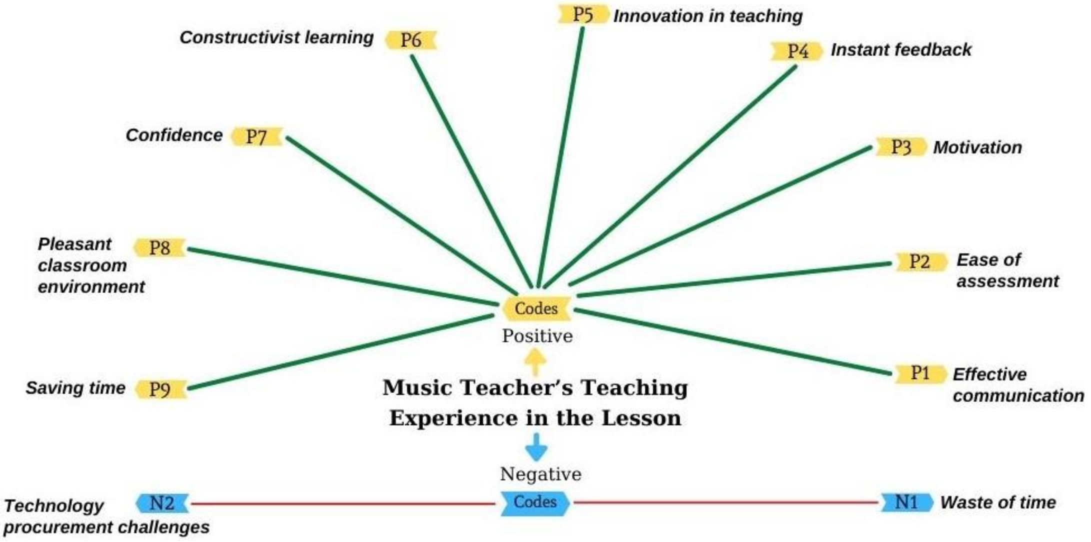
Figure 7. *Relationship map of students' station technique learning experiences.*


P5 Attractiveness
Students: *30* 5-26-16-3-10 24-2-28-9-12-19-25-22-7-11 14-21. Total: (18) *30* (66 word)
N=30 %60
P6 Effective communication
Students:: *11* 18-4-26-29-8-30-1 16-20-27-3. Total: (12) *11* (41 word)
N=30 %36,6
P4 Learning with fun
Students: *5* 26-3-16-30-13-1-8-20-5 12-29-6-10-18-14-25-4-9-24-17. Total: (21) *5* (55 word)
N=30 %70
P7 Productivity
Students: *26* 7-29-19-3-15-30-17 9-16. Total: (10) *26* (71 word)
N=30 %33,3
P3 Active participation
Students: *3* 2-10-11-30-26-12-16 24-8-27-1-9-20-13-25-29-5-7-26-22 18. Total: (22) *3* (55 word)
N=30 %73,3
P8 Originality
Students: *30* 11-3-26-16-3-29. Total: (7) *30* (66 word)
N=30 %23,3
P2 Motivation
Students: *5* 30-16-11-2-28-9-1 12-25-3-8-20-24-4-29-7-10-15-21 26-13-22. Total: (23) *5* (55 word)
N=30 %76,6
Codes
Positive
N=30 %16,6
N=30 %80
P9 Socialisation
Students: *26* 30-3-21-11. Total: (5) *26* (71 word)
Students' Station Technique Learning Experiences
P1 Cooperation
Students: *16* 5-30-10-1-15-26-9 12-21-18-7-2-19-8-25-23-3-11-28-4 22-6-14. Total: (24) *16* (138 word)
Negative
Codes
N2 Insufficient lesson time
N=30 %6,6
N=30 %13,3
N1 Noise
Students: *29* 30. Total: (2) *29* (39 word)
Students:: *17* 4-19-28. Total: (4) *17* (68 word)

3.7. Findings About the Seventh Research Question

In this section, the findings related to the seventh research question "What positive or negative teaching experiences did the music teacher gain as a result of music lesson activities?" are presented. In Figure 8, where the music teachers' opinions are presented, the visual options "code size reflects code frequency" and "line width reflects code frequency" were used, and the code frequency was not included. Figure 8 shows that the music teacher mostly went for the positive codes. A remarkable finding is that the teacher expressed both positive and negative views about the time. In addition, it is seen that there are related codes between teachers' teaching experiences and students' learning experiences. Both teachers and students, for instance, underlined that the work done increased their motivation under many themes.

Figure 8. *The relationship map of the music teacher's teaching experience in the lesson.*


P5 Innovation in teaching
Constructivist learning P6
P4 Instant feedback
Confidence P7
P3 Motivation
Pleasant classroom environment P8
P2 Ease of assessment
Codes
Positive
Saving time P9
P1 Effective communication
Music Teacher's Teaching Experience in the Lesson
Negative
Technology procurement challenges N2
Codes
N1 Waste of time

The opinions of the music teacher are as follows:

"I believe the Plickers platform is a very useful application. I think that this application raises the motivation of both the students and me to the next level. A more advantageous application than the Kahoot program. Because students do not need to use any digital device. These applications allowed me to communicate more effectively with students. It also gave the course process a new identity. Kahoot and Plickers applications made it easy for me to evaluate students' course performance. I was able to evaluate students' musical skills in a practical way. I never had to read the exam paper again. Changing the atmosphere of the lesson and making it fun also increased my self-confidence. In this process, being able to provide instant feedback to students helped me save time. I would like to draw attention to a significant disadvantage of such applications. Unfortunately, such activities before the lesson can take a few hours. This disadvantage is valid for all works with technology content. But it didn't take me long to get used to this kind of technology-based work. I gained practicality in a short time. I've had a little problem with Plickers apps. The sunlight reflected on the QR cards of the students sitting by the window. This prevented some cards from being scanned. After closing the curtains of the classroom, the problem disappeared (music teacher)."

4. DISCUSSION and CONCLUSION

This research revealed the effects of learning environments formed by the combination of blended learning, the ASSURE instructional design model, the learning at station method, mobile games and Web 2.0 tools in teaching basic music theory subjects, to both teachers and students. The action research procedures were carried out in both face-to-face and online environments. Related to the first research question, the current research suggests that blended learning offers significant opportunities for independent learning, as in the findings of Ruokonen and Ruismäki (2016). Emphasizing this, both the teacher and the students went for the "effective communica-tion" code. Adileh's (2012) study also suggests that the music lesson success and attitude scores of the students taught by blended learning methods are higher than those taught with face-to-face instruction. As opposed to these positive results, the negative code "anxiety" was identified in the findings of both Cheng and Lam's (2021) study and the current study. However, the findings of both studies showed that students experienced short-term anxiety in terms of blended learning activities. As a result, it can be said that blended learning not only provides unique learning experiences for the students but also offers significant benefits for distant instructional activities (Crawford, 2013).

The second research question showed that the students developed a new term “easy to use” for the Thinglink app unlike their other experiences because the Thinglink platform’s greatest strength includes simplicity and flexibility (Jeffery et al., 2022). However, students experienced a digital access problem while trying to follow the lesson activities on the Thinglink platform despite these positive results. According to Jeffery et al. (2022), it is possible to encounter a potential challenge such as digital access in the successful implementation of Thinglink-based learning resources.

As stated in the research conducted by Başal and Eryılmaz (2021) and Revenko (2021) during the COVID-19 pandemic, students went for the code “motivation” under both Kahoot and Plickers learning experiences related to the findings of the third and fourth research question. In the Plickers application, students used only answer cards designed with QR codes instead of mobile devices required in Socrative and Kahoot platforms (Pastor & López, 2018). This caused the students to go for the “easy to use” code in the first place. According to Pastor and López (2018), Plickers is a Web 2.0 tool that can be easily applied by someone who has never used technology, can be used at different education levels, is suitable for different subjects, and attracts both students’ and teachers’ attention.

When it comes to the fifth research question, it can be said that Rhythm Cat and NoteWorks mobile games facilitated students’ music learning (Della Ventura, 2017) and enabled them to learn while having fun (Baratè & Ludovico, 2013; Zhou et al., 2010). The current research findings are also similar to the findings of Paule-Ruiz et al. (2017) showing that mobile games increase the participants’ motivation. As a result of mobile learning-based activities, students participated in the lesson more. Sung et al. (2016) suggested that mobile devices reveal a potential power to strengthen students’ participation and motivation in the lesson. A few students experienced control problems based on features of the mobile device. However, Eren’s (2015) study proves that most secondary and high school students do not have any difficulties using tablets and they even use the touch screen easily. Moreover, another study shows that such experiences can be more effective by considering the weight or touchscreen sensitivity of mobile devices such as iPhones and tablets (Furió et al., 2013). Researchers should have considered these different experienced results so that they could take more preventive and early measures against such short-term problems.

Findings related to the sixth research question showed that the participants mostly went for positive codes among all learning experiences with the learning at stations method. The first code the students chose was “collaboration.” This finding proved that learning at station methods has the potential to improve collaborative learning, similar to Chien’s (2017) findings. The learning at stations method also created a fun classroom environment that promoted active participation, as stated in (Genç, 2013). Moreover, it was revealed that it was interesting for students, encouraged active learning (Li et al., 2021) and allowed them to socialize with each other (Davis et al., 2021). However, as Avcı (2015) stated, some students stated that noise that occurred during the application adversely affected them. It did not go unnoticed that the music teacher did not say anything about the noise, a common problem in the class.

In relation to the findings of the seventh research question, the music teacher drew attention to the fact that a pleasant learning environment was created in the classroom and the teaching became more innovative with the help of Web 2.0 tools, as stated in Coutinho and Mota (2011). The music teacher also discussed the concept of “time” both in terms of positive and negative aspects. The teacher drew attention to the time required to prepare the questions and the necessary exam environment in the Plickers application since it took approximately a few hours to prepare a question activity, which supports the findings of Chng and Gurvitch (2018). Teaching Plickers and similar applications to students can be long and boring in time (García, 2022). Mu-sic teachers pointed out that technology-supported lessons to be prepared during the pandemic should be organized more effortlessly (Adam & Metljak, 2022). For example, more time is needed while preparing or updating lesson activities on the Thinglink platform (Jeffery et al., 2022). While discussing the time issue, it should be noted that the Plickers and Thinglink plat-forms save time as a result of providing instant feedback to students (Chng & Gurvitch, 2018; García, 2022; Jeffery et al., 2022). As a result of this feedback, Plickers and Kahoot also facilitate the assessment of students' musical skills (García, 2022). Kahoot also has the potential to increase the interaction between teachers and students and the interaction among students (Wang & Tahir, 2020). Consistent with this finding, both the students and the music teacher drew attention to Kahoot's effective communication feature.

Both teachers and students evaluated modern teaching methods, mobile games, learning at stations method and Web 2.0 tools using mostly positive concepts. According to Sarıkaya (2021), integration between music teaching technologies and student-centered methods offer more meaningful learning activities for learners. At this point, Kibici (2022) concluded that the technological literacy of music teachers is high, while technology integration in the lesson process and general technological competencies are at a medium level. When students' opinions and teacher diaries were correlated, it was seen that there was consistency between the findings. Both teachers and students pointed out that the music lesson activities implemented during the action research procedure mostly increased their motivation and self-confidence. Another out-standing result is that the music teacher carried out the practices with fun like the students. In the current study, some of the findings that emerged in students' Kahoot learning experiences are remarkable. The positive codes of "fun", "intense participation", "interaction", "attractiveness" and "easy use" are in line with the findings of Gil et al.'s (2022) music education study, while the codes of "motivation" and "competitive environment" are in line with the findings of Škoro and Kir (2021) and Saraçoğlu and Kocabatmaz (2019). In addition to these results, it was observed that the emerged negative findings had a short-term effect.

The research findings should be evaluated in the light of various limitations. First, the NoteWorks, Rhythm Cat, Thinglink, Plickers, and Kahoot used within the technology combination in the procedure were a deliberate choice, providing specific data on the effectiveness of this combination. Second, the participants in the study represented a specific sample and consisted of a single music teacher. Therefore, the findings only represented a small area of participants. Future research can explore the effects of different modern teaching methods and different technology-based teaching tools on wider samples. Different instructional technology tools and modern pedagogies can be used in future applied studies. The findings of the current study and past studies have shown that in both modern pedagogies and technology-supported studies, students encounter self-regulation skill deficiencies in organising and managing the process while continuing their course work independently from the instructor (Chuang et al., 2018; Çakıroğlu & Öztürk, 2017; Lightner, 2016). It is recommended that students' self-regulation deficiencies should be taken into consideration in such new studies. In addition, practical and easy-to-use Web 2.0 tools should be included when planning such studies. This is very important in terms of saving teachers' time. Finally, in future studies, music teaching scenarios with augmented and virtual reality content, which is an important power of new technologies, can be included.

Acknowledgments

The authors would like to thank the blind reviewers for their useful comments and insightful suggestions. This study was presented as an abstract at the International EDUcongress (17-19 November 2022 / Akdeniz University).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Atatürk University/Institution, 19.03.202l-03.

Authorship Contribution Statement

Sevim Irmis Engizli: Visualization, Resources, Authors may edit this part based on their case. **Ali Korkut Uludag:** Investigation, Methodology, Software, Formal Analysis, and Writing-original draft, Supervi-sion, and Validation.

Orcid

Sevim Irmis Engizli https://orcid.org/0000-0001-6070-9642
Ali Korkut Uludag https://orcid.org/0000-0002-6164-5211

REFERENCES

Adam, T.B., & Metljak, M. (2022). Experiences in distance education and practical use of ICT during the COVID-19 epidemic of Slovenian primary school music teachers with different professional experiences. *Social Sciences & Humanities Open*, *5*(1), 100246. https://doi.org/10.1016/j.ssaho.2021.100246

Adileh, M. (2012). Teaching music as a University Elective Course through e-learning. *Australian Journal of Music Education*, 1, 71-79.

Aguirre, A., Salazar, C., Lema, A., & Martin, C. (2019). Use of affordable hardware and free web based tools for Control Systems laboratory experiments. *IFAC-PapersOnLine*, *52*(9), 74-78. https://doi.org/10.1016/j.ifacol.2019.08.127

Alacapınar, G.F.G. (2009). İstasyon tekniği ile ders işlemeye yönelik öğrenci görüşleri [Students' views on studying lessons with station Technique]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *9*(1), 137-147.

Alammary, A., Sheard, J., & Carbone, A. (2014). Blended learning in higher education: Three different design approaches. *Australasian Journal of Educational Technology*, *30*(4), 440-454.

AlNajdi, S.M. (2018). Design a blended learning environment to teach Arabic Alphabet for non-Arabic speaker children based on ASSURE model. *International Journal of Information and Education Technology*, *8*(2), 128-132.

Altın, M. (2021). Evaluation of the effectiveness of English language instruction based on the ASSURE model. *E-International Journal of Educational Research*, *12*(5). https://doi.org/10.19160/e-ijer.1018149

Al Fatihah, I., Ramli, M., & Rahardjo, D.T. (2022). The effect of STEM-ThingLink Learning Design on students' conceptual understanding of nutrition. *Biosfer: Jurnal Tadris Biologi*, *13*(1), 1-11. https://doi.org/10.24042/biosfer.v13i1.11920

Asatillayevna, U.G. (2022). The role of internet resources in language skills development. *Central Asian Journal of Mathematical Theory and Computer Sciences*, *3*(4), 30-32. https://doi.org/10.17605/OSF.IO/T39YH

Attard, C., & Holmes, K. (2020). "It gives you that sense of hope": An exploration of technology use to mediate student engagement with mathematics. *Heliyon*, *6*(1), e02945. https://doi.org/10.1016/j.heliyon.2019.e02945

Anohina, A. (2005). Analysis of the terminology used in the field of virtual learning. *Journal of Educational Technology & Society*, *8*(3), 91-102.

Avcı, H. (2015). İngilizce öğretiminde istasyon tekniğinin kullanımının akademik başarıya, tutumlara ve kalıcılığa etkisi [The effect of station technique usage on academic achievement, attitudes, and retention in English language teaching] (Master's thesis, Fırat Üniversitesi-Elazığ). Yükseköğretim Kurulu Ulusal Tez Merkezi.

Ayaz, A.H. (2019). Yabancı Dil Olarak Türkçe Öğretiminde Formatif Bir Test Olarak "Kahoot!" Uygulaması [The application of "Kahoot!" as a formative test in teaching Turkish as a foreign language]. *Hacettepe Üniversitesi Yabancı Dil Olarak Türkçe Araştırmaları Dergisi*, (5), 7-27.

Bajracharya, JR. (2019). Instructional design and models: ASSURE and Kemp. *Journal of Education and Research*, *9*(2), 1-9.

Baratè, A., & Ludovico, L. A. (2013, May). Serious games for music education. A mobile application to learn clef placement on the stave. In *International Conference on Computer Supported Education (CSEDU)* (pp. 234-237). https://pdfs.semanticscholar.org/3cf2/92d4e87aa2b7761ead506af3222e6497d09f.pdf

Başal, A., & Eryılmaz, A. (2021). Engagement and affection of pre-service teachers in online learning in the context of COVID-19: Engagement-based instruction with web 2.0 technologies vs direct transmission instruction. Journal of Education for Teaching, *47*(1), 131-133. https://doi.org/10.1080/02607476.2020.1841555

Batir, Z., & Sadi, Ö. (2021). A science module designed based on the ASSURE Model: Potential energy. *Journal of Inquiry Based Activities*, *11*(2), 111-124.

Batista, J.A.F.A., Souza, M.M.P., Barros, T.D., Gupta, N., & Reis, M.J.C.S. (2022). Using the ThingLink computer tool to create a meaningful environmental learning scenario. *EAI Endorsed Transactions on Smart Cities*, *6*(17), 1-8. http://dx.doi.org/10.4108/eai.21-2-2022.173457

Beirnes, S., & Randles, C. (2022). A music teacher's blended teaching and learning experience during COVID-19: Autoethnography of resilience. *International Journal of Music Education*, 02557614221091829. https://doi.org/10.1177/02557614221091829

Biasutti, M., Antonini Philippe, R., & Schiavio, A. (2022). E-learning during the COVID-19 lockdown: An interview study with primary school music teachers in Italy. *International Journal of Music Education*, 02557614221107190. https://doi.org/10.1177/02557614221107190

Bresler, L. (1995). Ethnography, phenomenology and action research in music education. *The Quarterly Journal of Music Teaching and Learning* *6*(3), 16-18.

Cain, T. (2008). The characteristics of action research in music education. *British Journal of Music Education, 25*(3), 283-313. https://doi.org/10.1017/S02650517

Calderón-Garrido, D., Gustems-Carnicer, J., & Faure-Carvallo, A. (2021). Adaptations in Conservatories and Music Schools in Spain during the COVID-19 Pandemic. *International Journal of Instruction, 14*(4), 451-462.

Cautreels, P. (2003). A personal reflection on scenario writing as a powerful tool to become a more professional teacher educator. *European Journal of Teacher Education*, *26*(1), 175-180.

Cheng, L., & Lam, C.Y. (2021). The worst is yet to come: The psychological impact of COVID-19 on Hong Kong music teachers. *Music Education Research*, *23*(2), 211-224. https://doi.org/10.1080/14613808.2021.1906215

Chien, C.W. (2017). Undergraduates' Implementations of Learning Stations as their Service Learning Among Elementary School Students. *International Journal of Primary, Elementary and Early Years Education*, *45*(2), 209-226.

Chng, L., & Gurvitch, R. (2018). Using Plickers as an assessment tool in health and physical education settings. *Journal of Physical Education, Recreation & Dance*, *89*(2), 19-25. https://doi.org/10.1080/07303084.2017.1404510

Chuang, H.H., Weng, C.Y., & Chen, C.H. (2018). Which students benefit most from a flipped classroom approach to language learning? *British Journal of Educational Technology*, *49*(1), 56-68. https://doi.org/10.1111/bjet.12530

Cooper, S., Dale, C., & Spencer, S. (2009). A tutor in your back pocket: Reflections on the use of iPods and podcasting in an undergraduate popular music programme. *British Journal of Music Education*, *26*(1), 85-97. https://doi.org/10.1017/S0265051708008280

Coutinho, C., & Mota, P. (2011). Web 2.0 technologies in music education in Portugal: Using podcasts for learning. *Computers in the Schools*, *28*(1), 56-74. https://doi.org/10.1080/07380569.2011.552043

Crawford, R. (2013). Evolving technologies require educational policy change: Music education for the 21st century. *Australasian Journal of Educational Technology*, *29*(5). https://doi.org/10.14742/ajet.268

Crawford, R. (2017). Rethinking teaching and learning pedagogy for education in the twenty-first century: Blended learning in music education. *Music Education Research*, *19*(2), 195-213. https://doi.org/10.1080/14613808.2016.1202223

Creswell, J.W. (2019). *Nitel araştırmacılar için 30 temel beceri* [30 essential skills for qualitative researchers] (2th edition). Anı Publication.

Crompton, H., & Burke, D. (2020). Mobile learning and pedagogical opportunities: A configurative systematic review of PreK-12 research using the SAMR framework. *Computers & Education*, *156*, 1-15. https://doi.org/10.1016/j.compedu.2020.103945

Çakıroğlu, Ü., & Öztürk, M. (2017). Flipped classroom with problem based activities: Exploring self-regulated learning in a programming language course. *Journal of Educational Technology & Society*, *20*(1), 337-349. https://www.jstor.org/stable/10.2307/jeductechsoci.20.1.337

Çetinkaya, M. (2017). Fen eğitiminde modelleme temelinde düzenlenen kişiselleştirilmiş harmanlanmış öğrenme ortamlarının başarıya etkisi [Effect of Personalized Blended Learning Environments Arranged on the Basis of Modeling to Achievement in Science Education]. *Ordu Üniversitesi Sosyal Bilimler Enstitüsü Sosyal Bilimler Araştırmaları Dergisi*, *7*(2), 287-296.

Daubney, A., & Fautley, M. (2020). Editorial Research: Music Education in a Time of Pandemic. *British Journal of Music Education*, *37*(2), 107-114. https://doi.org/10.1017/S0265051720000133

Della Ventura, M. (2017). Creating inspiring learning environments by means of digital technologies: A case study of the effectiveness of WhatsApp in music education. In *E-Learning, E-Education, and Online Training*, *4*(14), 36-45. https://doi.org/10.4108/eai.26-7-2017.152906

Dziuban, C.D., Patsy, M., & Joel, L.H. (2004). Higher education, blended learning, and the generations: Knowledge is power-No more. *LIB*, (118), 1-17.

Davis, E., Flavin, A., Harris, M.M., Huffman, L., Watson, D., & Weller, K.M. (2021). Addressing Student Isolation During the Pandemic: An Inquiry into Renewing Relationships and Reimagining Classroom Communities on Remote Instruction Platforms. *Journal of Practitioner Research*, *6*(1), 1-9. https://doi.org/10.5038/2379-9951.6.1.1199

Edward, C.N., Asirvatham, D., & Johar, M.G.M. (2018). Effect of blended learning and learners' characteristics on students' competence: An empirical evidence in learning oriental music. *Education and Information Technologies*, *23*, 2587-2606.

Edwards-Smith, A. (2022). Learning through exploration and experience using ThingLink. *100 Ideas for Active Learning*. https://doi.org/10.20919/OPXR1032/102

Eilks, I. (2002). " Learning at Stations" in Secondary Level Chemistry Lessons. *Science Education International*, *13*(1), 11-18.

Eren, E. (2015). Perceptions and opinions of middle and high school students about tablet computers in education. *Journal of Kirsehir Education Faculty*, *16*(1), 409-428.

Ernst, H., Harrison, J., & Griffin, D. (2013). Anywhere, anytime, with any device: scenario-based mobile learning in biomedical sciences. *International Journal of Mobile Learning and Organisation 11*, *7*(2), 99-112.

Eroğlu, Ö. (2022). Integrating movable numbers into fixed-do system in solfege class: An action research study. *Music Education Research*, *24*(1), 70-82. https://doi.org/10.1080/14613808.2021.2015311

Faryadi, Q. (2007). Instructional Design Models: What a revolution! *Online Submission*.

Fazal, M., & Bryant, M. (2019). Blended learning in middle school math: The question of effectiveness. *Journal of Online Learning Research*, *5*(1), 49-64.-https://www.learntechlib.org/primary/p/183899/.

Fehrle, C.C., & Schulz, J. (1977). *guidelines for learning stations*. Missouri University.

Furió, D., GonzáLez-Gancedo, S., Juan, M.C., Seguí, I., & Costa, M. (2013). The effects of the size and weight of a mobile device on an educational game. *Computers & Education*, *64*, 24-41. https://doi.org/10.1016/j.compedu.2012.12.015

García, N.J.L. (2016). Evaluation and ITC in primary education: using plickers to evaluate musical skills. *Revista de la Facultad de Educación de Albacete*, *31*(2), *81-90. https://revista.uclm.es/index.php/ensayos/article/view/1131/pdf_1*

García, N.J.L. (2022). Kahoot! Plickers y Socrative: recursos TIC para evaluar contenidos educativo-musicales en educación primaria. *Apertura*, *14*(1), 6-25. http://doi.org/10.32870/Ap.v14n1.2134

Gay, G., Stefanone, M., Grace-Martin, M., & Hembrooke, H. (2001). The effects of wireless computing in collaborative learning environments. *International Journal of Human-Computer Interaction*, *13*(2), 257-276. https://doi.org/10.1207/S15327590IJHC1302_10

Genç, M. (2013). Prospective Teachers' Views About Using Station Technique at Environmental Education Course. *Erzincan University of Education Faculty*, *15*(2), 188-203.

Gibson, S.J. (2021). Shifting from offline to online collaborative music-making, teaching and learning: perceptions of Ethno artistic mentors. *Music Education Research*, *23*(2), 151-166. https://doi.org/10.1080/14613808.2021.1904865

Gil, D.G., Vallés, C.B., Villarroel, C.A., & Otero, I.R. (2022). Kahoot! in Music and Physical Education Disciplines in Higher Education. *Aloma: Revista de Psicologia, Ciències de l'Educació i de l'Esport*, *40*(1), 45-54. https://doi.org/10.51698/aloma.2022.40.1.45-54

Göksün, D.O., & Gürsoy, G. (2019). Comparing success and engagement in gamified learning experiences via Kahoot and Quizizz. *Computers & Education*, *135*, 15-29. https://doi.org/10.1016/j.compedu.2019.02.015

Gündüzalp, C., & Yıldız, E.P. (2020). The effect of a course designed with ASSURE model on student's attitudes towards the use of information communication technologies and computer anxiety levels. *Ekev Akademi Dergisi*, *24*(83), 107-136.

Heinich, R., Molenda, M., Russell, J.D. & Smaldino, S.E. (2002). *Instructional media and technologies for learning* (7th ed.). Merrill Prentice Hall. Inc.

Hietanen, L., & Ruismäki, H. (2017). The use of a blended learning environment by primary school student teachers to study music theory. *The European Journal of Social & Behavioural Sciences*, *19*(2), 2393-2404.

Hiğde, E., & Aktamış, H. (2021). The investigation of the effectiveness of the problem-based blended learning environment and students' attitudes. *Manisa Celal Bayar University Journal of the Faculty of Education*, *9*(1), 81-103. https://doi.org/10.52826/mcbuefd.884752.

Ibrahim, A.A. (2015). Comparative analysis between system approach, Kemp, and ASSURE instructional design models. *International Journal of Education and Research*, *3*(12), 261-270.

Jeffery, A.J., Rogers, S.L., Pringle, J.K., Zholobenko, V.L., Jeffery, K.L., Wisniewski, K.D., ... & Emley, D.W. (2022). Thinglink and the laboratory: interactive simulations of analytical instrumentation for HE science curricula. *Journal of Chemical Education*. 99(6), 2277–2290. https://doi.org/10.1021/acs.jchemed.1c01067

Joseph, D., & Lennox, L. (2021). Twists turns and thrills during COVID-19: Music teaching and practice in Australia. *Music Education Research*, *23*(2), 241-255. https://doi.org/10.1080/14613808.2021.1906852

Karaduman, B., Memnun, D.S., & Çakır, C. (2019). ASSURE öğretim tasarımı modeli ile olasılık kavramının öğretimine yönelik bir öneri [A Recommendation For The Teaching Of Possibility Concept With The Assure Teaching Design Model]. 2nd International Congress on New Horizons in Education and Social Sciences (ICES-2019), Proceedings Book, 456-468. https://pdfs.semanticscholar.org/e25c/186e91f9dd5ea47384dbe4f995817691e579.pdf

Kibici, V.B. (2022). An investigation into music teachers' perceptions of technological competencies. *International Journal of Technology in Education and Science*, *6*(1), 111-123. https://doi.org/10.46328/ijtes.344

Kim, E. (2013). Music technology-mediated teaching and learning approach for music education: A case study from an elementary school in South Korea. *International Journal of Music Education*, *31*(4), 413-427. https://doi.org/10.1177/0255761413493369

Kim, D., & Downey, S. (2016). Examining the use of the ASSURE model by K–12 teachers. *Computers in the Schools*, *33*(3), 153-168. https://doi.org/10.1080/07380569.2016.1203208

Korkmaz, Ö., & Tetik, A. (2018). Örgün ve uzaktan eğitim öğrencilerinin derslerde kahoot ile oyunlaştırmaya dönük görüşleri [Formal and Distance Education Students' Views on Gamefication with Kahoot in Lessons]. *Journal of Instructional Technologies and Teacher Education*, *7*(2), 46-55.

Krause, J.M., O'Neil, K., & Dauenhauer, B. (2017). Plickers: A formative assessment tool for K-12 and PETE professionals. *Strategies*, *30*(3), 30-36. https://doi.org/10.1080/08924562.2017.1297751

Kristianti, Y., Prabawanto, S., & Suhendra, S. (2017, May). Critical thinking skills of students through mathematics learning with ASSURE model assisted by software autograph. In *Journal of Physics: Conference Series* (pp. 012063). https://doi.org/10.1088/1742-6596/895/1/012063

Li, L., Xu, L.D., He, Y., He, W., Pribesh, S., Watson, S.M., & Major, D.A. (2021). Facilitating online learning via zoom breakout room technology: a case of pair programming involving students with learning disabilities. *Communications of the Association for Information Systems*, *48*(1), 12. https://doi.org/10.17705/1CAIS.04812

Lightner, C.A., & Lightner-Laws, C.A. (2016). A blended model: Simultaneously teaching a quantitative course traditionally, online, and remotely. *Interactive Learning Environments*, *24*(1), 224-238. https://doi.org/10.1080/10494820.2013.841262

Livari, N., Sharma, S., & Ventä-Olkkonen, L. (2020). Digital transformation of everyday life–How COVID-19 pandemic transformed the basic education of the young generation and why information management research should care? *International Journal of Information Management*, *55*, 102183. https://doi.org/10.1016/j.ijinfomgt.2020.102183

Mishra, L., Gupta, T., & Shree, A. (2020). Online teaching-learning in higher education during lockdown period of COVID-19 pandemic. *International Journal of Educational Research Open*, *1*, 100012. https://doi.org/10.1016/j.ijedro.2020.100012

Ogata, H., & Yano, Y. (2004, March). Context-aware support for computer-supported ubiquitous learning. In *The 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education, 2004. Proceedings.* (pp. 27-34). IEEE.

OuYang, Y., Yin, M.C., & Wang, P. (2010). Cognitive load and learning effects of mobile learning for the students with different learning styles. *International Journal of Mobile Learning and Organisation*, *4*(3), 281-293.

Özerbaş, A., Şahin, Ç., Köse, E., Özkan, H.H., Bahar, H.H., Özbek, R., Yeşil, R., Genç, S.Z. (2010). Bilimsel araştırma yöntemleri [Scientific research methods]. In R.W. Kıncal (Ed.), Veri toplama teknikleri içinde [In data collection techniques] (1st ed., pp. 121-179). Nobel Yayın Dağıtım.

Paule-Ruiz, M., Álvarez-García, V., Pérez-Pérez, J.R., Álvarez-Sierra, M., & Trespalacios-Menéndez, F. (2017). Music learning in preschool with mobile devices. *Behaviour & Information Technology*, *36*(1), 95-111.

Park, S.I., & Kihl, T.S. (2012). Rhythm game design for effective music education. *Journal of Korea Game Society*, *12*(1), 33-42.

Pastor, R.M.S., & López, O.C. (2018, May). PLICKERS: una gran alternativa en el Flipped Learning. *In Actas de las Jornadas Virtuales de Colaboración y Formación Virtual USATIC 2018* (pp. 104-122). Zaragoza.

Pozo, J.I., Echeverría, M.P.P., Casas-Mas, A., López-Íñiguez, G., Cabellos, B., Méndez, E., ... & Bano, L. (2022). Teaching and learning musical instruments through ICT: the impact of the COVID-19 pandemic lockdown. *Heliyon*, *8*(1), e08761. https://doi.org/10.1016/j.heliyon.2022.e08761

Prasad, P.W.C., Maag, A., Redestowicz, M., & Hoe, L.S. (2018). Unfamiliar technology: Reaction of international students to blended learning. *Computers & Education*, *122*, 92-103. https://doi.org/10.1016/j.compedu.2018.03.016

Rasheed, R.A., Kamsin, A., & Abdullah, N.A. (2020). Challenges in the online component of blended learning: A systematic review. *Computers & Education*, *144(1)*, 103701. https://doi.org/10.1016/j.compedu.2019.103701

Ronimus, M., Eklund, K., Westerholm, J., Ketonen, R., & Lyytinen, H. (2020). A mobile game as a support tool for children with severe difficulties in reading and spelling. *Journal of Computer Assisted Learning*, *36*(6), 1011-1025. https://doi.org/10.1111/jcal.12456

Rosen, D., Schmidt, E.M., & Kim, Y.E. (2013, June). Utilizing music technology as a model for creativity development in K-12 education. In *Proceedings of the 9th ACM Conference on Creativity & Cognition* (pp. 341-344). https://doi.org/10.1145/2466627.2466670

Ruokonen, I., & Ruismäki, H. (2016). E-learning in music: A case study of learning group composing in a blended learning environment. *Procedia-Social and Behavioral Sciences*, *217*, 109-115. https://doi.org/10.1016/j.sbspro.2016.02.039

Salvador, K., Knapp, E.J., & Mayo, W. (2021). Reflecting on the 'Community'in Community Music School after a transition to all-online instruction. *Music Education Research*, *23*(2), 194-210. https://doi.org/10.1080/14613808.2021.1905623

Saraçoğlu, G.K., & Kocabatmaz, H. (2019). A study on Kahoot and Socrative in line with preservice teachers' views. *Educational Policy Analysis and Strategic Research*, *14*(4). https://doi.org/10.29329/epasr.2019.220.2

Sarıkaya, M. (2022). An investigation of music teachers' perceived self-efficacy for technology integration. *International Journal of Technology in Education and Science (IJTES)*, *6*(2). https://doi.org/10.46328/ijtes.369

Sastre, J., Cerdà, J., García, W., Hernández, C.A., Lloret, N., Murillo, A., ... & Dannenberg, R. B. (2013, September). New technologies for music education. In *2013 Second International Conference on E-Learning and E-Technologies in Education (ICEEE)* (pp. 149-154). IEEE. https://ieeexplore.ieee.org/abstract/document/6644364

Sears, M.E. (2007). Designing and Delivering Learning Center Instruction. *Intervention in School and Clinic*, *42*(3), 137-147.

Škoro, D., & Kir, I. (2021). The Application of Digital Tools in Listening to Music in Music Culture Education. *Školski vjesnik: časopis za pedagogijsku teoriju i praksu*, *70*(2), 415-433. https://doi.org/10.38003/sv.70.2.18

Smaldino, S.E., Lowther, D.L., Mims, C., & Russell, J.D. (2015). Öğretim teknolojileri ve öğrenme araçları, (Çev. Ed. A. Arı). Eğitim Kitabevi.

Sung, Y.T., Chang, K.E., & Liu, T.C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252-275. https://doi.org/10.1016/j.compedu.2015.11.008

Şimşek, Ö., Bars, M., & Zengin, Y. (2017). Matematik öğretiminin ölçme ve değerlendirme sürecinde bilgi ve iletişim teknolojilerinin kullanımı [The Use of Information and Communication Technologies in the Assessment and Evaluation Process in Mathematics Instruction]. *Uluslararası Eğitim Programları ve Öğretim Çalışmaları Dergisi*, *7*(13), 190-207.

Palazón, J., & Giráldez, A. (2018). QR codes for instrumental performance in the music classroom. *International Journal of Music Education*, *36*(3), 447-459. https://doi.org/10.1177/0255761418771992.

Pringle, J.K., Stimpson, I.G., Jeffery, A.J., Wisniewski, K.D., Grossey, T., Hobson, L., ... & Rogers, S.L. (2022). eXtended Reality (XR) virtual practical and educational eGaming to provide effective immersive environments for learning and teaching in forensic science. *Science & Justice*, *62*(6),696-707. https://doi.org/10.1016/j.scijus.2022.04.004

Qiu, H., Li, Q., & Li, C. (2020). How technology facilitates tourism education in COVID-19: Case study of Nankai University. *Journal of Hospitality, Leisure, Sport & Tourism Education*, 100288. https://doi.org/10.1016/j.jhlste.2020.100288

Revenko, V. (2021). Education and Music Culture in the Context of Web 2.0. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(10), 96-107. https://doi.org/10.3991/ijet.v16i10.19693

Roman, M., & Plopeanu, A.P. (2021). The effectiveness of the emergency eLearning during COVID-19 pandemic. The case of higher education in economics in Romania. *International Review of Economics Education*, *37*, 100218. https://doi.org/10.1016/j.iree.2021.100218

Thorgersen, K.A., & Mars, A. (2021). A pandemic as the mother of invention? Collegial online collaboration to cope with the COVID-19 pandemic. *Music Education Research*, *23*(2), 225-240. https://doi.org/10.1080/14613808.2021.1906216

Torun, F., & Dargut, T. (2015). A proposal on the feasibility of the flipped classroom model in mobile learning environments. *Adnan Menderes Üniversitesi Eğitim Fakültesi Eğitim Bilimleri Dergisi*, *6*(2), 20-29.

Türe, Z.G., Yalçın, P., & Yalçın, S.A. (2020). Investigating the use of case-oriented station technique in teaching socio-scientific issues: A mixed method study. *PEGEM Journal of Education and Instruction*, *10*(3), 929-960. https://doi.org/10.14527/pegegog.2020.029

Valverde-Berrocoso, J., & Fernández-Sánchez, M.R. (2020). Instructional design in blended learning: Theoretical foundations and guidelines for practice. *Blended Learning: Convergence between Technology and Pedagogy*, 113-140. https://doi.org/10.1007/978-3-030-45781-5_6

Viera, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: The kappa statistic. *Fam Med*, *37*(5), 360-363.

Wang, A.I. (2015). The wear out effect of a game-based student response system. *Computers & Education*, *82*, 217-227. https://doi.org/10.1016/j.compedu.2014.11.004

Wang, A.I., & Tahir, R. (2020). The effect of using Kahoot! for learning–A literature review. *Computers & Education*, *149*, 103818. https://doi.org/10.1016/j.compedu.2020.103818

Yalçın, S. (2021). Arapça kelime öğretiminde Web 2.0 araçlarının önemi ve işçiliği hazırlama uygulamaları [The importance of Web 2.0 tools and application examples in teaching Arabic vocabulary]. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *1*(1), 517-538.

Yıldırım, A., & Şimşek, H. (2006). Sosyal bilimlerde nitel araştırma yöntemleri [Qualitative research methods in the social sciences]. Seçkin Publication.

Zhou, Y., Percival, G., Wang, X., Wang, Y., & Zhao, S. (2010, October). Mogclass: a collaborative system of mobile devices for classroom music education. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 671-674). https://doi.org/10.1145/1873951.1874048

APPENDIX

Semi-structured interview questions:

Q1: Can you tell us about your learning experiences in music class? Did you encounter any advantages or disadvantages?

Q2: What do you think about music lesson activities prepared with Web 2.0 tools? Did you encounter any advantages or disadvantages?

Q3: What do you think about music lesson activities prepared with mobile games called Rhythm Cat and NoteWorks based on mobile learning? Did you encounter any advantages or disadvantages?

Q4: What do you think about the learning at stations method you used in the music lesson? Did you encounter any advantages or disadvantages?

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 331–344

https://doi.org/10.21449/ijate.1238250

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

The validity and reliability of the Turkish version of revised metacognitive awareness of reading strategies inventory

Rabia Börekci [1,*], Caner Börekci [2]

[1]Ministry of National Education, Balıkesir, Türkiye
[2]Balıkesir University, Research Center for Distance Learning, Balıkesir, Türkiye

ARTICLE HISTORY

Received: Jan. 17, 2023
Revised: May 06, 2023
Accepted: May 26, 2023

Keywords:
Metacognition,
Metacognitive awareness,
Reading comprehension,
Reading strategies,
Reliability and validity of Turkish version.

Abstract: This study aims to adapt the revised Metacognitive Awareness of Reading Strategies Inventory (MARSI- R) into Turkish. MARSI-R is a self-report instrument designed to assess students' metacognitive awareness of reading strategies and perceived strategy use when reading school-related materials. 525 students (65% female, 35% male, M_{age} = 13 years old.) from multiple school types and degrees participated in this study. A stepwise validation procedure was used to translate and produce a Turkish version of the inventory. Evidence of structural and external aspects of validity for the inventory was collected. The 15-item inventory had a three-factor solution (global reading strategies, problem-solving strategies, and support reading strategies), as supported by confirmatory factor analysis. Turkish version scores were positively correlated with students' perceived reading ability, which provides evidence of MARSI-R's external validity. The coefficient of stability was calculated using data from 85 students who took the Turkish version of the MARSI-R twice in a five-week interval. The study's overall results provided evidence of the reliability and validity of the inventory. According to the results presented in this study, the Turkish version of the inventory can be implemented to assess the students' metacognitive awareness of reading strategies and perceived strategy use. The findings show that the adapted inventory can be used to obtain valid and reliable results for Turkish lower and upper secondary school students.

1. INTRODUCTION

Researchers have been highly interested in the metacognitive aspect of reading comprehension for almost three decades The source of this growing interest in metacognition stems from its effects on the reading process because metacognition, which is defined as "knowledge that takes as its object or regulates any component of any cognitive attempt" (Flavell, 1979, cited in Kusiak, 2013, p. 67), encourages readers to think about thinking and monitor reading comprehension because thinking and monitoring the process affect the nature of reading (Alderson & Alderson, 2000). In that way, readers identify what they already know, monitor the progress in the process, define the problematic areas, and repair them (Grabe & Stoller, 2013). Thus, while assessing reading comprehension, metacognition affects reading

*CONTACT: Rabia Borekci ✉ rabia_borekci@hotmail.com Ministry of National Education, Balıkesir, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

comprehension, and this effect has been the subject of several studies in language learning (Aghaie & Zhang, 2012; Köse & Günesli, 2021; Phakiti, 2008; Wenden, 1998).

Reading comprehension is "a process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (RAND Reading Study Group, 2002, p. 11), and based on this perspective, reading comprehension is interactive and involves multiple processes, knowledge, and procedures. Skilled readers are better comprehenders and efficient strategy users (Zhang et al., 2017). Metacognitive reading strategies are implemented by skilled readers and are referred to as strategic competence (Bachman & Palmar, 2010). Thus, metacognitive reading strategies enable readers to comprehend better while monitoring and regulating the cognitive processes (Devine, 1993), and skilled readers implement more than one strategy at a time (Grabe & Stoller, 2013) to take control of their learning by planning, monitoring, and evaluating the process (Brown, 2000). Moreover, metacognitive reading strategies enable readers to identify possible ways to overcome some reading comprehension-related problems (Villanueva Aguilera, 2014). By reflecting upon reading, monitoring the process by asking comprehension questions, and taking charge of the process (Soto et al., 2019), skilled readers can handle these issues by implementing multiple strategies.

Readers' awareness of metacognitive reading strategies improves the reading comprehension process, and this level of consciousness supports learners in paying specific attention to the process. This awareness is also the indicator of knowledge of metacognition, the ability to use that knowledge and strategy (Zhang et al., 2017). But how to assess language learners' awareness or "perceived use of reading strategies" as stated by Moktari et al. (2018) is a fundamental question. And this brings us to the Revised Metacognitive Awareness of Reading Strategies Inventory (MARSI-R). This inventory is widely used due to its ease of use, better interpretation of responses, and speed of inventory screening in comparison to other instruments for measuring metacognitive awareness of reading strategies, such as the SORS (Mokhtari & Sheorey, 2002) or the Metacomprehension Inventory (Soto et al., 2018). Although the inventory was revised in 2018, the widespread use of the old version in the Turkish context indicated a gap and the need for a new adaptation study, which was also the motivation for this study. In other words, the adapted version of MARSI by Öztürk in Turkish (2012) has been widely used in the Turkish context (e.g., Boyraz, & Altinsoy, 2017; Sarıçoban & Behjoo, 2017), and although the revised version has been available since 2018, the old version is still implemented in some recent studies (e.g., Bagcı & Unveren, 2020; Erdoğan & Yurdabakan, 2018; Köse & Günesli, 2021; Tamin & Büyükahska, 2020). In conclusion, before giving detailed information about the methodology of the study, a piece of brief information will be given for a better understanding of the revision and enhancement of the old form into a revised 15-item inventory.

Metacognitive Awareness of Reading Strategies Inventory developed by Mokhtari and Reichard (2002) comprises three latent factors: global reading strategies (13 items), problem-solving strategies (9 items), and support reading strategies (8 items), for a total of 30 items. It is a Likert survey on a five-point scale (1 = "I never or almost never do this," and 5 = "I always or almost always do this"). This inventory was designed to assess the metacognitive awareness of reading strategies of adult or adolescent language learners. Thus, the target group in the main study is adults and adolescents, but only adolescents are included in this study because the term "adolescent" covers the period of life between childhood and adulthood, between the ages of 10 and 19 (Encyclopedia Britannica, 2023), which is also the age range of the participants in this study. In the relevant literature, the development of metacognition can also be seen as an early stage of adolescence (Kolić-Vehovec et al., 2010; Wall, 2008). MARSI has been translated into various languages, and Turkish is one of them. As mentioned above, Öztürk (2012) conducted a study to adapt the inventory to Turkish. For the piloting stage, 29 6th-grade students, and for the main study, 250 5th-grade students participated in the study. Fit indices were examined, and the values were consistent with the standard values. The results of the confirmatory factor analysis of the adaptation study conducted by Öztürk demonstrated that the Turkish version of the MARSI was consistent with the main study of Mokhtari and Reichard (2002).

Researchers and practitioners have implemented MARSI in their studies and provided feedback, which has been used to shape the revised version. The initial aim of revising the inventory is to make it more effective in assessing metacognitive reading strategy awareness and perceived use of strategy. Mentioned as "enhancements", the focus of these changes is on the readability and comprehensibility of the items related to strategy awareness and implementation for even fourth-grade students, allowing them to easily understand and complete the form. One of the changes in wording is illustrated with an example; "I try to get back on track when I lose concentration" (Problem-solving Strategies) evolved to "getting back on track when getting sidetracked or distracted." The format of the inventory and the response type were improved for a better interpretation of the results. Likert-type scale items were revised to assess participants' degree of knowledge of reading strategies. This enhancement was explained in the article as "5", which means "I always or almost do this" in the old version turned into "I know this strategy quite well, and I often use it when I read." Mokhtari, Dimitrov, and Reichard conducted another study to provide validity and reliability for the revised instrument. The revised version of the instrument has fifteen items, five for each of three latent factors same as in the old version: global reading strategies, problem-solving strategies, and support reading strategies. The inventory takes about 15-20 minutes to complete, and there is some background knowledge information that aims to identify each participant's age, level, and type of school, as well as a reader scale that asks participants to label themselves as readers.

This study aims to adapt the Revised Metacognitive Awareness Reading Strategies Inventory into Turkish and provide evidence for valid and reliable results for Turkish lower and upper secondary school students while evaluating their awareness and perceptions of metacognitive reading comprehension strategies. The old version was adapted into Turkish and implemented in several studies, but the revised version of the inventory hasn't been adapted into Turkish yet. The revised version was strengthened based on feedback from on-site implementations, the reflections of the implementers, and the statistical analysis of the data gathered with the old version. Thus, a more recent, readable, and comprehensible version is available. Researchers of this study aim to contribute to the related literature by adapting the inventory into Turkish. In this way, MARSI-R can be administered in multiple classrooms or to multilevel readers in the Turkish context with fewer items, improved wording, and scale instruction, and it is more convenient for multiple screenings.

2. METHOD

2.1. Sample

This study involved the voluntary participation of Turkish lower and upper secondary school students. The study group was selected by convenience sampling and consisted of students from multiple school types and levels. After the ethics committees of Çanakkale Onsekiz Mart University and the Ministry of Education approved the study, the survey was conducted in face-to-face classes. The scales were applied to a total of 537 students, and the data of 12 students were removed from the data set as a result of the detection of outliers. In this way, the sample was reduced to a total of 525 (346 female, 179 male) Turkish students (M_{age} = 13 yr., *SD* = 2.2, range 10 to 19). These students attended public lower and upper secondary schools in Turkey and had different grade levels. The sample included 305 lower and 220 upper secondary school students.

2.2. Instrument

MARSI-R, which is a self-report instrument, is used to measure students' metacognitive awareness of reading strategies while they read school-related materials. (Moktari et al., 2018). The inventory's revised version consists of 15 items categorized as three factors and graded on a five-point scale (1. "I have never heard of this strategy before." 2. "I have heard of this strategy, but I don't know what it means." 3." I have heard of this strategy, and I think I know what it means." 4. "I know this strategy, and I can explain how and when to use it." 5. "I know this strategy quite well, and I often use it when I read."). The original study's three-factor structure had a loading of at least .40 (the factor loadings of the original version are shown in Appendix A). The internal consistency and reliability of MARSI-R were measured by Cronbach's alpha coefficient, which was equivalent to .850. Per the latent factors, the support reading, problem-solving, and global reading strategies' alpha values were .703, .693, and.743, respectively. The fit indices of the 3-factor scale were CFI = .972, TLI = .966, WRMR = 1.188, and RMSEA = .046, with 90% CI [.016 .027], respectively. The three broad categories of strategies measured by the MARSI-R include:(1) Global reading strategies (GRS), also described as generalized or global reading strategies, are meant to prepare readers for the reading process (e.g., setting a purpose for reading, previewing text content, predicting what the text is about, etc.). (2) Problem-solving Strategies (PSS), which are localized, focused problem-solving or repair strategies used when difficulties are encountered in comprehending textual information (e.g., verifying one's comprehension when confronted with conflicting information, rereading for clarity, etc.). (3) Support reading strategies (SRS), which provide the mechanisms or tools necessary to keep readers' responsiveness (e.g., the use of reference materials such as dictionaries and other support systems). When used in the process of deriving meaning from text, these three subcategories of strategies interact and reinforce each other. The inventory also provides a reader scale that has four options to self-report what level of reader they are. These are: I consider myself (1) an excellent reader, (2) a good reader (3) an average reader, and (4) a poor reader. Moktari and colleagues (2018) suggest interpreting the scores on the instrument as (1) high level of awareness (3.5 points or above) (2) medium level of awareness (2.5 -3.4) and (3) low level of awareness (2.4 points or under). The scores of the items for each reading strategy are summed and divided by five for each subscale score, and all the items' scores in the inventory are summed and divided by the number of items for a composite score. They advise administering the MARSI-R instrument two or three times per school year to track student metacognitive awareness of and use of reading strategies in relation to overall reading performance.

2.3. Procedure and Data Analysis

The MARSI-R was translated and adapted into Turkish using a step-by-step validation approach. The items ' semantic equivalence was established through a translation and back translation procedure (Mallinckrodt & Wang, 2004). A group of Turkish lower and upper secondary school students provided feedback on the items' clarity. Experts in reading comprehension and metacognitive awareness who are native speakers of Turkish and proficient in English were consulted to determine the content equivalency of the Turkish version of the inventory. The measurement's criterion equivalence was investigated following the establishment of construct validity. Below is a detailed explanation of the validation and data analysis process.

Step 1: Two EFL instructors in the school of foreign languages at the university independently translated the MARSI-R items into Turkish. The translated items were then given to two other EFL instructors who were fluent in both languages for back translation. These instructors translated the items into Turkish and then back into English independently.

Step 2: The researchers reviewed the back-translated items and contrasted them with the original MARSI-R items. Researchers examined the back-translated items to determine whether they were semantically equivalent to the original items and made sure that the translation procedure had maintained the original items' intended meaning. A draft version of the MARSI-R in Turkish was created based on the suggested elements.

Step 3: Two experts, one in ELT and the other in curriculum and instruction, evaluated the first draft of the Turkish version of MARSI-R for content equivalence and cultural appropriateness. The wording of a few items in the Turkish version's initial draft was changed based on feedback from experts to better convey the concepts that the inventory developers intended. The field experts' revision suggestions served as the basis for the second draft of the Turkish MARSI-R.

Step 4: To test the items' readability and clarity for the targeted users, the MARSI-R's second draft in Turkish was given to five lower secondary and five upper secondary school students. Students who took part in this pilot study were asked if they thought the items' meanings were clear. Additionally, students were encouraged to offer substitutes for any phrases or items that they felt were unclear. Two items underwent minor adjustments in response to student comments. The Turkish version's final form was created based on the student's feedback.

Step 5: To establish the validity and reliability of the inventory scores in the Turkish sample, the final form of the MARSI-R was administered to lower and upper secondary school students in face-to-face classes. Unidimensional Confirmatory Factor Analysis, Standard Confirmatory Factor Analysis, Bifactor Confirmatory Factor Analysis, and Exploratory Structural Equation Modelling (ESEM) were used to determine the inventory's structural validity. MPlus (ver. 8.1) and Jamovi (ver. 2.3.21) open statistical software, were used to conduct the analyses (Gallucci & Jentschke, 2021) on the three-factor, 15-item model (see Figure 1) to inspect the structural validity. Jamovi is built on the R statistical language and the *lavaan* package used for FA (Rosseel, 2019). Factor analysis required several presumptions to be met before the analysis could begin. Univariate normality, univariate outliers, multivariate normality, and multivariate outliers were thus investigated. Based on kurtosis and skewness values, as well as z standard scores, univariate normality and univariate outlier analyses were performed. Multivariate outliers and multivariate normality were determined using Mahalanobis distance and residual calculations. According to MacCallum et al. (1999), sample size can affect the accuracy of parameter estimates, the model's fit to the data, the influence of observable variables on the proportion of variance explained by common factors, and all these aspects in factor analytic models. For this reason, it was intended to reach the highest possible number of participants. The sample size (N = 525) is considered suitable for FA. After checking the assumptions, it was seen that the most appropriate methods for the analysis of the data were Maximum Likelihood (ML) and Robust Diagonally Weighted Least Squares (known as one of the categorical variable modeling alternatives). It is an alternative to the DWLS model for ordinal data, especially when the responses also have a high degree of skewness, kurtosis, or both (Distefano & Morgan, 2014; Muthén, 1993). The fit indices were based on conventional guidelines introduced by Hu and Bentler (1998). CFI and TLI values in the region of 0.95 indicate a good model fit, but values around 0.90 may be acceptable. The RMSEA should be equal to or less than .07 and .05 to reflect acceptable and good model fits respectively. In the ESEM model, cross-loadings are 'targeted' to be as close to zero as possible to reflect the confirmatory approach of ESEM (Morin et al., 2020). Average Variance Extracted (AVE) measures were also used to determine the scale's convergent and discriminant validity for each dimension. The square roots of all AVEs are displayed on the diagonal in Table 4 to establish discriminant validity. The correlation between the latent factors (GRS, PSS, and SRS) was examined for convergent validity. The reader variable, used as an external measure of perceived reading ability, was used to measure the Pearson correlation coefficients between the student's scores on each of the three MARSI-R latent factors (GRS, PSS, and SRS). Internal consistency and coefficient of stability approaches were used to calculate the measurement's reliability. The Cronbach's Alpha, McDonald's Omega, and Composite Reliability scores were used as indicators of internal consistency. The test-retest method was preferred to examine the coefficient of stability. A total of 85 students took part in the test-retest procedure. The retest was done after 5 weeks, 45 of them were lower secondary and 40 were upper secondary school students.

3. RESULTS

Descriptive Statistics: Table 1 presents the means and standard deviations for the overall MARSI-R score by gender, school type, and sample size of 525 students. According to the descriptive statistics, the sample's MARSI-R mean score was 3.59 (*SD* = .70). In addition, the descriptive statistics of the items and the correlation between the items are presented in Appendix B.

Table 1. *Means and standard deviations of MARSI-R scores by gender, school type, and the total sample.*

		GRS		PSS		SRS		Total	
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender									
Female	346	3.56	.84	3.80	.82	3.61	.74	3.66	.69
Male	179	3.44	.88	3.63	.89	3.26	.88	3.44	.77
School Type									
L. Secondary	305	3.46	.91	3.68	.85	3.47	.79	3.54	.74
U. Secondary	220	3.61	.78	3.81	.85	3.52	.83	3.65	.71
Total	525	3.52	.86	3.74	.85	3.50	.81	3.59	.70

Note. *N* = Sample size, *M* = Mean, *SD* = Standard deviation

3.1. Examining the Factorial Validity of the MARSI-R

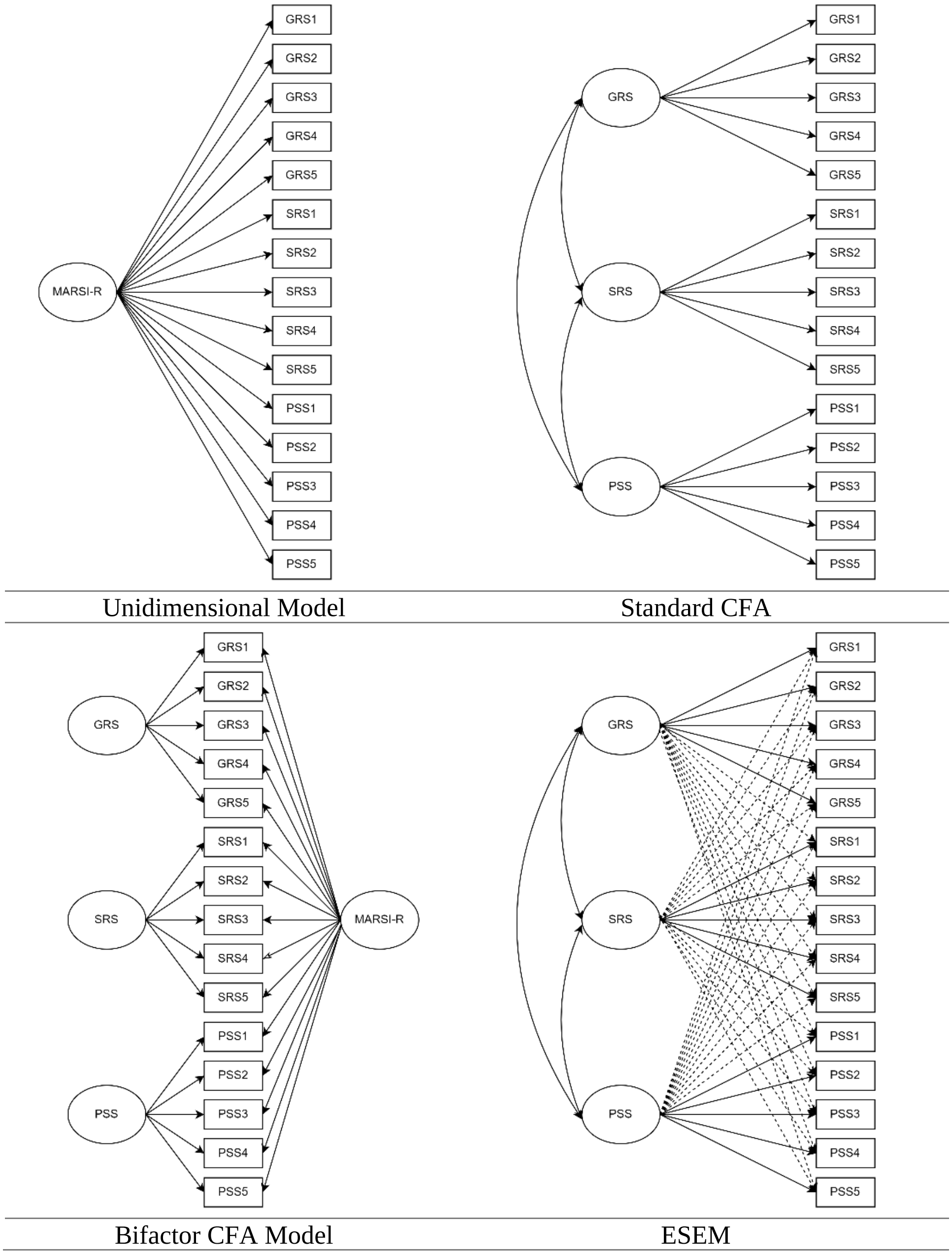
Table 2 shows the calculated normality values for the items for univariate normality. For multivariate normality, Mardia's multivariate skewness and kurtosis coefficients and p values were calculated. Mardia's Test determines whether a group of variables follows a multivariate normal distribution (Von Eye & Bogat, 2004).

Table 2. *Skewness and kurtosis values of items.*

Item	Kurtosis (s.e.)	Skewness (s.e.)	Item	Kurtosis	Skewness
GRS1	-.876 (.213)	-.128 (.106)	SRS1	-.182	-.499
GRS2	-.045	-.954	SRS2	-.637	-.583
GRS3	-.833	-.515	SRS3	-.842	-.409
GRS4	-.939	-.590	SRS4	.249	-1.11
GRS5	-.942	-.429	SRS5	-1.03	-.014
PSS1	-.686	-.565	GRS	-.113	-.537
PSS2	-.372	-.920	PSS	.155	-.686
PSS3	-1.23	-.315	SRS	-.026	-.579
PSS4	.140	-1.02	MARSI-R	.250	-.594
PSS5	-.109	-.932			

From the results, both the skewness ($\gamma 1_p = 2.96$, $p = 0.1772$) and kurtosis ($\gamma 2_p = 18.1$, $p = 0.112$) estimates indicate multivariate normality. Therefore, according to Mardia's MVN test, this data set follows a multivariate normal distribution. The assumption of multicollinearity, variance inflation factor (VIF), and tolerance values were analyzed (Tabachnick & Field, 1996). As a result of the analysis, the VIF and tolerance values were found to be within acceptable ranges (GRS = 1.77 - .56; PSS = 2.10 - .47, SRS = 190 - .52). This indicates that there is no multicollinearity problem as the VIF value is not greater than 10 and the tolerance value is not less than .10. The fit values and factor loadings of the four models were calculated after testing the assumptions. The models are illustrated in Figure 1.

Figure 1. *Visual representation of the four models tested in the present study.*



Unidimensional Model

Standard CFA

Bifactor CFA Model

ESEM

As shown in Table 3, the values obtained for the unidimensional model are acceptable and for the others are good (Hu & Bentler, 1998; Schumacher & Lomax, 2010). Although the analyses showed good model fit, it was important to inspect the factor loadings of all models to see how each solution functioned in estimating model parameters. The factor loadings of bifactor and

ESEM solutions many items are below .40 in bifactor and ESEM analyses (see Appendix A). The factor loadings for items indicate how much of the average respondent's answer to that item is due to his or her general interest in agentic goals, as opposed to something unique to that item. In other words, the factor loadings are an indication of how well the items represent the underlying factor. Therefore, it can be said that using unidimensional and standard CFA models is more appropriate.

Table 3. *Model fit indices for four measurement models of the MARSI-R.*

Model	X^2	*p*	*df*	RMSEA (%95 CI Low-High)	CFI	TLI
Unidimensional CFA	237	<.001	90	.056 (.047 - .064)	.912	.897
Standard CFA	219	<.001	87	.053 (.044 -.062)	.988	.966
Bifactor CFA	114	<.001	69	.035 (.023 - .046)	.973	.956
ESEM	71.2	.153	60	.019 (.010 - .034)	.993	.998

3.2. Discriminant Validity

Average Variance Extracted (AVE) values are shown in Table 4. All AVE values were higher than .40, which provides evidence for the convergent validity of the scale (Fornell & Larcker, 1981). The square roots of all AVE values (diagonally shown) were higher than the correlations shown below them or to their left, supporting the discriminant validity of the scale (Hair et al., 1995).

3.3. Convergent Validity

The factors under MARSI-R were found to be correlated: (1) r = .624 between global reading and problem-solving strategies, (2) r = .570 between global reading and support reading strategies, and (3) r = .658 between problem-solving strategies and support reading strategies (Table 4).

3.4. External Validity

The correlation between the latent factors scores and total scores on the MARSI-R with the scores on the variable reader was calculated as part of the process of gathering evidence relating to the external aspect of validity (Moktari et al., 2018). The Pearson Correlations between the factor scores under the MARSI-R and reader scores of the students are all statistically significant, (1) r = .382 between reader and global reading strategies, (2) r = .346 between reader and problem-solving strategies, (3) r = .320 between reader and support reading strategies, and (4) r = .406 between reader and the total score on the MARSI-R (Table 4).

Table 4. *Correlations, average variance extracted, and composite reliability values for each factor (N = 525, p < .001).*

	AVE	CR	GRS	PSS	SRS	Total Score	Reader Score
GRS	.50	.83	(.71)				
PSS	.50	.83	0.624*	(.71)			
SRS	.49	.71	0.570*	0.658*	(.70)		
Total Score			0.851*	0.883*	0.855*	—	
Reader Score			0.382*	0.346*	0.320*	0.406*	—

Note: * *p*<.001, AVE is average variance extracted; CR is composite reliability

3.5. Reliability and Item Analysis

By latent factors, the alpha values for global reading, problem-solving, and support reading strategies were .763, .693, and .743, respectively. McDonald's ω for the sample was .849. By latent factors, the ω values for global reading, problem-solving, and support reading strategies were .771, .705, and .752, respectively. The Stratified Alpha calculated for the whole scale is .848. These scores were an indication of the consistency of the participants' responses to the inventory items (McNeish, 2018; Taber, 2018). Reliability estimates of the adapted inventory were compatible with the original study. Composite Reliability (CR) scores are shown in Table 4. All CR values (GRS_{CR} =.83, PSS_{CR} = .83, SRS_{CR} = .71) were higher than .70. Values greater than .60 are generally considered acceptable (Bagozzi & Yi, 1988). Inter-item correlations and corrected item-total correlations ranged from .45 to .56 and .48 to .61, respectively. The corrected item-total correlations were above the minimum level of 0.3. The inter-item correlations were also within the acceptable range (greater than 0.3). Test-retest reliability is a measure of the stability of scores on a stable construct from the same person on two or more separate occasions. The coefficient of stability was calculated using data from 85 students who took the Turkish version of the MARSI-R twice in a five-week interval. The Pearson correlation was r = .82, again demonstrating strong reliability.

4. DISCUSSION and CONCLUSION

This study examined the validity and reliability of the Turkish adaptation of MARSI-R using a sample of Turkish lower and upper secondary school students. Factor analysis results supported the structure of the 15-item MARSI-R for Turkish lower and upper secondary school students. The results provide evidence about (1) the structural aspect of validity, with a three-factor structure (GRS, PSS, and SRS), and (2) the external aspect of validity, with correlations between the students' scores on each of the three MARSI-R latent factors (GRS, PSS, SRS) and their scores on the reader scale as an external measure of perceived reading ability. The Turkish version of the inventory's reliability estimations were within the acceptable range, and they were equivalent to the reliability coefficients reported in the original study. Several adaptation studies of MARSI-R have been conducted in various contexts, such as Iranian by Amini et al. (2020), Spanish by Ondé et al. (2022), Vietnamese by Do and Phan (2021), Hungarian by Tary and Molnár (2022), and this study in the Turkish context is one of them. According to the results presented in this study, the Turkish version of the inventory can be implemented to assess the students' metacognitive awareness of reading strategies and perceived strategy use when reading school-related materials. This version is valid and reliable for Turkish lower and upper secondary school students.

Metacognitive reading strategy awareness and perceived reading strategy use are also important in higher education. Also, adult language learners may be the target group of another study. Reading comprehension strategy awareness and perceived strategy use affect the comprehension process. Different proficiency levels in the target language, reading habits, perceptions of themselves as readers, and multiple learners of different ages could be the subject of further studies. Scale development studies can be conducted for these groups, and their development can be monitored. Furthermore, this adapted scale can be tested for measurement invariance between genders or different groups, and it can be investigated whether there is consistency in interpreting the reading strategy statements between these groups. In terms of the convergent validity of the inventory, it can be compared with scales that measure similar characteristics. In addition, the scores obtained from the scale can be compared with students' performance in Turkish, Turkish language and literature, and foreign language courses, where their performance in reading comprehension is crucial.

This study has several potential limitations. First, this study is limited to the Turkish (EFL) context. Second, the grade level of the participants was limited to 5th to 12th grade. This means that the age range of the students was limited to those between the ages of 10 and 19. Last, the size of the sample that was chosen for the study could also be considered a limitation of the study.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Çanakkale Onsekiz Mart University, E-84026528-050.01.04-2200276328.

Authorship Contribution Statement

Authors are expected to present author contributions statement to their manuscript such as; **Rabia Börekci**: Conception, Design, Supervision, Materials, Data Collection and Processing, Literature Review, Writing. **Caner Börekci**: Conception, Design, Materials, Data Collection and Processing, Analysis and Interpretation, Literature Review, Critical Review.

Orcid

Rabia Börekci https://orcid.org/0000-0001-5678-7365
Caner Börekci https://orcid.org/0000-0001-5749-2294

REFERENCES

Aghaie, R., & Zhang, L.J. (2012). Effects of explicit instruction in cognitive and metacognitive reading strategies on Iranian EFL students' reading performance and strategy transfer. *Instructional Science*, *40*(6), 1063-1081. https://doi.org/10.1007/s11251-011-9202-5

Alderson, C.J., & Alderson, J.C. (2000). Assessing reading. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Amini, D., Hosseini Anhari, M., & Ghasemzadeh, A. (2020). Modelling the relationship between metacognitive strategy awareness, self-regulation and reading proficiency of Iranian EFL learners, *Cogent Education*, *7*(1), 1-17. https://doi.org/10.1080/2331186X.2020.1787018

Bachman, L.F. & Palmer, A.S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. OUP.

Bagci, H., & Unveren, D. (2020). Investigation of the relationship between metacognitive awareness of reading strategies and self-efficacy perception in reading comprehension in mother-tongue: Sample of 8th graders. *International Journal of Educational Methodology*, *6*(1), 83-98. https://doi.org/10.12973/ijem.6.1.83

Bagozzi, R.P., & Phillips, L.W. (1982). Representing and testing organizational theories: A holistic construal, *Administrative Science Quarterly*, *27* (September): 459-489. https://doi.org/10.2307/2392322

Brown, H.D. (2000). *Principles of language learning and teaching* (Vol. 4). Longman.

Boyraz, S., & Altinsoy, E. (2017). Metacognitive awareness of reading strategies in EFL context. *International Journal of Language Academy*, *5*(5), 159-167. https://doi.org/10.18033/ijla.3655

Devine, J. (1988). The relationship between general language competence and second language reading proficiency: Implications for teaching. *Interactive approaches to second language reading*, 260-277. https://doi.org/10.1017/CBO9781139524513.024

Distefano, C., & Morgan, G.B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 425- 438. https://doi.org/10.1080/10705511.2014.915373

Do, H.M., & Phan, H.L.T. (2021). Metacognitive awareness of reading strategies on second language Vietnamese undergraduates. *Arab World English Journal*, *12*(1) 90-112. https://dx.doi.org/10.24093/awej/vol12no1.7

Encyclopædia Britannica, inc. (2023, March 28). Adolescence. Encyclopædia Britannica. Retrieved May 1, 2023, from https://www.britannica.com/science/adolescence

Erdoğan, T., & Yurdabakan, İ. (2018). Adaptation of metacognitive awareness of reading strategies inventory: Turkish higher education sample. *Electronic Turkish Studies, 13*(19), 669-680.

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi:10.1177/002224378101800104

Gallucci, M., & Jentschke, S. (2021). *SEMLj: Jamovi SEM Analysis*. [jamovi module].

Grabe, W.P., & Stoller, F.L. (2013). *Teaching and researching: Reading.* Routledge. https://doi.org/10.4324/9781315833743

Hair, J.F., Jr., Anderson, R.E., Tatham, R.L., & Black, W.C. (1995). *Multivariate data analysis with readings* (4th ed.). Prentice-Hall.

Hu, L., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to under parameterized model misspecification. *Psychological Methods*, *3*, 424-453. https://doi:10.1037/1082-989X.3.4.424

Kolić-Vehovec, S., Bajšanski, I., & Zubković, B.R. (2010). Metacognition and reading comprehension: Age and gender differences. *Trends and Prospects in Metacognition Research*, 327-344. https://doi.org/10.1007/978-1-4419-6546-2_15

Köse, N., & Günes, F. (2021). Undergraduate students' use of metacognitive strategies while reading and the relationship between strategy use and reading comprehension skills. *Journal of Education and Learning*, *10*(2), 99-108. https://doi.org/10.5539/jel.v10n2p99

Kusiak, M. (2013). *Reading comprehension in Polish and English: Evidence from an introspective study*. Jagiellonian University Press.

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, *4*(1), 84. https://doi.org/10.1037/1082-989X.4.1.84

Mallinckrodt, B., & Wang, C.-C. (2004). Quantitative methods for verifying semantic equivalence of translated research instruments: A Chinese version of the experiences in close relationships scale. *Journal of Counselling Psychology, 51*(3), 368–379. https://doi.org/10.1037/0022-0167.51.3.368

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Mokhtari, K., Dimitrov, D.M., & Reichard, C.A. (2018). Revising the metacognitive awareness of reading strategies inventory (MARSI) and testing for factorial invariance. *Stud. Sec. Lang. Learn. Teach.*, *8*, 219–246. https://doi.org/10.14746/ssllt.2018.8.2.3

Mokhtari, K., & Reichard, C.A. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology*, *94*(2), 249. https://doi.org/10.1037/0022-0663.94.2.249

Mokhtari, K., & Sheorey, R. (2002). Measuring ESL students' awareness of reading strategies. *Journal of Developmental Education*, *25*(3), 2-11.

Morin, A., Myers, N., & Lee, S. (2020). *Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM) and bifactor-ESEM*. In G. Tenenbaum, & R. C. Eklund (Eds.), Handbook of Sport Psychology (4th Edition). Wiley.

Muthén, B.O. (1993). *Goodness of fit with categorical and other nonnormal variables*. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 205–243). Sage.

Ondé, D., Jiménez, V., Alvarado, J.M., & Gràcia, M. (2022). Analysis of the structural validity of the reduced version of metacognitive awareness of reading strategies inventory. *Frontiers in psychology, 13*, 894327. https://doi.org/10.3389/fpsyg.2022.894327

Öztürk, E. (2012). The validity and reliability of the Turkish version of the metacognitive awareness of reading strategies inventory. *Elementary Education Online, 11*(2), 292-305.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25*(2), 237-272. https://doi.org/10.1177/0265532207086783

RAND Reading Study Group. (2002). Reading for understanding: Towards an R&D program in reading comprehension. Report prepared for OERI.

Rosseel, Y. (2019). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Sarıçoban, A., & Behjoo, B.M. (2017). Metacognitive awareness of Turkish EFL learners on reading strategies. *The journal of Social Sciences Institute of Ataturk University, 21*(1), 159-172.

Schumacher, R.E., & Lomax, R.G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). Routledge Pub.

Soto C., Gutierrez de Blume A.P., Asún R., Jacovina M., Vásquez C. (2018). A deeper understanding of metacomprehension in reading: Development of a new multidimensional tool. *Frontline Learning Research, 6(1)*, 31-52. https://doi.org/10.14786/flr.v6i1.328

Soto, C., Gutierrez de Blume, A.P., Rodríguez, M.F., Asún, R., Figueroa, M., & Serrano, M. (2019). Impact of bridging strategy and feeling of knowing judgments on reading comprehension using COMPRENDE: An educational technology. *TechTrends, 63*(5), 570-582. https://doi.org/10.1007/s11528-019-00383-5

Tabachnick, B.G., & Field, L.S. (1996). *Using multivarete statistics*. Harper Collins Publishers

Taber, K.S. (2018). The Use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48, 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Tary, B., & Molnár, E.K. (2022). A MARSI-R kérdőív magyar adaptációja–olvasási stratégiák vizsgálata anya-és idegen nyelven egyetemi hallgatók körében. *Iskolakultúra, 32*(5), 57-75. https://doi.org/10.14232/ISKKULT.2022.5.57

Tamin, İ.B., & Büyükahıska, D. (2020). Reading strategy instruction on metacognitive awareness: the case of Turkish high school students. *The reading matrix: An International Online Journal, 20*(2), 85-97. https://www.readingmatrix.com/files/23-f957795j.pdf

Villanueva Aguilera, A.B. (2014). *Strategy intervention to enhance reading comprehension of 15-year-old students in Mexico* [Doctoral dissertation, University of York].

Von Eye, A., & Bogat, G.A. (2004). Testing the assumption of multivariate normality. *Psychology Science*, 46, 243-258.

Wall, K. (2008). Understanding metacognition through the use of pupil views templates: Pupil views of learning to learn. *Thinking Skills and Creativity, 3*(1), 23-33. https://doi.org/10.1016/j.tsc.2008.03.004

Wenden, A.L. (1998). Metacognitive knowledge and language learning1. *Applied linguistics, 19*(4), 515-537. https://doi.org/10.1093/applin/19.4.515

Zhang, L., Zhang, L., & Liu. (2017). *Metacognitive and cognitive strategy use in reading comprehension.* Springer. https://doi.org/10.1007/978-981-10-6325-1

APPENDIX

APPENDIX A: FACTOR LOADINGS

Turkish Item [English]	Present Study							Moktari et. al, 2018
Global reading strategies (GRS)	Uni. CFA	CFA	Bif. CFA		ESEM			
			S-β	G-β	β	β	β	
GRS1. Okurken aklımda bir amaç vardır. [Having a purpose in mind when reading.]	.44	.60	.01	.45	**.32**	.19	.27	.54
GRS2. Metni okumadan önce ne hakkında olduğunu görmek için metni gözden geçiririm. [Previewing text to see what it is about before reading.]	.54	.65	-.01	.49	**.44**	.06	.42	.58
GRS3. Metnin içeriğinin okuma amacıma uygun olup olmadığını kontrol ederim. [Checking to see if the content of the text fits my purpose for reading.]	.49	.77	.12	.56	**.59**	.38	.31	.64
GRS4. Önemli bilgileri ayırt etmek için koyu renk yazı tonu ve italik gibi yazımsal yardımcıları kullanırım. [Using typographical aids like boldface and italics to pick out key information.]	.43	.70	-.20	.46	**.21**	.08	.09	.63
GRS5. Okuduğum metinleri eleştirel olarak analiz eder ve değerlendiririm. [Critically analyzing and evaluating the information read.]	.53	.80	.13	.60	**.52**	.07	.38	.67
Problem-solving strategies (PSS)								
PSS1. Dikkatim dağıldığında ya da kafam karıştığında dikkatimi tekrar toplayabilirim. [Getting back on track when getting sidetracked or distracted.]	.61	.63	-.07	.49	.24	**.30**	.29	.60
PSS2. Okuduğum metne göre okuma ritmimi veya hızımı ayarlarım. [Adjusting my reading pace or speed based on what I'm reading.]	.49	.80	-.04	.61	.13	**.29**	.22	.66
PSS3. Okuduklarımı düşünmek için zaman zaman okumaya ara veririm. [Stopping from time to time to think about what I'm reading.]	.46	.84	.05	.57	.22	**.51**	.43	.66
PSS4. Okuduğumu anladığımdan emin olmak için tekrar okurum. [Re-reading to help ensure I understand what I'm reading.]	.60	.56	-.30	.50	.12	**.30**	.06	.59
PSS5. Bilmediğim kelime ve deyimlerin anlamını tahmin ederim. [Guessing the meaning of unknown words or phrases.]	.54	.65	-.02	.52	.29	**.24**	.67	.52
Support reading strategies (SRS)								
SRS1. Okurken not alırım. [Taking notes while reading.]	.57	.62	.28	.42	.20	.25	**.47**	.56
SRS2. Metni yüksek sesle okumak okuduğumu anlamama yardımcı olur. [Reading aloud to help me understand what I'm reading.]	.49	.68	.37	.26	.27	.32	**.45**	.53
SRS3. Anlayıp anlamadığımı kontrol etmek için okuduklarımı başkalarıyla tartışırım. [Discussing what I read with others to check my understanding.]	.56	.81	-.04	.64	.12	.16	**.24**	.68
SRS4. Metindeki önemli bilgilerin altını çizer ya da daire içine alırım. [Underlining or circling important information in text.]	.47	.66	.49	.24	.50	.43	**.70**	.69
SRS5. Okumamı desteklemek için sözlük gibi kaynakları kullanırım. [Using reference materials such as dictionaries to support my reading.]	.52	.71	.12	.48	.15	.27	**.23**	.73

Target ESEM factor loadings are indicated in bold.

APPENDIX B: ITEM CORRELATIONS

		Std.															
	Mean	Deviation	GRS1	GRS2	GRS3	GRS4	GRS5	SRS1	SRS2	SRS3	SRS4	SRS5	PSS1	PSS2	PSS3	PSS4	PSS5
GRS1	3,08	1,271															
GRS2	3,94	1,233	,231**														
GRS3	3,55	1,332	,301**	,358**													
GRS4	3,58	1,421	,256**	,243**	,206**												
GRS5	3,47	1,343	,274**	,319**	,378**	,270**											
SRS1	3,49	1,093	,257**	,217**	,273**	,371**	,229**										
SRS2	3,57	1,285	,133**	,197**	,166**	,278**	,139**	,263**									
SRS3	3,40	1,298	,270**	,282**	,359**	,245**	,416**	,318**	,230**								
SRS4	4,10	1,156	,196**	,233**	,134**	,372**	,184**	,383**	,330**	,268**							
SRS5	2,92	1,319	,244**	,194**	,224**	,238**	,270**	,336**	,266**	,373**	,199**						
PSS1	3,63	1,274	,191**	,232**	,251**	,222**	,274**	,261**	,258**	,317**	,186**	,275**					
PSS2	3,88	1,333	,268**	,343**	,360**	,309**	,349**	,288**	,280**	,334**	,269**	,320**	,333**				
PSS3	3,27	1,469	,289**	,228**	,310**	,225**	,361**	,305**	,197**	,418**	,185**	,396**	,258**	,341**			
PSS4	4,01	1,186	,169**	,282**	,183**	,252**	,205**	,250**	,309**	,266**	,346**	,331**	,253**	,247**	,256**		
PSS5	3,91	1,245	,194**	,257**	,316**	,237**	,326**	,299**	,195**	,301**	,244**	,253**	,286**	,360**	,299**	,216**	
**. Correlation is significant at the 0.01 level (2-tailed).																	

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 345–375

https://doi.org/10.21449/ijate.1267223

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

Enhancing critical thinking and English-speaking skills of gifted students through philosophy for children approach

Feride Acar[1,*], **Recep Şahin Arslan**[2]

[1]Pamukkale University, Institute of Educational Sciences, Denizli, Türkiye
[2]Pamukkale University, Faculty of Education, Department of English Language Teaching, Denizli, Türkiye

ARTICLE HISTORY

Received: Mar. 18, 2023

Accepted: June 22, 2023

Keywords:

Philosophy for Children (P4C),
Gifted students,
English speaking skill,
Critical thinking.

Abstract: This study aimed at investigating the effectiveness of "Philosophy for Children (P4C)" approach on the critical thinking and English-speaking skills of twenty-three 7th grade gifted students learning English as a foreign language at a Science and Art Center (SAC) in the city of Denizli, Türkiye. In the study a mixed methods research design was adopted with the participation of randomly assigned experimental and control groups. While quantitative data were collected through the Cornell Critical Thinking Test (CCTT) X level and the Cambridge English Speaking Test (A2 level for Schools), qualitative data were collected through reflective diaries written by the experimental group participants and evaluation forms. While the control group followed the usual SAC English speaking lessons, English speaking lessons for the experimental group were based on the P4C approach. According to the quantitative data findings, there was an increase in the scores of both groups in terms of speaking skills in English, however this difference was not statistically significant. In terms of critical-thinking skills, there was a slight decrease between the pre-test and post-test scores of the control group and an increase in the scores of the experimental group and this difference was statistically significant. In addition, the qualitative data revealed that the experimental group participants generally provided positive feedback on P4C practices and speaking lessons based on the P4C approach created positive effects on their critical thinking and speaking skills in English, while such an application did not create a significant effect on their speaking skills in English.

1. INTRODUCTION

Current trends in the globalized world and language learning field have highlighted the necessity of adapting the developments and integrating the innovations into English as a foreign language (EFL) classroom (Tosuncuoğlu, 2018). Thus, within the developments witnessed in education in the 21st century, the roles of language teachers cannot be considered limited to teaching only the linguistic aspects of a language or improving language proficiency of their students as they are also supposed to provide opportunities that can enable their students to gain 21st century skills such as questioning, reasoning or critical thinking (CT) and to express

*CONTACT: Feride Acar ✉ feride_dag@hotmail.com Pamukkale University, Institute of Educational Sciences, Denizli, Türkiye

e-ISSN: 2148-7456 /© IJATE 2023

themselves to cope with the challenges they can encounter in their lives (Yang et al., 2005). It has therefore become indispensable for language learners to have such skills within learner-centered pedagogies unlike the traditional teaching methods where the teacher is the source of information and presents the content and where the students are passive receivers of information. Consequently, creating changes in EFL classrooms by implementing activities that enable the learners to use and improve CT skills provides contributions not only for a more effective language learning environment, but also for more independent, autonomous learners, and good thinkers (Yang & Gamble, 2013).

Gifted individuals who are the subject of the present study play a critical role in nations' future and their education has gained significance over the years. Although the issue has critical significance, research concerning to what extent special programs for the education of gifted individuals is effective is not crystal clear. The present study therefore attempts to provide contributions to have an understanding of the education of gifted students in an EFL context in Türkiye. Specifically, this present study attempts to investigate the effectiveness of Philosophy for Children (P4C) approach on English speaking and CT skills of 7th grade gifted students who studied at a Science and Art Center (SAC) in Denizli, Türkiye. To this end, the present study attempts to find out answers to the questions as follows: 1. Can P4C-enhanced English-speaking practices be effective in fostering critical thinking skills of gifted students?; 2. Can P4C-enhanced English-speaking practices be effective in fostering English speaking skills of gifted students?; and 3. What are the participating gifted students' opinions about P4C-enhanced English language speaking practices?

1.1. Critical Thinking

Critical thinking (CT) can be defined as the thinking process that enables the individuals to formulate and construct new knowledge through questioning, searching for reliable information, and asking relevant questions. From Lipman's (2003) point of view, CT is defined as involvement "in all responsible interpretation (the production of meaning), and all responsible translation (the preservation of meaning)" (as cited in McGregor, 2007, p. 192). Ennis (1987) defines CT as "reasonable, reflective thinking that is focused on deciding what to believe or do" (p. 10). In similar definitions, being a critical thinker requires using higher-level thinking skills such as analyzing, reasoning, making evaluation based on standards or proven evidence, and reaching reasonable and rational decisions (Facione, 2015; Kuhn, 2015).

The "Framework for 21st Century Learning" consists of "4Cs"; namely, "critical thinking, communication, collaboration, and creativity" (Zahrani & Elyas, 2017, p. 134). CT has been one of the crucial skills that provide contributions also to the language learning practices of learners by making them be involved in the interactive process of getting and reflecting on information through questioning, assessing, analyzing, and evaluating. As the language is the tool to communicate with people and express ideas or beliefs, the close relationship of using the language as a means of promoting CT becomes clear (Tosuncuoğlu, 2018). It is therefore possible for the language learners to use their CT skill in each of the language skills; namely, speaking, writing, listening, and reading. For instance, language learners can analyze the stance of a writer in a reading class, join a debate in a speaking class, question or criticize a story they watch in a listening/speaking class, and they can also express their opinions on a discussion/debate/video in a writing class. Thus, it can be seen that CT should not be taught as an isolated skill in a different class; however, it should be integrated into all language skills as a vital 21st skill that an individual should develop in our digital and globalized age.

At this point, teachers have a key role in acting as a facilitator to help the learners learn how to think critically by raising good questions, introducing the strategies they can use to become good thinkers, sustaining their motivation to think and reflect on the ideas and to collaborate, and having interaction with their peers**.** For language learning process, CT also has gained significance when the close relationship between the use of language and the requirement of thinking critically within learner-centered methods has been taken into account (Yang & Gamble, 2013).

There have been some studies that search for alternative methods to enhance CT skills of EFL learners and focus on possible outcomes of CT in EFL classes. Davidson and Dunham (1997) investigated the effect of Ennis-Weir Critical Thinking practices to assess the CT progress of Japanese students in an experimental study. The researchers implemented a year of intensive academic English instruction with control and experimental groups. The control group was provided with only content-based intensive English instruction, while the treatment group received additional training in CT. The results demonstrated that the experimental group outperformed the control group significantly in the test which suggested that CT skills can be gained as part of academic EFL / ESL instruction. In another study, Yang and Gamble (2013) explored the effects of a course based on CT-integrated EFL instruction at a university in Taiwan on the English proficiency, CT, and academic achievement of the participants. The participants in the experimental group were involved in CT activities such as information literacy and critical reading (reading), critical reflection/sharing, article critique/peer feedback, debate (listening/speaking), argumentative writing, and peer critique with an emphasis on CT skills (writing). However, the control group received effective language learning practices by following the textbook without emphasizing CT activities. Both of the groups wrote essays about global warming as the final product and their essays were analyzed using the 'Holistic Critical Thinking Scoring Rubric' (HCTSR) to assess their CT skills. Quantitative data demonstrated that the experimental group scored significantly better than the control group did in both overall English proficiency and CT skills. As a result, it was found out that CT-integrated instruction created positive effects on English proficiency and academic achievement of the participants. Similarly, Báez (2004) aimed to investigate the effectiveness of an implementation designed on tasks which could enhance CT skills of 33 students in three groups in Colombia. The researcher examined meaning construction and meta-cognitive processes developed by the students, to what extent these kinds of tasks create an impact on interaction of the students, and the impact of this interaction on English proficiency of the students through CT - enhanced activities. The results of the study demonstrated that the students could make associations and interpretation with the help of their background knowledge, experiences, and beliefs and teacher's role is crucial in motivating the students to explore the depth of the texts. Furthermore, positive developments were observed thanks to the use of feedback, however sometimes there was asymmetry of interaction. In terms of language proficiency, the students showed improvement especially in their lexical and discursive competencies. Nevertheless, they needed support in terms of syntactic, socio-pragmatic, and strategic competence.

The impact of CT on writing skill was searched by Rashtchi and Khoshnevisan (2020), who presented reflections of the classroom practices of a writing class designed to foster CT skill (CT) in an EFL context in Iran as they introduced some sample tasks and provided suggestions on the ways of integrating CT skill into writing and thinking classes for the teachers in their study. Since writing skill requires good thinking and organization of the ideas, a process-based writing approach needed to be employed with the outlining, drafting, and revising steps. By paying attention to the important elements such as questioning, cooperation, and employing organizational skills, the writers provided some suggestions that could make writing classes more effective in a way to foster CT of the students. In the Turkish context, Tosuncuoğlu (2018) looked into the issue from the students' perspectives and investigated the awareness and knowledge of 79 undergraduate students of English Language and Literature Department at Karabük University, Türkiye about CT. In this study, research findings showed that students' awareness about thinking critically was not at a desirable level. As a result, an education program that serves to the contribution of CT improvement is suggested by the researcher. In another study carried out with the participation of 34 undergraduate university students at Karadeniz Technical University, Türkiye, Arslan and Yıldız (2012) examined whether a literature-based critical-thinking program could create a positive impact on CT skills of students' and teachers' beliefs about the literature course. Within the treatment, the participants were encouraged to engage in dealing with various literary works and the CT activities were designed in a way that enabled the participants to practice the cognitive levels of Bloom's taxonomy. According to the results of the study, it was found out that there was a significant change in CT levels of the participants considering pre-test and post-test scores. The participating instructors also reported positive attitudes towards the implementation as it contributed to their professional development and made a more student-centered learning environment possible. The results of the study might be considered significant as it provides an example of how to integrate literature practices that can enhance CT skills into Turkish education system at tertiary level.

1.2. Philosophy for Children (P4C) Approach

The Philosophy for Children (P4C) approach has its roots from Socratic legacy which is based on reasoning as Socrates used his methods of questioning and dialogue to enhance good thinking skills of people (Chamberlain, 1993). P4C approach has therefore been considered as an effective way of improving thinking skills as well as emotional and social skills of especially gifted individuals (Sutcliffe, 2004) and has also been seen as a response to the needs of people who have to overcome the challenges and hardships in the 21st century.

P4C practices generally begin with the use of a stimulus which can be a story, a poem or a video that involves philosophically rich content or an object to take the attention and make pupils ask logical questions to start a discussion. Children are involved in a 'community of inquiry' in which they are encouraged to show respect to each other's ideas or opinions and to reflect more deeply than their usual ways. Moreover, children carry out discussions by expressing agreement or disagreement, making suggestions, realizing the various perspectives, creating a connection between the concepts and their own experiences or emotions. In this way, with the help of the constructed dialogues between them, the social aspect of the youngsters is also improved. The teacher in this environment acts as a facilitator but does not interrupt the youngsters' speech or direct them, does not express his/her opinion, or does not provide the answer in a short time when the youngsters do not have responses. The key point for the teacher is to first create a peaceful environment in which the youngsters feel comfortable to express their opinions or emotions and strange ideas are welcomed. In addition, the youngsters are encouraged to cooperate with each other rather than being involved in a competitive environment. In this way, they gain the opportunity of constructing the knowledge and concepts within a cooperative learning environment and enhance their discussion skills. The idea of designing EFL classes as not only language learning environments but also social interaction settings which favor thinking deeply and discussing on life issues was first proposed by Pishghadam (2011) in a seminal paper (Dabbagh & Noshadi, 2016). When language learners are exposed to philosophical questions related to different domains of their daily life, they have been observed to be more motivated to answer and participate in discussions or debates.

When the possible outcomes of the P4C approach are taken into consideration, there have been many studies that seek to find out the effectiveness of P4C in improving CT skills of the learners. Some of them were carried out in EFL contexts within the Philosophy-Based Language Teaching (PBLT) in order to investigate whether P4C can be an effective tool to improve CT and English proficiency or not. In a pioneering research, Shahini and Riazi (2011) carried out a study in EFL classrooms in Iran to assess the development of students' speaking, writing, and thinking skills. The experimental group was involved in practices in which the PBLT techniques such as asking alternative views, clarifications and reasons were used while the control group was exposed to ordinary or non-philosophical questions. According to the results of the study, a significant difference was found between the experimental and control groups in speaking and writing skills. The results of the study showed that PBLT may be an effective tool to foster students' CT skills in an ELT context as it enables students to respect different views, explain concepts, apply reflective thinking, and think critically.

Based on the assumption that children have the potential to philosophize starting at an early age, Lam (2013) investigated whether Lipman's P4C approach could be effective in fostering CT skills of Chinese Secondary 1 students in Hong-Kong or not. During the treatment, P4C sessions were carried out by using the philosophical novel "Harry Stottlemeier's Discovery" (Lipman, 1982). After the sessions, the experimental group students were supposed to write on the My Thinking Log as a follow-up activity. The control group was exposed to the traditional language classes by using novels with a similar level without an emphasis on philosophy or reasoning. According to the results of the study it was found that the participants of the study who received philosophy-based instruction were found to have the capability of conducting debates, good thinking, reasoning, and reflective thinking abilities through the implementation of P4C approach and showed a greater performance in displaying CT and reasoning skills compared to that of the control group. In a more recent study, Lam (2020) also examined whether CT and English proficiency can be enhanced through philosophy in ESL classrooms or not. Employing the basic principles of P4C approach, the study was conducted based on a program called Philosophy in Schools (PIS). 62 Chinese secondary students from a school and 57 secondary students who used English as a second language from another school participated in the study. The results of the study supported the previous research which suggested that philosophy-based instruction can be an effective alternative to engage the individuals in a social community of inquiry by making them learn how to think, reason, question, and have interaction with their peers and develop their CT skills. The study also highlights the close relationship between the use of language and thinking skills by putting an emphasis on especially the speaking skill that plays a key role in cognitive development.

1.3. Improving English Speaking Skills

One of the main concerns of this study is to see the possible effects of implementing philosophy-enhanced approach on English speaking skills of EFL learners. Speaking skill constitutes the basis of communication and also is the most difficult skill to be improved by foreign language learners (Oradee, 2012) as EFL learners generally have difficulties in expressing themselves orally and using the language for communicative purposes. It is highly possible that this stems from lack of exposure to real life situations through which they can have the opportunities of practicing and using the language as a response to their needs (Afrizal, 2015). In addition, they do not have the chance of having interaction with native speakers as a way to encounter cultural properties of the target language and improve their oral proficiency skills. As an alternative to create communicative language learning environments as much as possible under these conditions, investigations into the effects of various methods on speaking performance of the language learners have become a prevalent research concern. As Lipman (2003) states, discussions and also cooperative activities within P4C can enable the learners to use higher-level skills, construct their own beliefs, and as a result strengthen their CT and reasoning abilities. As a vehicle to foster these skills, learners use the 'language' in expressing their opinions and getting feedback while they are involved in the community of inquiry. Thus, it has been clear that there is a close relationship between the use of language and philosophy-based instruction.

As an alternative method, improving English speaking skills of EFL learners through philosophy-based language teaching approach has become a research concern recently (Dabbagh & Noshadi, 2016). There are some important key factors that should be taken into consideration while implementing the P4C approach in language classes. The classrooms are considered as a 'community of inquiry' in which the members cooperate and discuss to find a solution to a common problem or accomplish a philosophical task. This component of P4C enables the students to take the responsibility of their own learning which contributes to their autonomy. Another component is that the students are involved in philosophical dialogues which refer to some problems that can be solved through deep thinking. The students can also use the language more effectively while expressing their opinions, gain analysis skills while coping with the challenges and problems, have persuading and communication skills when they express agreement or disagreement, support their ideas and try to convince their peers, learn to be more respectful towards different or extraordinary ideas or beliefs, and learn how to support their group partners or the other students to reach a consensus on a common problem (Rustam et al., 2018).

When the related literature is examined, it can be seen that various methods are used to enhance English speaking skills of the learners. While in some studies, the effects of conducting discussions or debates in the classroom on speaking skill of the learners were investigated, in some studies whether employing digital tools that may provide the opportunity of practicing speaking for the learners alone and prevent the anxiety of speaking in a community could be influential or not was studied by the researchers. One of the most directly related empirical studies with the present study was carried out by Rustam et al. (2018), who aimed to find out whether group-discussions within the philosophy-based language teaching approach (PBLT) can be an effective alternative to promote English oral proficiency of the undergraduate students in Indonesia or not. Data obtained in their study showed that implementation of PBLT approach in language classes provided contributions to English speaking proficiency and also to social and emotional development of the students. Moreover, students' attitudes towards English changed in the positive way and their motivation to participate in the discussions increased.

Using discussion or debate has been considered as one of the effective ways of improving speaking skills of the language learners. In this context, Afrizal (2015) investigated the effectiveness of classroom discussion on English speaking ability of students from Almuslim University in Kudus and found out that classroom discussion placed a positive impact on speaking performance of the participants. In the study, discussions were mentioned as methods that increased motivation of the students to express themselves orally, engaged the students in the learning process actively, and provided the opportunity of having interaction with their peers and teachers. A more recent study carried out by Haryanti et al. (2021) aimed to foster English oral fluency of 64 students at the 11th grade in Indonesia through Three Steps Interview Technique. Being one of the cooperative learning strategies, this technique is mentioned as a communicative activity carried out with an interlocutor by asking and answering questions and having discussions. Findings obtained in the study suggested that Three Steps Interview Technique is useful in improving students' speaking ability effectively. The positive outcomes of the treatment were considered to have stemmed from the fun atmosphere during the implementation of the technique and cooperation which led to an increase in students' motivation to express themselves orally. In the light of the studies, it can be stated that there have been numerous attempts to foster and facilitate speaking skills of the language learners in EFL or ESL contexts. Most of the employed methods or techniques were reported to be effective in order to reach the aim. However, there has been no research reported in the literature in which gifted students were the subjects participating in the Philosophy for Children program with the aim of enhancing their CT or English-speaking skills in the Turkish context.

Thanks to the present study, it is expected that the handled findings might be enlightening for the education of gifted students in Türkiye. Furthermore, this study has been considered significant as it aims to provide contribution to the development of 21st century skills of gifted students by making them use their higher-level thinking skills. It is also assumed that P4C-enhanced speaking practices also can create a positive impact on English speaking skills of the gifted students with the help of an interesting and challenging discussion environment.

2. METHOD

2.1. Research Design

This study is a quasi-experimental study with pre-test and post–test administration and involves the participation of experimental and control groups. Convenience sampling method was adopted in choosing the participants of the study. One of the groups was the researcher's group within the regular program in SAC and it was assigned as the experimental group. The researcher did not have another 7th grade group at the time of the study. As a result, the researcher created an elective English-speaking class and the students who wanted to take part in this class formed the control group.

Since critical thinking, one of the main focuses of the current study, has been considered as complex and multifaceted to be measured correctly, many psychologists and educators have centered upon the difficulties in understanding or effectively assessing CT (Arter & Salmon, 1987; Chamberlain, 1993). To be able to reach more valid and reliable findings in measuring CT, collecting both quantitative and qualitative data to provide a greater depth of understanding and to increase the "strength and rigor" of an investigation (Patton, 1990, p. 60) was preferred by the researcher. Considering the emphasis on the importance of employing multi-methods design in data collection process of the research; in this study, quantitative data were collected through critical thinking and English-speaking tests and qualitative data of the study were gathered through student-reflective journals and responses of the participants to an evaluation form which has four questions based on the main focus of the study.

The sample of the present study comprised 23 seventh grade gifted individuals who were accepted to a Science and Art Center in Denizli, Türkiye. Science and Art Centers (SACs) are the main institutions for gifted individuals as after the school institutions in Türkiye that create a response to the needs and expectations of exceptional children from 2nd, 3rd or 4th grades till the end of high school. SACs aim to make gifted students become aware of their talents and discover their potential, foster their skills, and assist them for their future progress. As well as the other courses, the students have two English classes within their weekly schedule and can also prefer to join elective foreign language classes. In SACs, there is no established curriculum but a program that suggests some themes and objectives. Teachers have the freedom of designing and conducting their teaching practices considering the needs of their students and making efforts to improve the students' skills, project management process, and also intellectual development. It is clear that teachers aim to foster the 21st skills of the students such as creativeness, questioning, evaluating objectively, and also CT.

The center where this study was implemented was located in the center of the city and had 645 students at the time of data collection. Some students who lived in the urban areas and identified as gifted came to the center generally at the weekend, because they had classes in their schools on weekdays, and it could not be possible for them to catch the classes due to transportation problems.

The age range of the students in the study was between 12 and 13, and they were studying at different public and private schools in the province of Denizli. All the students spoke Turkish as their mother tongue, and they learned English as a foreign language both in their schools and in the SAC. The students were accepted to the SAC within the General Talent area, and they were at the *recognizing individual abilities* (RIA-2) stage during the study. The students chose English as a main field that they would be studying in SAC till the end of high school. Information regarding the gender of the participants is presented in Table 1.

Table 1. *Gender distribution of the participants.*

Gender	Group		TOTAL
	Experimental Group	Control Group	
Female	7	6	13
Male	5	5	10
TOTAL	12	11	23

Students' English proficiency levels showed differences as they were involved in different programs in their schools. While the students studying at public schools had four English classes per week at the seventh grade, some of the students studying at private schools joined around 10 to 15 English classes. The students attended classes in SAC once or twice a week according to their school's weekly schedule. They had two English classes per week in SAC, however if they demanded, it was possible for them to join elective English classes as well.

The researcher had 12 students in one of the groups within the regular program in SAC. As implementing P4C-enhanced approach requires a certain level of English proficiency level for the participants and the researcher did not have another 7th grade group in that education term, the researcher created an elective course under the name of *English-Speaking Course* in order to carry out the experiment. The researcher announced it to the 7th grade students in the center and 11 students applied to take part in the course. As Cohen et al. (2007) state, "the larger the sample, the better, as this not only gives greater reliability but also enables more sophisticated statistics to be used" (p. 101). However, as the program in the *recognizing individual abilities (RIA-2)* stage in SAC does not allow grouping students in high numbers, it was not possible for the researcher to have more students in the groups. In addition, as the participants belonged to a minority group (gifted individuals) in the society, and they were required to have some certain qualities in order to be involved in the study, the representativeness and generalizability of the results might be expected to be much higher. The researcher separated the students into two groups as sessions 16.30 and 18.10, and the students preferred to join one of the classes considering their weekly schedule.

2.2. Quasi-Experimental Procedure

This study employed a pre-test-post-test control and experimental group design. While the P4C-enhanced speaking course participants were assigned as the experimental group, English speaking elective course participants were assigned as the control group. Design of the study can be seen in Table 2.

Table 2. *Quasi-experimental design of the study.*

Group	Pre-test	Treatment	Post-test
Experimental Group (E)	P1. E	+	P2. E
Control Group (C)	P1. C	-	P2. C

In this study, while the experimental group was exposed to P4C-enhanced speaking activities, the control group was involved in speaking classes within the regular program of SAC (See Appendix 1 and Appendix 2, respectively). The duration of the sessions was the same for both groups as 10 weeks, which included two classes of 40 minutes per week. Before starting the implementation, parents of all the participants were requested to sign a parental permission form as the participants were under the age of 18. The participants were also informed about the study in terms of ethical concerns. After the Ethical Committee Report was provided from Pamukkale University and the necessary permission was taken from The Ministry of National Education, Critical Thinking Tests and English-Speaking Tests were administered in both of the groups as pre-test during the first week. Before beginning the sessions, the participants were provided with information concerning the procedure and also the scope and sub-skills of critical thinking.

The first session was implemented in mother tongue (Turkish) of the participants as the researcher wanted them to gain some knowledge about the treatment. The terms such as "facilitator, participants, community of inquiry" were introduced and the principles of the discussions were identified and written down on a paper with the students and the paper was hung on the board as Wartenberg (2009) suggests. This board was created for the research on the wall of the classroom and included some common phrases and questions that they would be using, and the list of rules that they were supposed to consider during the discussions. The rules that were defined by the experimental group and the facilitator included: *We show respect to others' opinions, beliefs and ideas; We learn from each other; We listen to each other actively; We question, explain, and reason; We make an explanation for our opinions/choices; We are not biased; We criticize others in a polite and objective way; We accept others' fair criticisms towards us; We express our opinions freely; We evaluate from various perspectives;* and *We can construct and conduct and argument.*

In order to create a peaceful atmosphere and to establish a rapport between the researcher and the students in the classroom, they were also instructed that there was no authority, everybody was free to express their ideas or opinions in a respectful way, and they were to show respect to different ideas. While giving the instruction during the sessions, the researcher used Turkish in some critical points that would create an effect on comprehension of the participants or lead to misunderstanding. The researcher acted as a facilitator and a moderator during the discussions.

In a P4C session, the facilitators need a stimulus such as a story, a picture, a video, or a poem that has the potential of taking the attention of the participants at the beginning phase. In order to create a deeper understanding of how to apply this approach in classes for educators or teachers, a description of a session implemented by the researcher is provided here (see Appendix 1 for more details). In a P4C implementation constructed on the concept of freedom and social identity, the facilitator starts the session with a short video from Robinson Cruise, who survived in a deserted island after a plane crash. After the students watch the video, they are instructed to suppose that they are lost in an island and stay alone, need to explore the island, and find the ways of surviving on the island. At this point, they are requested to be groups of 3-4 people with a poster, pencils, and crayons and are told that they are going to create their own society by discussing and having common decisions as a group. Then, an instruction paper that involves information about what they should do is handed out. They are supposed to meet on a common point in terms of the name, population, currency, income sources, governmental issues, freedom limits, and rules of the society and reflect them on the posters. While the group members discuss and study on their posters, the facilitator observes the groups without any inducement. After they complete their posters, they are requested to choose a leader who is going to introduce their work. During the presentations, the facilitator addresses philosophical questions to the group members such as: "How did you share decision-making?, How did you reach a consensus?, Is it possible for a society to be completely free?, and Would you like to live and govern in the society you create?". In this way, a big group discussion is carried out by making them to think deeper on the process and on the concepts. The students are required to provide a reason or explanation for their choices. The facilitator does not interrupt with their answers or express his/her opinion on the issue. S/he summarizes the participants' expressions when needed and takes the attention to the matters that the participants cannot agree on and cannot reach a common point. If the group members do not participate in the discussions actively, the facilitator creates dilemmas and addresses controversial questions to make them express their opinions. At the end of the session, the participants are requested to construct a meaning of freedom on their own.

As it can be seen from the session described, the facilitator is prepared in terms of addressing the effective philosophical questions at the right time. And also, the facilitator is aware of the potential responses of the participants before the session which enables her to address new provoking questions during the flow of the discussion. As the participants are supposed to explain their choice or disagreement, they handle the opportunity of fostering their reasoning skills as well. In this session, the objective of the session is not only to make the participants to collaborate with a group-work, a big discussion is also carried out after the presentations through which deep thinking and meaning construction of the concept can be realized. As a conclusion, it can be stated that this P4C implementation creates a contribution to the development of discussion, collaboration, leadership, creativity, reasoning, and social skills of the participants as well as using the target language for communicative purposes.

2.3. Data Collection Instruments

In this study two instruments were used to collect quantitative data on the effect of P4C instruction; namely, *Cambridge A2 Speaking Test for Schools* to measure English speaking skills of the participants and *Cornell Critical Thinking Test Level X* to measure critical thinking skills of the participants. Cambridge speaking test was preferred in the study as it is a widely used test around the world. It also includes the rubric which is scored over five points for each criterion. The rubric involves *grammar and vocabulary*, *pronunciation*, *interactive communication*, and *global achievement* criteria. The speaking tests were recorded in video format as the researchers wanted to watch each one several times to increase the objectivity and reliability in scoring. Moreover, the researchers administered the tests with another teacher as an assessor to increase interrater reliability and compared the results.

Cornell Critical Thinking Test (CCTT) X level was used to measure critical thinking skills of the participants in pre-test and post-test. CCTT, one of the most widely used CT tests all over the world, was developed by Ennis and Millman (1985) in multiple choice format (three choices) and involved 75 questions in total. Five of the questions were already answered as examples; as a result, test-takers were required to answer 71 questions. X level was suitable for individuals graded between four and 14. Maximum time for the administration of the test was 80 minutes based on the grade of the test-takers. Reliability values of the test range between 0.67 and 0.90 according to the international studies (Ennis, Millman & Thomko, 2005). In the Turkish context, Akar (2007) translated CCTT into Turkish in his doctoral dissertation, and the Cronbach Alpha reliability value of the test was found to be 0.71 in the pilot administration. Sub-skills that were measured in the test included: *Deduction through inductive reasoning (23 items); Deduction through deductive reasoning (14 items); Questioning credibility and reliability of the sources (24 items); and Identification of assumptions (10 items).*

In order to eliminate the risk of misunderstanding resulting from language for the test-takers, CCTT was administered in the Turkish format. The Company Palindrom is the legal representative of the CCTT in Türkiye, and the tests were purchased from the company. The company sent the printed tests to the researchers, and after the administration, these tests were sent back to the company. Scoring of the CCTT was made by the Palindrom, and the statistical analysis of the tests was carried out by the researchers.

In this study, two different instruments were used to gather qualitative data. One of the instruments was the student-reflective journals. The participants were instructed to write in their journals in the last five or 10 minutes of the sessions, which enabled the participants to write as a way of reflection on the discussions, materials, or stimulus presented at the beginning of the practices. The participants were instructed to write freely, and in this way, they gained an opportunity of sharing their feelings, ideas or thoughts about the implementation, especially for the ones who did not prefer expressing themselves orally during the discussions. They could also write about some private issues that they were hesitant to discuss with their classmates. While they were told that they would feel free to write anything they wanted, they were also instructed for the issues they needed to take into account in reflection on the sessions with an aim to determine a framework for their writing so as to carry out the content analysis in a more accurate way. Another contribution of journal writing for the participants was also providing the chance of summarizing the discussions and thinking over the talk during the sessions. As a result, the researcher was able to explore the themes and emerging issues and get an insight concerning the effects of the implementation on the participants.

The other instrument employed to collect qualitative data was a written evaluation form including four questions that were designed by the researcher to enable the participants to reflect on the effects of the treatment overall at the end of the process. 12 participants in the experimental group were requested to explain their answers in order to provide in-depth data and give them the chance of reflecting on their experience in their own words. Written forms were preferred instead of an interview with an aim for the participants to eliminate the risk of not sharing what they were thinking exactly with the teacher. It would also be probable that they would not make up their minds and organize their ideas within the flow of communication and in a definite amount of time if they were interviewed. The participants were also informed that they were not supposed to write their names on the forms to make them feel more secure and free to share their opinions. The questions of the form were constructed on the variables of the research and are as follows: 1. Which of your skills do you think P4C practices fostered? (Critical thinking, questioning, analyzing, evaluation, showing respect to different opinions, etc.) Please explain; 2. Do you think that P4C practices which were implemented in English created a positive or negative impact on your English-speaking skill? Please explain; 3. Do you think that P4C practices have some drawbacks for you? Please explain; and 4. Do you think that you could look from different perspectives and think differently from the other people in your daily life after joining P4C sessions? Please explain.

2.4. Data Analysis

As the copyright owner of the CCTT, scoring of the tests was done by the Company Palindrom in Türkiye, and the results were sent back to the researcher. Each multiple-choice item of the test has one correct answer; however, one item in the test can measure more than one CT sub-skill. For this reason, as it can be seen in the tables of the results section, total score of the test is higher than the sum of all CT sub-skill scores. The company stated that they had a special algorithm for calculating the test which was not publicly available. The company had the answer key for the test and checked the answer sheets of the participants. Before deciding which test to run in the data analysis of the CCTT, normality of the data was checked. The results of the normality test can be seen in Table 3. As the next step, results of the CCTT were analyzed by employing independent sample *t*-test via the Statistical Package for Social Science (SPSS) program.

Table 3. *Normality test results of the CCTT pre-test and post-test scores of the groups.*

	Group	Kolmogorov-Smirnov^a			Shapiro-Wilk		
		Statistic	*Df*	Sig.	Statistic	*Df*	Sig.
Pre-test	Control-group	.217	11	.155	.922	11	.339
	Experimental group	.134	12	.200*	.978	12	.972
Post-test	Control group	.215	11	.167	.925	11	.364
	Experimental group	.180	12	.200*	.962	12	.806

* $p < .005$

While determining normality of the data, as Büyüköztürk et al. (2014) and Demir et al (2016) state, the number of the sample should be taken into consideration. If the sample size is big, the Kolmogorov-Smirnov test is recommended, however if the sample size is small, the Shapiro-Wilk test is recommended. Since the number of participants in this study was below 30, Shapiro-Wilk value was taken into consideration in evaluating normality of the data. As it can be seen from Table 4, the pre-test of both the groups can be considered as normally distributed (*p* > 0.05). After examining histograms and QQ plots for both of the groups as well, it became possible to assume normality and decide to use a parametric test.

Table 4. *Normality test results of the pre-test and post-test scores of the groups.*

	Group	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	*Df*	Sig.	Statistic	*Df*	Sig.
Pre-test Scores	Control Group	.189	11	.200*	.898	11	.172
	Experimental Group	.193	12	.200*	.918	12	.271
Post-test Scores	Control Group	.125	11	.200*	.966	11	.847
	Experimental Group	.170	12	.200*	.929	12	.372

* *p* < .005

In the analysis of the quantitative data collected through Cambridge Speaking Test, speaking test performances of the participants were assessed by the interlocutor (researcher) and the assessor (participating teacher), who took part in the test implementations. In order to determine interrater reliability of scoring the instrument, performances of the participants were recorded in video format and watched several times after the implementation. The results obtained by the two scorers were compared, and it was seen that there was a high consistency between the scores. In the last step, arithmetic means of the points were calculated and the final scores of the participants were identified. At the first step of the data analysis, normality tests were run via SPSS software program to check whether the data were distributed normally or not. After it was seen that the data were distributed normally, independent sample *t*-test was used to find out whether there was a significant difference between post-test scores of the control and experimental groups or not.

Since the number of participants was limited in this study, Shapiro-Wilk value was taken into consideration in evaluating normality of the data. As it can be seen from Table 6, both of the groups' pre-test can be considered as normally distributed (*p* > 0.05). After examining histograms and QQ plots for both of the groups as well, it became possible to assume normality and decide to use a parametric test.

In this study, to analyze the qualitative data gathered through student-reflective journals and evaluation forms filled by the experimental group students, content analysis was carried out. Content analysis is defined as "any technique for making inferences by systematically and objectively identifying special characteristics of messages" (Holsti, 1968, p. 604). The aim was to find out common thematic elements throughout the research, so, as the first step, the reflection written on the student journals and the responses given in the evaluation forms were initially examined several times and then similar statements were coded and categorized. In the coding step, the common and emerging answers based on the themes were found out. In order to identify the data, the most common responses were categorized basically as concepts, key themes were found, and the frequencies of the concepts were analyzed. The most common responses were illustrated by tables for the evaluation forms.

To strengthen the reliability of the content analysis, stability, reproducibility, and accuracy factors were also taken into consideration (Palmquist, 2012). In terms of reproducibility and accuracy factor, the data were examined at different times by the researchers. Furthermore, while presenting the findings in the results section, direct quotations that belong to the participants were shared to illustrate the emerging themes and establish the credibility of qualitative data.

3. RESULTS

3.1. P4C-enhanced English Speaking Practices in Fostering Critical Thinking Skills of Gifted Students

In order to find an answer to the first research question of the study, CCTT was administered as pre-test and post-test. The test scores were calculated by the Company Palindrom and the researcher carried out the data analysis by using SPSS software program. Before deciding which analysis test to run, normality of the pre-test and post-test scores of the groups was checked. After it was found out that the data were distributed normally, an independent sample *t*-test as a parametric test was employed to test whether there was a significant difference between pre-test and post-test scores of the groups or not. The results of the test can be seen in Table 5 and Table 6.

Table 5. *CCTT Results of the independent sample t-test.*

	Group	N	Mean	Std. Deviation	Std. Error Mean
Pre-test	Control-group	11	39.3636	5.31550	1.60268
	Experimental group	12	41.4167	5.05350	1.45882
Post-test	Control-group	11	38.6364	6.56160	1.97840
	Experimental group	12	43.8333	5.07818	1.46594

Table 6. *CCTT Results of the independent sample t-test.*

Variable	Group	N	M	SD	*t*	*p*
Critical thinking	Experimental Group	12	43.8333	5.07818	-2.111	.048
	Control Group	12	38.6364	6.56160		

* $p < .005$

When the results of the independent sample *t*-test are considered, it can be seen that mean values of the pre-test scores of the control group and experimental group were similar (Experimental group = 41.41, Control group = 39.36). Even if there was a difference, it was not found statistically significant ($p > 0.05$). This result shows that both of the groups had similar critical thinking skills test scores at the beginning of the study.

On the other hand, mean values of the post-test scores of the groups show difference, namely the control group= 38.63 and the experimental group= 43.83. While the experimental group showed progress in post-test results, the control group's post-test scores were found to be lower than their pre-test scores. As the last step of the analysis, the difference between post-test scores of the groups was checked and found to be statistically significant with the p value .048 ($p > 0.05$). These results suggest that the experimental group which received treatment showed a higher performance in critical thinking skills test when compared to the control group which followed the routine speaking course and did not receive the treatment. In the light of the findings, it can be concluded that P4C-enhanced speaking practices created a positive impact on CT skills of the gifted students.

3.2. P4C-enhanced English-Speaking Practices in Fostering English Speaking Skills of Gifted Students

In order to investigate the answer for the second research question of the study, several steps were to be followed to identify whether P4C-enhanced speaking practices crated a positive impact on English speaking skills of the experimental group or not. For this aim, first of all, normality of pre-test and post-test scores of the groups was tested via the SPSS software program. After it was determined that the data of the English-speaking tests were distributed normally, independent sample *t*-test was run to examine the difference between pre-test and post-test scores of the participants. The results can be seen in Table 7 and Table 8.

Table 7. *Independent sample t-test results of the Cambridge Speaking Test.*

	Group	N	Mean	Std. Deviation	Std. Error Mean
Pre-test	Control Group	11	16.2727	1.23215	.37151
Scores	Experimental Group	12	16.5833	1.24011	.35799
Post-test	Control Group	11	16.9545	1.57249	.47412
Scores	Experimental Group	12	17.6667	1.51257	.43664

Table 8. *Independent sample t-test results of the Cambridge Speaking Test.*

Variable	Group	N	Mean	SD	*t*	*p*
English speaking skill	Experimental Group	12	17.6667	1.51257	1.105	.282
	Control Group	12	16.9545	1.57249		

* $p < .005$

According to the independent *t*-test results, it can be seen that mean values of the pre-test scores of the groups show quite similarity (Control group = 16.27; Experimental group = 16.58). No significant difference was found between pre-test scores of the control and experimental groups ($p > 0.05$). As a result, both of the groups had similar English-speaking proficiency at the beginning of the study. When the difference between post-test scores of the groups is compared, it can be reported that experimental group (M= 1.08 / difference) outperformed the control group (M= 0.68 / difference) in the English speaking test; however, the difference between post-test and pre-test results of the experimental and control groups was not found statistically significant considering the p value .282 ($p > 0.05$). These results indicate that, P4C-enhanced speaking practices fostered English speaking skills of the experimental group when compared to those of the control group which did not receive the treatment.

3.3. Participating Gifted Students' Opinions about P4C-enhanced English-Speaking Practices

In this study, student-reflective dairies were used to collect qualitative data as a way of revealing opinions of the experimental group participants about the treatment they received for ten weeks within this study, which aimed to support the quantitative data to increase the possibility of obtaining more reliable results. At the end of the each P4C session, for 5 or 10 minutes, experimental group students (N=12) wrote about their opinions, feelings or ideas considering the session they experienced in their journals in English. They got the opportunity to reflect on the sessions by writing freely and expressing their opinions. In some sessions, they mentioned the challenges or difficulties they encountered during the discussions. As the teacher did not interrupt with students' speech or force them to speak, some students could explain their choices, opinions or preferences concerning the studied theme or dilemma by writing in their journals. However, some students were not so eager to give detailed information and they preferred to write shortly. Journal writings were analyzed with content analysis to find out the main categories with emerging points, patterns, and tendencies of the participants.

Considering students' pieces of writing in their journals, it was seen that they occasionally used journal writing as a way of making a summary and concluding their opinions after the sessions. From a general perspective, they generally wrote about the new English words they learned, the points that were confusing and challenging for them, the skills that they could foster during the discussions, whether they shared the same opinion with their group members or not and the new perspectives they could realize. After the analysis, two main categories were identified based on the student-reflective journals: *Positive experiences and advantages of P4C practices* and *negative experiences and disadvantages of P4C practices*. Statements of the participants concerning the main categories that reflect their opinions are shared in the next part.

3.3.1. First Category: Positive Experiences of the Participants / Advantages of P4C Practices

Based on the analysis results of the student-reflective journal writings, it can be concluded that most of the participants reported positive comments regarding different dimensions of the P4C sessions. Six themes were identified under the first category in the data analysis process. These themes were based on the outcomes of the P4C practices which were: *Fostering collaboration skills, contributing to gaining new perspectives, enhancing discussion skills, making people think deeply and question the truth, teaching new English words, and providing a different experience from the participating students' regular school environment.*

By addressing the first theme of the first category, *fostering collaboration skills*, five of the participants stated that P4C discussions contributed to their collaboration skills through group-work and group-discussions. Following quotations of the participants can be given as examples to illustrate:

> *(1) I learned how to live in a community and designed a new island thanks to the collaboration with my group friends. (P1/2nd session)*[†]
> *(2) … we tried to make a common decision with our group friends, collaboration was useful. (P2/2nd session)*

For the second theme, *contributing to gaining new perspectives*, which emerged from the first main category, six of the participants expressed their opinions about the new perspectives they could realize during the P4C practices. They mentioned that they did not look from that perspective earlier or they could realize something new about their point of view during the discussions. The following quotations can be helpful to support this finding:

> *(3) I realized that in order to trust someone, I need to know him/her. (P3/6th session)*
> *(4) I realized that equality and justice are not the same things. (P2/2nd session)*
> *(5) I noticed that some bad things may be good for us or for our environment depending on the situation… (P10/1st session)*

Third, fourth and the fifth statements support each other in terms of reflecting the participants' satisfaction that they could discover something new about themselves. They came through these results by responding to a questionnaire about 'How emphatic I am' during the P4C session. On the other hand, the fourth quotation indicates that the participants could distinguish between concepts and construct meanings of their own regarding the main themes that focused on the P4C sessions. This finding serves the objectives of the P4C pedagogy which aims to assist the community of inquiry to think deeply and construct new knowledge through questioning and discussing.

Based on the third theme which was *enhancing discussion skills*, as a natural property of the group discussions, eight of the participants expressed opinions that they experienced agreeing or disagreeing with their friends during the discussions. Some of them mentioned that they changed their ideas after listening to their friends and accepted that their friends could be right from that point of view:

> *(6) I changed my idea after we discussed with my friends; it was hard to decide… (P7/8th session)*
> *(7) We had some common ideas with my friends; we shared the same idea today. (P5/2nd session)*
> *(8) … I agreed with my friends in general. (P12/8th session)*

[†] (P… is used to show the number of the participant, and the number of the session for which the participants expressed opinion is given between the parentheses).

These remarks suggest that the participants attained the opportunity of expressing agreement or disagreement during the discussions which contributed to the development of their discussing skills. In this way, they could express their opinions freely and learned to defend their ideas in a respectful way.

Based on the fourth theme, *making people to think deeply and question the truth,* which arose from the main categories of the findings, four of the participants reported that P4C practices enabled them to think deeply and question the knowledge they encountered. Following statements might be striking to support this finding:

> *(9) I thought deeply everything in this class… (P5/6th session)*
> *(10) …the story was nice; it consistently directs people to question… (P7/7th session)*
> *(11) It was hard for me to decide between valid and sound statements. I was confused when I questioned them … (P1/4th session)*

These expressions are considered extremely valuable as they indicate that the P4C practices encouraged the participants to question the truth they encounter and think deeply. Even if they mentioned these concerns as hardships or troubles they experienced in the sessions, they are in fact the challenges that the community of inquiry members are supposed to struggle with and lead them to think deeply during the P4C inquires. These findings suggest that P4C practices provided contributions to the development of higher-level thinking skills on the part of the participants.

Nine of the participants also stated that they learned new words during the sessions about various subjects that they did not learn in regular classrooms as a supporting data for the fifth theme which was *teaching new English words*. Some of the participants also mentioned that they were motivated to speak English as the issues focused on during the debates were challenging and interesting for them even if they did not consider themselves proficient in speaking English. The following examples can show their opinions:

> *(12) I learned what equality and justice are… (P6/2nd session)*
> *(13) I learned new words such as prejudice and disabled in this session. (P3/3rd session)*
> *(14) Today we learned new words about epistemology… (P1/4th session)*
> (15) *I don't speak English well, but I tried to speak because the topic was interesting… (P7/2nd session)*

The sixth theme based on the *positive experiences of the participants / advantages of P4C practices* category reflected favoring opinions of the participants about the treatment. Six of the participants reported positive remarks about the practices. The following quotations show their satisfaction as they enjoyed the practices in general being involved in a different learning environment through which they could get the chance of expressing their opinions:

> *(16) I had fun in this class, it was useful, we had a good practice… (P5/6th session)*
> *(17) I realized that I am an emphatic person in fact, I liked it. Thanks teacher! (P5/3rd session)*
> *(18) … it was a good and beneficial lesson. (P7/6th lesson)*
> *(19) … we discussed on a very important subject (empathy), it was fun. (P11/3rd session)*

3.3.2. Second Category: Negative Experiences of the Participants / Disadvantages of P4C Practices

According to the findings obtained through the student-reflective journals, a low number of participants reported negative experiences or disadvantages of P4C. Only four themes could be identified within this category which were: *Feeling confused while making a decision; having difficulty in finding reasons for their choices and provide an explanation for their preferences; getting bored when they were addressed various questions and when they disagreed with others*; and *some of their friends being not respectful enough to listen to them.*

Concerning the first theme of the second category, *they felt confused while making a decision;* the participants stated that they struggled with making up their minds and coming through a decision during the inquiries. The following quotations may be helpful to show this finding:

> *(1) Today, we focused on ethics, it was hard and complicated. (P7)*
> *(2) When we were asked many questions, I couldn't make up my mind. (P5)*
> *(3) We studied critical reasoning today and I felt confused. (P3)*

For the second theme, *they had difficulty in finding reasons for their choices and provide an explanation for their preferences,* the participants expressed that:

> *(4) I couldn't explain my ideas, it was challenging. (P10)*
> *(5) Sometimes it was difficult for me to explain my choice, even though I had a preference. (P8)*

Based on the third theme which was *they got bored when they were addressed various questions,* the participants reported that they would feel bored as it was challenging for them to respond to many provoking questions during the discussions. One of the participant's quotations can be given as an example for this finding:

> *(6) When we were asked too many questions, I started to feel bored. (P12)*

The fourth theme of the second category, which was *some of their friends were not respectful enough to listen to them,* illustrates one of the weaknesses of participants especially at the beginning of the treatment. One of the goals of P4C practices was also strengthening discussion skills of the students such as listening to others actively, expressing agreement or disagreement in a polite way, and showing respect towards strange opinions. This expression of the participant shows her disturbance when the others did not listen to her during the discussion:

> *(7) Some friends were disrespectful towards my opinion, but no problem! (P2)*
> *(8) I hope everybody will learn how to discuss! (P1)*

As mentioned before, the participants consider these challenges of the P4C practices as hardships or drawbacks. However, in a community of inquiry, it is something required and favored for the participants to feel confused and indecisive while struggling with dilemmas and conflicting ideas.

3.4. Evaluation Forms

At the end of the implementation process, the participants were instructed about filling out an evaluation form that was designed to elicit their overall opinions about the treatment they received. The form consisted of four open-ended questions to make the participants provide more detailed data. First of all, the participants were requested to respond to the questions in their mother tongue as the main objective was not to assess language proficiency. Moreover, the researcher wanted to enable them to express their opinions freely without dealing with translation from Turkish to English and eliminate the risk of language mistakes that would cause problems in data analysis process while finding out the categories and themes. After they finished filling out the form in Turkish, the participants translated their responses into English with the help of the researcher. In the analysis of the data, the researcher examined both Turkish and English versions of the forms to check accuracy.

The same data analysis process was followed as in the examining of student-reflective journals. The participating researcher took part in the analysis of the evaluation form as another rater as well, and the results were compared before reaching a consensus. As mentioned before, the responses of the participants in the form were not long enough and not so detailed; as a result, and a high consistency was found between the obtained results of the researchers. Only a theme was added to the fourth question and two codes were omitted from the second question after discussion. Table 9 shows the obtained categories and codes from the evaluation forms based on the four questions of the instrument.

Table 9. *Opinions of the participants towards P4C-enhanced speaking practices.*

Question	Theme		*f*
1. Which of your skills do you think P4C practices fostered? (Critical thinking, questioning, analyzing, evaluation, showing respect to different opinions, etc.) Please explain.	Analyzing		5
	Critical thinking		5
	Questioning		4
	Showing respect to different opinions		4
	Discussing		4
	Defending opinions in a polite way		3
	Thinking deeply		2
	Changing minds		2
	Explaining		2
	Distinguishing between concepts		1
	Comparing and contrasting		1
2. Do you think that P4C practices which were implemented in English created a positive or negative impact on your English-speaking skill? Please explain.	Positive	Learning new words	7
		Encouraging to speak English	6
		Pronunciation	1
	Negative	No	12
3. Do you think that P4C practices have some drawbacks for you? Please explain.	Boring		4
	Feeling confused during the discussions		3
	Asking many questions on the same issue		2
	Too much discussion		2
4. Do you think that you could look from different perspectives and think differently from the other people in your daily life after joining P4C sessions? Please explain.	No		6
	Thinking different from others		3
	Showing tolerance to others		3
	Showing respect to others		2
	Looking from other perspectives		2
	Breaking the prejudices		2
	Questioning more than the others		2
	Using empathy more often		2

In the light of the findings demonstrated in Table 9, it can be stated that the participants generally mentioned positive outcomes or experiences concerning P4C-enhanced speaking practices. For the first question, they reported various skills that were fostered through P4C sessions. The highest percentage belongs to critical thinking and analyzing skills, which has vital importance for the participants to realize the effects of the treatment on their thinking and discussing skills. They also think that P4C practices created a positive impact on their questioning and discussion skills. The following quotations can be given as examples reflecting their opinions:

> *(1) In my opinion; P4C improved my critical thinking and questioning skills. (P4)*
> *(2) I learned to think in detail and how to discuss properly. I showed respect to others' opinions… (P2)*
> *(3) I learned to be more polite during the discussions… (P10)*
> *(4) I feel like P4C developed my questioning skill most. At the beginning of the sessions, I had an idea, but sometimes I changed my ideas as the story progressed and my friends expressed their opinions. (P1)*

Responses that were given for the second question were generally positive in terms of the effects of P4C practices on English speaking skills of the participants. While all of the participants stated that there was no negative effect of P4C on their English-speaking skills, most of them reported learning new English words as a positive outcome of the treatment. In addition, half of the participants believed that implementing P4C sessions in English and challenging discussion issues were encouraging for them to express themselves orally and were useful to enhance their English-speaking skills. To illustrate these opinions, the following responses can be considered:

> (5) *I think, it affects English speaking skill positively, because we had to speak English during the sessions. (P8)*
> *(6) Yes, I think it improved my English. I learned new words, tried to speak on interesting topics. (P11)*
> *(7) Too positive, it motivated me, it was nice. Moreover, it was better for us to practice in English during the discussions. (P2)*

When the responses for the third question which addressed the drawbacks of the P4C implementations (if any) for the participants are considered, it can be seen that four of the participants found P4C practices boring for some cases and two of the participants felt confused during the discussions. As the P4C pedagogy was mainly implemented through addressing philosophical questions, it was accepted normal that the participants felt bored and confused to give a response to the questions or provide an explanation for their choices. Unfortunately, teenagers in our age mostly do not want to think over some issues deeply and prefer giving up finding reasons for their ideas after a short time. It was observed by the observer especially at the first sessions until the participants were accustomed to the implementations. The following responses of the participants can be given as examples:

> *(8) When I couldn't find an idea or decide, sometimes I lost my motivation… (P3)*
> *(9) When you discuss a lot, it becomes tiring… (P4)*
> *(10) When we were asked too many questions, I started to feel bored. (P12)*

When it comes to the last question of the evaluation form which addressed whether the participants could observe positive changes in their thinking skills in their lives after P4C practices or not, half of them provided negative responses. While three of the participants stated that they probably thought differently from others, three of them mentioned that they could show more tolerance to others in their daily lives. It should be considered significant that they were aware of the improvements in their thinking and daily life skills. Quotations that support this finding are as follows:

> *(11) Yes, I think my opinion and point of view towards others has changed positively; I can break my prejudices and I can think from different perspectives. (P5)*
> *(12) I started to question the reasons for some events more often. (P1)*
> *(13) In my opinion, the implementation did not affect me in my daily life. (P4)*
> *(14) I can establish empathy more often; I try to use these skills in my daily life. (P8)*

4. DISCUSSION and CONCLUSION

The major purpose of the present study was to investigate the effectiveness of P4C-enhanced speaking practices on English speaking and CT skills of gifted students. Moreover, another objective of the study was to shed light upon the experimental group students' opinions about the P4C treatment they received within the study. In the light of the related literature, it was revealed that there had been no study which focused on examining whether P4C approach can be effective in EFL classes for gifted students or not in the Turkish context. In this sense, implementing a method that had not been employed in this specific field before was considered to yield enlightening implications for the related literature.

4.1. P4C-enhanced English-Speaking Practices in Fostering Critical Thinking Skills of Gifted Students

Before comparing CTTT scores of the groups, the levels of the groups for CT skills are considered necessary to be examined and discussed. While the control group had a 39.36 mean value according to the pre-test scores of the CCTT, the mean value of the experimental group was 41.41. These results show that gifted students involved in the present study had an acceptable level of CT skills when the findings of the study carried out by Ennis et al. (2005) are taken into consideration. In the study of Ennis, the mean value of the 7th, 8th and 9th grade samples' CCTT scores was found to be 38. As CCTT was developed by Ennis and Millman (1985), findings of their study can be accepted as a criterion to make comparison for the present study. Furthermore, in his doctoral study, Akar (2007) found out that mean value of the CCTT scores of the 6th grade samples was 29. Although the samples' grade was not the same with the samples of the present study, it can be suggested that 6th grade regular classroom students had a low level of CT skills and gifted students who participated in this study had a higher level of CT skills. As a consequence, it can be indicated that gifted students had the potential of using their CT skills to some extent before implementing the treatment in the present study.

The results of the experimental procedure in which CCTT was used to test CT skills of the students showed that there was a significant difference between the groups in favor of the experimental group students. As an interpretation of this finding, it is known that fostering CT skills requires a long period of time for individuals and has been considered to be difficult to measure. Departing from this assumption, the treatment was implemented for 10 weeks in total. Furthermore, different P4C sessions which were designed with the aim of fostering different sub-skills of CT on different philosophical concepts can be another factor that affected obtaining positive outcomes in the study. In this way, it is believed that the participants could handle the opportunity of being engaged in various P4C sessions through which they could question the truths they were presented, think deeply, and experience discussion and reasoning. The participants were also observed to improve their discussion skills in time; there was a considerable positive change in their habits of active listening, manner of agreeing or disagreeing, and producing sensible reasons. They were fairly better at discovering relationships, exploring alternatives, interacting with group members and responding to 'Why do you think so/can you explain' questions, and they generally provided their answers without waiting for those questions towards the end of the treatment. Consequently, the treatment was supposed to be effective to create a positive impact on CT skills of the participants. This finding is in line with the results of the research of Lam (2013) who examined the effectiveness of Lipman's P4C Approach in enhancing CT skills of the 28 Chinese Secondary 1 students in Hon-Kong. In this study, it was found that the experimental group participants who received philosophy-based instruction were found to have the capability of conducting debates, good thinking, reasoning, and reflective thinking abilities through the implementation of P4C approach and outperformed the control group in CT and reasoning skills.

Another possible key factor that was supposed play a role on this result of the study was the fact that this study was carried out with the participation of 7th grade gifted students who had a certain level of English proficiency and a potential of using higher-level thinking skills. Since they had been identified as highly talented individuals, from a general perspective, they were expected to have the potential of using CT and English language skills compared to their peers. The participants of the present study displayed a pleasing performance in English during the P4C sessions after they gained enough experience on the method; they thought deeply on philosophical concepts, carried out discussions by providing reasons or explanations for their choices or preferences, and constructed meanings on the focused concepts. If this study was conducted with the participation of younger students from regular classrooms who had not been identified as gifted and who did not have a certain level of English proficiency, it would not be possible to carry out P4C discussions in English language effectively and reach meaningful results. According to the results of a related study carried out by Chamberlain (1993) who aimed to evaluate the effect of Lipman's story –Harry Stottlemeier's Discovery- on the critical thinking skills of 80 fourth and fifth grade gifted students over 12 weeks, it was found that experimental group’s scores were significantly higher than the control group’s scores on the New Jersey Test. It was reported that interaction among students in the philosophy group increased more than that among the literature group and the students could focus on logic, metacognition, and thinking deeply.

Another study that reached similar results with the present research was carried out by Zulkifli and Hashim (2020) who investigated the effectiveness of P4C approach on enhancing CT skills of 61 secondary grade students in moral education classes in Malaysia. The results of their study indicated that experimental group which received P4C treatment displayed a higher performance than the control group did. As a result, it was reported that P4C implementations contributed to the development of CT skills of the participants which supported the results of the present study. The results of the present study also show parallelism with Türksoy’s (2020) study in which positive outcomes of P4C approach were obtained in terms of CT skills of secondary school students in his master thesis. The participants reported positive attitudes towards CT skills after they received P4C treatment which lasted for eight weeks. Even if Türksoy’s study was carried out within science lessons, the tested variable, the participants, and the treatment were similar with the present study. Even if the participants were not gifted students, in a similar study, Pala (2022) also reached supporting findings in line with the present study in terms of the positive effects of P4C instruction on CT skills of 5th grade secondary school students in Türkiye. According to the findings of the study, the experimental group displayed a higher performance in Critical Thinking Skills Scale than the control group and the experimental group participants reported positive outcomes of P4C practices.

4.2. P4C-enhanced English-Speaking Practices in Fostering English Speaking Skills of Gifted Students

The other dependent variable of the presents study was English speaking skills of the participants. Results of the Cambridge Speaking Test for Schools A2 revealed no significant differences between the groups in terms of English-speaking skills of the participants. A possible reason for this finding might be attributed to the fact that experimental group joined P4C sessions regularly every week, and they knew what they were supposed to do during the discussions even though the activities and the discussion topics were not the same with the previous weeks. As the qualitative findings of the study demonstrate, the participants would feel bored and confused when they were addressed philosophical questions in every session. In some cases, they did not want to speak, but they were not forced to express their ideas by the facilitator. As the discussion topics and created dilemmas were challenging and interesting for them, they were encouraged to speak and discuss with their group friends. However, the genre and terminology that they used during the sessions were generally limited to the vocabulary related to discussion, explaining, reasoning or questioning expressions.

Considering the qualitative findings, the participants reported that they did not find the sessions enjoyable or fun to feel eager to speak English during the activities. Most of the teenagers of our age are observed to have changing characteristics such as being impatient, ambitious to reach their aims in a short time without enough effort or having the desire to be in the forefront among people. In this sense, since they were not accustomed to this type of discussion experiences in their classroom environments, having the gifted students be engaged in the P4C practices actively was a challenging task for the researcher as well. On the other hand, the control group joined different activities every week such as playing games, making presentations, writing and acting dialogues, and studying Phonetics. They carried out debates in some classes as well in order to make them experience the discussion process like the experimental group. As is known, teenagers are keen on joining enjoyable activities and more motivated to show participation in these learning environments. In this way, they handled the opportunity of using English in different contexts in different settings which was supposed to contribute to their English-speaking skills. As a result, both of the groups showed a higher performance in post-test scores compared to their pre-test scores, but the difference was not statistically significant, which shows inconsistency with the findings obtained by Rustam et al. (2018) who examined the effectiveness of philosophy-based language teaching approach (PBLT) in fostering English oral fluency of undergraduate students. The findings of their study demonstrated that PBLT approach might provide contributions for the development of English-speaking skills of the students.

On the other hand, contrary to the findings of the present study, Shahini and Riazi (2011) provided evidence on positive outcomes of P4C in terms of English-speaking skill. They aimed to investigate the effect of PBLT techniques on students' speaking, writing, and thinking skills in EFL classrooms in Iran. A significant difference was found between the experimental and control groups' performance in English speaking and writing skills. Their study indicated that PBLT may be an effective tool to enhance students' speaking skills in an ELT context. When compared with these studies, it can be interpreted that the participants, their ages, settings, materials or the duration of the studies may play a role in obtaining different results from the present study.

4.3. Participating Gifted Students' Opinions about P4C-Enhanced English Language Speaking Practices

Qualitative findings of the present study which were gathered through student-reflective journals and evaluation forms filled by the experimental group students demonstrated that the participants mostly reported positive comments on P4C discussions and underlined the contribution of P4C practices to their CT and English-speaking skills. This finding supports the fact that gifted students are considered to be willing and motivated to carry out challenging tasks and activities. In the light of the qualitative findings, it can be suggested that even though the participants did not consider themselves as proficient in speaking English, they tried to speak English during the sessions since the presented stimulus was interesting and the discussion topics were absorbing which encouraged them to share their opinions. The participants also expressed that P4C practices helped them to enhance their CT and higher-level thinking skills such as reasoning, explaining, questioning the credibility of the information, analyzing, or evaluating. It was satisfying that the participants realized the positive effects of P4C on their skills. By conforming the previous findings proposed by Lim (2006) who reported positive attitudes and opinions of gifted 7th grade students towards P4C lessons in Singapore by examining P4C lesson transcriptions, participating students' journal entries, survey responses and interviews; the evidence found in the present research points to the potential of P4C approach in enabling the gifted students to learn thinking critically, reflecting on philosophical issues, and collaborating and constructing meaning.

A low percentage of the participants reported negative comments on P4C practices that they felt bored and confused when they were addressed many compelling and open-ended philosophical questions. As they were not familiar with thinking deeply on philosophical concepts and challenging tasks that enabled them to use their CT skills in their regular classrooms at schools, it was not something surprising for the researcher. This situation can result from the fact that Turkish education system does not embrace implementation of CT skills in the education programs effectively. Even though the Turkish curriculum witnessed a reform from Behaviorist Approach to Constructivist Approach starting in 2004, implementation of CT in the education program has not seemed promising. There may be several reasons for this situation; one of them is that there has been no detailed information for teachers about how to implement CT in different disciplines and the teachers lack practical information to be able to use CT-enhancing classroom activities. The teachers might gain the necessary skills to be critical thinkers through qualified training programs designed by the Turkish MoNE. In addition, the course books have not been designed as a way of enabling the students to use their CT skills. A report by the Education Reform Initiative also suggests that the reform can be considered effective on theoretical basis; however, it has not been designed by taking the specific needs and characteristics of Turkish teachers and students into account (Gürkaynak et al., 2004).

Another factor that can play a role in students' lack of motivation to think deeply and respond to provoking philosophical questions and use their CT skills is teachers' capabilities in adopting CT-enhancing classroom activities. Studies conducted in the Turkish context have focused on to what extent pre-service teachers can use CT skills (Akar, 2007; Gülveren, 2007; Kökdemir, 2003; Kürüm, 2002), and have shown that CT skills of the pre-service teachers were not at a sufficient level. When the teachers do not have an adequate level of CT skills, it cannot be expected that they aim at fostering CT skills of their students and design their teaching procedure in accordance with this aim. In conclusion, the participants of the present study were not accustomed to CT instruction from their regular classrooms and found it difficult to respond to challenging philosophical questions during the treatment.

4.4. Conclusions

The overall goal of the present study was to explore whether P4C approach can be an effective tool in enhancing English speaking and CT skills of gifted students. Findings obtained indicate that P4C has the potential of stimulating gifted students to think deeply on some important issues and concepts and use their higher-level thinking skills when compared to the traditional methods.

In our globalized world, if educators, teachers, or families expect the students and teenagers to learn how to think, listen to the other people actively and respectfully, realize alternative options in their lives, seek reasons and solutions for the problems they face, show respect to different or interesting views, develop arguments and explore underlying concepts, and eventually become critical thinkers, fundamental changes and revisions have to be made in education programs as well. P4C might create a response to the needs and interests of students by allowing them to ask the questions they develop, satisfying their curiosity in exploring the nature of the world, and providing a room for them to develop meaning.

In order to implement this approach in our classrooms in an effective way, teachers need to get training on how to ask philosophical questions, creating dilemmas, choosing appropriate stimulus, and preparing P4C lessons consistent with the objectives of the program. Learning to be a facilitator and integrating P4C approach into our disciplines is not a simple task and requires hard work and time. Based on the observations in the light of the findings, the P4C approach has been considered to take the attention of the students to a very different or undetected point even in daily life matters. As people are already accustomed to many preconceptions in their daily lives, they do not even think of the possibility of questioning or inquiring about the truths or information they are presented. In our age, where accessing knowledge has become easy and quick, students mostly accept the truths or information they encounter as true without searching for its credibility and source. In this sense, the P4C approach may be beneficial in making our students gain 21st century skills such as questioning, searching reliability of the information, distinguishing between fact and opinion, and thinking critically.

In the light of the obtained findings of the study and observations and experiences of the researcher, some suggestions can also be proposed. This quasi-experimental study was carried out by the participation of 23 seventh grade gifted students and the treatment lasted for 10 weeks of 2 classes per week. Future studies that will last for longer periods to examine the effects of P4C approach can be carried out. This study involved a control and experimental group to reach more reliable results; however, the number of the samples was limited within the circumstances of the term in which the study was conducted. A higher number of samples can be involved in future studies. The most challenging aspect of conducting the present study for the researcher was to find a CT test appropriate for the age and level of the participants. In the Turkish context, no appropriate CT test was found for the sample of the present study. Foreign CT tests carry the risk of language comprehension and cultural differences for other nations, thus Turkish researchers or academicians who want to contribute to the literature can develop CT tests suitable for different levels and ages. In this study, the dependent variables were English speaking and CT skills of the gifted students. In further studies, the impact of P4C approach on different variables can be investigated. In addition, this research was conducted with the participation of 7th grade gifted students and the results cannot be generalized to other students; further studies which will involve regular classroom students or different grade gifted students may yield beneficial implications to the literature and make it possible to compare the performance of gifted students and their peers in regular classrooms. In terms of the integration of P4C approach into regular classrooms, P4C can be implemented as a stand-alone activity within the program or can be fully integrated into the curriculum. As an independent course, teachers migh have one-hour practice of P4C weekly to support curriculum-connected enquiry. Even in this way, it would be possible to engage the students with listening actively, thinking philosophically and gaining effective discussion skills. As a second alternative, the curriculum needs to be designed completely as a way to “plan P4C around the conceptual content in schemes of work” (Sapere P4C, 2023, p. 1). It may therefore be possible to create philosophical dialogues with the students in every discipline and in every subject; for instance, in a math class, a teacher might prompt the students to deeper thinking by asking “Was mathematics invented or discovered?”; in a language class, teachers can ask “Could there be a perfect translation?” to make the students look from different perspectives with a high level of mindfulness. As a conclusion, researchers in the field of program development can focus on integrating P4C approach into different disciplines and Turkish education program starting from an early age as well as producing new materials and course books which include open-ended questions, compelling activities or stories that have philosophically rich content.

4.5. Suggestions for Practitioners

P4C does not mean teaching philosophy; teachers and practitioners from many disciplines can implement it in their own classrooms as a method that encourages the students to question the truth, construct meaning, gain appropriate discussion skills, and become critical thinkers. Teachers and practitioners should also get training in how to be a facilitator; otherwise, they cannot implement P4C approach in an effective and appropriate way. Institutions or universities can organize in-service training for volunteer teachers or educators who want to integrate the P4C approach into their teaching procedures. Especially kindergarten teachers should get training in P4C and be encouraged to employ this approach in their classrooms based on the assumption that philosophizing should start at an early age to become a life-long habit. Furthermore, with a purpose to observe the effects of P4C implementation on cognitive abilities of the students such as higher-level thinking, discussion and CT skills require a long period of time considering the related studies which include a long process of intervention (Colom et al., 2018; Niklasson et al., 1996). When it is practiced on a regular basis, it is highly possible that such a practice will enable the students to perceive long-term significant gains in communication and discussion skills and self-confidence as well.

4.6. Suggestions for Further Study

The treatment of the study lasted 10 weeks within the study; however, more reliable results can be achieved in longer studies since improving higher-level thinking skills and creating a change on CT and English-speaking skills of students require a long period of time to realize.

Additionally, this study was carried out in a specific context as the participants were 7th grade gifted students studying at a SAC and having a certain English proficiency to carry out the philosophical discussions; thus, for future research, studies involving a higher number of students from different grades can be suggested.

Acknowledgments

This paper is produced from the part of the first author's master's dissertation prepared under the supervision of the second author. This study has been supported by Pamukkale University Scientific Research Projects Coordination Unit under grant number 2022EĞBE004.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). **Ethics Committee Number**: Pamukkale University/ Social Sciences and Humanities Research and Publication Ethics Committee, 19.11.2021-E.131257.

Authorship Contribution Statement

Feride Acar: Methodology, Data Collection, Data Analysis, and Writing the Manuscript. **Recep Şahin Arslan**: Design, Supervision, and Writing the Manuscript.

Orcid

Feride Acar https://orcid.org/0000-0001-6215-5700
Recep Şahin Arslan https://orcid.org/0000-0002-2475-5884

REFERENCES

Afrizal, M. (2015). Improving English speaking ability through classroom discussion. *Lentera, 15*(14), 1-9. https://doi.org/10.33603/rill.v2i3.2127

Akar, Ü. (2007). *Öğretmen adaylarının bilimsel süreç becerileri ve eleştirel düşünme beceri düzeyleri arasındaki ilişki [The relationship between student teachers' scientific proces skills and critical thinking skills]* [Unpublished master's thesis]. Kocatepe University.

Anwar, F.Z. (2015). Enhancing students' speaking skill through Gallery Walk technique. *Register Journal*, *5*(1), 253-568. https://doi.org/10.18326/rgt.v8i2.384

Arslan, R.Ş., & Yıldız, N. (2012). Enhancing critical thinking at the tertiary level through a literature based critical thinking program. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 21*(2), 19-36. https://dergipark.org.tr/tr/pub/cusosbil/issue/4390/60352

Arter, J.A., & Salmon, J.R. (1987). *Assessing higher order thinking skills. A consumer's guide.* Oregon: Northwest Regional Educational Laboratory Test Center.

Báez, C.P. (2009). Critical thinking in the EFL classroom: The search for a pedagogical alternative to improve English learning. *Íkala Revista de Lenguaje y Cultura*, *9*(15), 45-80. https://www.researchgate.net/publication/277834572

Büyüköztürk, Ş., Çokluk, Ö., & Köklü, N. (2014). *Sosyal bilimler için istatistik [Statistics for the social sciences] (15th ed.).* Pegem Akademik.

Chamberlain, M.A. (1993). *Philosophy for children program and the development of critical thinking of gifted elementary students* [Unpublished doctoral dissertation]. University of Kentucky.

Chan, D.W. (2015). Education for the gifted and talented. *International Encyclopaedia of the Social & Behavioural Sciences, 2nd edition*, *7*, 158-164. http://dx.doi.org/10.1016/B978-0-08-097086-8.92137-8

Cinquino, D. (1981). An evaluation of a philosophy program with 5th and 6th grades academically talented students. *Thinking the Journal of Philosophy for Children, 2*(3 & 4), 79-83. https://doi.org/10.5840/thinking19812324

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education (6th ed.)*. Routledge.

Colom, R., Moriyón, F.G., Magro, C., & Morilla, E. (2018). The long-term impact of Philosophy for Children: A Longitudinal Study (Preliminary Results). *Analytic Teaching and Philosophical Praxis*, *35*(1), 50-56. https://journal.viterbo.edu/index.php/atpp/article/view/112

Dabbagh, A., & Noshadi, M. (2016). Philosophy-based language teaching approach on the horizon: A revolutionary pathway to put Applied ELT into practice. *Journal of Language Teaching and Research*, *7*(5), 1022-1028. https://doi.org/10.17507/jltr.0705.25

Davidson, B.W., & Dunham, R.A. (1997). Assessing EFL student progress in critical thinking with the Ennis-Weir Critical Thinking Essay Test! *Jalt Journal*, *19*(1), 43-57. https://files.eric.ed.gov/fulltext/ED403302.pdf

Demir, E., Saatcioğlu, Ö., & İmrol, F. (2016). Examination of educational researches published in international journals in terms of normality assumptions. *Current Research in Education*, *2*(3), 130-148. https://atif.sobiad.com/index.jsp?modul=makale-goruntule&id=AWY_Ip3eHDbCZb_mQzvn

Ennis, R.H., & Millman, J. (1985). *Cornell critical thinking test, level X (3rd ed.)*. The US: Midwest Publications.

Ennis, R.H. (1987). A taxonomy of critical thinking dispositions and abilities in J.B. Baron and R.J. Sternberg (Eds.). *Teaching thinking skills: Theory and practice*. W.H. Freeman.

Ennis, R.H., & Millman, J., & Tomko, T.N. (2005). *Cornell critical thinking test, level X (5th ed.)*. Seaside, CA: The Critical Thinking Company.

Facione, P. (2015). Critical thinking: What it is and why it counts. *Insight Assessment*, *1*, 1–30. https://www.student.uwa.edu.au/__data/assets/pdf_file/0003/1922502/Critical-Thinking-What-it-is-and-why-it-counts.pdf

Gülveren, H. (2007). *Eğitim fakültesi öğrencilerinin eleştirel düşünme becerileri ve bu becerileri etkileyen eleştirel düşünme* faktörleri *[Critical thinking skills of education faculty students and factors influencing critical thinking skills]* [Unpublished doctoral dissertation]. Dokuz Eylül University.

Gürkaynak, İ, Üstel, F., & Gülgöz, S. (2008). *Eleştirel düşünme: Erg raporları [Critical thinking: Erg reports]*. Sabancı Üniversitesi İstanbul Politikalar Merkezi, İstanbul. https://www.egitimreformugirisimi.org/wp-content/uploads/2017/03/Elestireldusunme_0.pdf

Haryanti, D.U., Indah, R.N., & Wahyuni, S. (2021). Enhancing oral proficiency using Three Steps Interview technique for eleventh graders. *JOLLT Journal of Languages and Language Teaching*, *9*(1), 61-68. https://doi.org/10.33394/jollt.v%vi%i.3271

Holsti, O.R. (1968). Content analysis. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology (2nd ed.)*, 596-692. Amerind Publishing Co.

Hughes, J. (2014). *Critical thinking in the language classroom.* Italy: Eli Publishing.

Kökdemir, D. (2003). *Belirsizlik durumlarında karar verme ve problem çözme [Decision making and problem solving in situations of uncertainty]* [Unpublished doctoral dissertation]. Ankara University.

Kuhn, D. (2015). Thinking together and alone. *Educational Researcher*. *44*(1), 46–53. https://doi.org/10.3102/0013189x15569530

Kürüm, D. (2002). *Öğretmen adaylarının eleştirel düşünme gücü [Critical thinking abilities of teacher trainers]* [Unpublished master's thesis]. Anadolu University.

Lam, M. (2013). An empirical study of the effectiveness of Lipman's Philosophy for Children Programme on promoting children's critical thinking in Hong Kong, China. *Childhood, Philosophy and Open Society*, 67-120. https://doi.org/10.1007/978-981-4451-06-2_4

Lam, C. (2020). Integrating philosophy into English curriculum: The development of thinking and language competence. *The Journal of Educational Research, 112*(6), 700-709. https://doi.org/10.1080/00220671.2019.1696273

Lim, T.K. (2006). Gifted students in a community of inquiry. *Journal of Educational Policy, 3*(2), 67-80. https://www.proquest.com/openview/4442a444684609172622831f371e50d2/1?pq-origsite=gscholar&cbl=946348

Lipman, M. (1982). *Harry Stottlemeier's discovery*. United States of America: Institute for the Advancement of Philosophy for Children, University Press of America Inc.

Lipman, M. (2003). *Thinking in education.* Cambridge University Press.

Marland, S. P. (1972). Education of the gifted and talented. *Report to the Congress of the United States by the Commissioner of Education.* Washington, DC: US Government Printing Office, 72-502. https://www.valdosta.edu/colleges/education/humanservices/document%20/marland-report.pdf

McGregor, D. (2007). *Developing thinking; developing learning.* England: Open University Press.

MoNE, (2009). *İlköğretim Türkçe dersi öğretim programı ve kılavuzu*. Milli Eğitim Bakanlığı, Ankara. Retrieved from http://web.deu.edu.tr/ilyas/ftp/turkce2009.pdf on 20.01.2021

MoNE, (2021). *Resource Rooms Directive*. http://meb.gov.tr

Niklasson, J., Ohlsson, R., & Ringborg, M. (1996). Evaluating Philosophy for Children. *Thinking, 12*(4), 17-23.

Oradee, T. (2012). Developing speaking skills using three communicative activities (Discussion, problem-solving, and role playing). *International Journal of Social Science and Humanity, 2*(6), 533-535. https://doi.org/10.7763/IJSSH.2012.V2.164

Pala, F. (2022). The effect of Philosophy Education for Children (P4C) on students' conceptual achievement and critical thinking skills: A mixed method research. *Education Quarterly Reviews, 5*(3), 27-41. https://doi.org/10.31014/aior.1993.05.03.522

Patton, M.Q. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA: Sage.

Philosophy 592 (Pre-College Philosophy) Class Members (2013). *Philosophy for children: Lesson plans.* Michael Davidson (Ed), (1-74). The USA: The University of North Carolina. https://philosophy.unc.edu/wp-content/uploads/sites/122/2013/10/Philosophy-for-Children-Lesson-Plans.pdf

Rashtchi, M., & Khoshnevisan, B. (2020). Lessons from critical thinking: How to promote thinking skills in EFL writing classes. *European Journal of Foreign Language Teaching, 5*(1), 34-47. https://doi.org/10.46827/ejfl.v5i1.3153

Rustam, U., Anwar, A., & Amzah, A. (2018). Implementing Philosophy-based language teaching approach to improve students' speaking skill. *Eternal, 4*(1), 127-145. https://doi.org/10.24252/Eternal.V41.2018.A10

Sapere P4C. (2023, May 12). *P4C and the curriculum.* https://www.sapere.org.uk/why-sapere-p4c/p4c-and-the-curriculum/

Shahini, G.H., & Riazi, A.M. (2011). A PBLT approach to teaching ESL speaking, writing, and thinking skills. *ELT Journal, 65*(2), 170- 179. https://doi.org/10.1093/elt/ccq045

Sutcliffe, R. (2004). Philosophy for children – a gift from the gods? *Gifted Education International, 19*, 5-12. https://doi.org/10.1177/026142940401900103

Tosuncuoğlu, I. (2018). Place of critical thinking in EFL. *International Journal of Higher Education, 7*(4), 26-32. https://doi.org/10.5430/ijhe.v7n4p26

Türksoy, N. (2020). *Çocuklar için felsefe (P4C) eğitiminin Ortaokul öğrencilerinin bilimsel sorgulamaya yönelik görüşlerine ve eleştirel düşünme becerilerinin gelişimine katkısı: Bir karma yöntem araştırması [The contribution of philosophy (P4C) education for secndary school students to scientific inquiry views and development of critical thinking* *skills: A mixed method research]* [Unpublished master's thesis]. Alanya Alaattin Keykubat University.

Wartenberg, T.E. (2009). *Big ideas for little kids: Teaching philosophy through children`s literature*. Lanham: The Rowman & Littlefield Publishing Group.

Yang, Y., Newby, T.J., & Bill, R.L. (2005). Using Socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments. *American Journal of Distance Education, 19*(3), 163-181. https://doi.org/10.1207/s15389286ajde1903_4

Yang, Y.C., & Gamble, J. (2013). Effective and practical critical thinking-enhanced EFL instruction. *ELT Journal*, *67*(4), 398-412. https://doi.org/10.1093/elt/cct038

Zahrani, B.S., & Elyas, T. (2017). The implementation of critical thinking in a Saudi EFL context: Challenges and opportunities. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *1*(2), 133-142. https://doi.org/10.21093/ijeltal.v1i2.21

Zulkifli, H., & Hashim, R. (2020). Philosophy for Children (P4C) in improving critical thinking in a secondary moral education class. *International Journal of Learning, Teaching and Educational Research*, *19*(2), 29-45. https://doi.org/10.26803/ijlter.19.2.3

APPENDICES

Appendix 1. *Process of implementation with the experimental group.*

Weeks	Implementation	Lesson Procedure
1st Week	Critical Thinking Test (Pre-test) English Speaking Test (Pre-test) 2 hours (40'+40')	
2nd Week	**Ethics** - Identifying the principles for the community of inquiry *What makes an action right or wrong?* P4C Practice – 2 hours	Stimulus: Heinz's Dilemma (Socratic Story) Group discussion on ethics (in Turkish) Should Heinz steal the medicine for his ill wife? Does stealing the medicine for his ill wife make him an immoral person?
3rd Week	**Freedom / Social Identity** *How can people become a society?* *What are the limits of freedom?* P4C Practice – 2 hours	Stimulus: Robinson Cruise (A short movie) Constructing a new society on a deserted island What is the name/rules/flag of the new society? How do they meet a common decision? Describing their choices/preferences with reasons
4th Week	**Being aware of empathy, breaking down the prejudices** *Do you have prejudices?* *How emphatic are you?* P4C Practice – 2 hours	Stimulus: A short movie "The Present" about disabled people Did you think that the boy was a bad character at the beginning of the video? Did you change your mind about the boy after watching the end of the video? Do you think that you have prejudices in your lives? Joining a questionnaire: *How empathic are you?*
5th Week	**Epistemology / Critical Reasoning** *How do you know what you know?* *What counts as a valid reason?* P4C Practice – 2 hours	Stimulus: "What's Your Reason" game Writing three reasons for the truth they write Writing three reasons for the falsehood they write Group discussion on whether the reasons are valid or not Attention Test: Do you trust on your senses
6th Week	**Altruism/Environmental Ethics** *Is it possible to love without any expectations?* *Does expecting something from somebody mean selfishness?* P4C Practice – 2 hours	Stimulus: The story of "The Giving Tree" Discussion on pure love, happiness and environmental ethics When you give something to someone, do you expect something in return? Do you think that there is someone who loves without expecting anything from us?
7th Week	**Ethical Responsibility** *Am I always responsible for my actions?* *Can we change our nature?* P4C Practice – 2 hours	Stimulus: The story "The Frog and the Scorpion" Discussion on personal and ethical responsibility. If something is in our nature, can we control it?
8th Week	**Morality** *How do you know what is right and wrong?* *What would you do if you were invisible?* P4C Practice – 2 hours	Stimulus: The story "The Ring of Gyges" Discussion on social morality Would you do good things if there were no authority? Is there a parallelism between what you do and what you expect from others?
9th Week	**Bravery/Cowardice/Fear** *Should you always do what your community, family or friends ask of you?* *What is bravery?* *If you are fearless, are you braver?* *P4C Practice – 2 hours*	Stimulus: The story of "Three Brothers" who join the army Which boy is braver? Can you neglect your duty and still do good?
10th Week	**Justice and Equality**	Stimulus: The story of "Winnie the Pooh's Cake"

	How can we share something? *Should we consider justice or equality?* P4C Practice – 2 hours	Discussion on how they should share the cake Should they consider equality, justice or needs?
11th Week	**Ethical Responsibility** *Should we help or do a favour for everyone, under every circumstance?* *P4C Practice – 2 hours*	Stimulus: The story of “New Trainers” Would you give up something important for you to help others? Who is doing wrong in terms of ethical responsibility?
12th Week	English Speaking Test (Post-test) Critical Thinking Test (Post-test)	English Speaking Test (Post-test) Critical Thinking Test (Post-test)

Appendix 2. *Process of implementation with the control group.*

Weeks	Implementation	Lesson procedure
1st Week	Critical Thinking Test (Pre-test) English Speaking Test (Pre-test) 2 hours (40'+40')	Critical Thinking Test (Pre-test) English Speaking Test (Pre-test) 2 hours (40'+40')
2nd Week	**Getting to know each other better** Regular SAC speaking class program 2 hours	"Find Someone Who" Activity Implementing a questionnaire with the classmates (data collection) Analysing the data Sharing the results with the group by using percentages/rate
3rd Week	**Studying on basic Phonetics** Regular SAC speaking class program 2 hours	Focusing on different accents in English Introducing the International Phonetic Alphabet Studying on basic principles in English pronunciation
4th Week	**Name three people / places /things** Regular SAC speaking class program 2 hours	A writing and speaking activity Writing three people/places or things on a ready worksheet Sharing the answers
5th Week	**Carrying out a debate** Regular SAC speaking class program 2 hours	Debate on "the pros and cons of the technology" Grouping the class into two groups Allowing 20 minutes for preparation Carrying out the debate Peer-assessment
6th Week	**Taboo Game** *Regular SAC speaking class program* *2 hours*	Grouping the class into three/four groups Giving ready taboo cards to the speakers Telling the given words without using the forbidden words and making the group friend find the secret word to get points
7th Week	**Making Presentations** *Regular SAC speaking class program* *2 hours*	Making presentations about different cultures to the class (4 students who got ready before the class) Question & answer
8th Week	**Making Presentations** *Regular SAC speaking class program* *2 hours*	Making presentations about different cultures to the class (4 students who got ready before the class) Question & answer (See overleaf)
9th Week	**Carrying out a debate** *Regular SAC speaking class program* *2 hours*	Debate on "Distance or face-to-face education?" Grouping the class into two groups Allowing 20 minutes for preparation Carrying out the debate Peer-assessment
10th Week	**Writing dialogues on a silent movie and acting it out** *Regular SAC speaking class program* *2 hours*	Watching part of the movie- 10 minutes "Life School" without any sound Writing dialogues for the characters according to the context Getting ready to act it out Acting out the movie
11th Week	**Studying on basic Phonetics** *Regular SAC speaking class program* *2 hours*	Studying on Short Vowel Words Studying on Blends and Digraphs Practice on a ready worksheet
12th Week	English Speaking Test (Post-test) Critical Thinking Test (Post-test)	English Speaking Test (Post-test) Critical Thinking Test (Post-test)

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 376–394

https://doi.org/10.21449/ijate.1155679

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

Classroom assessment that tailor instruction and direct learning: A validation study

Wai Kei Chan [1], Li Zhang [2], Emily Pey-Tee Oon [3*]

[1]University of Macau, Faculty of Education, Macau, China
[2]University of Macau, Faculty of Education, Macau, China
[3]University of Macau, Faculty of Education, Macau, China

ARTICLE HISTORY

Received: Aug. 04, 2022
Revised: Apr. 26, 2023
Accepted: June. 21, 2023

Keywords:
Classroom assessment,
Constructing measures,
Tailor instruction,
Direct learning.

Abstract: We report the validity of a test instrument that assesses the arithmetic ability of primary students by (a) describing the theoretical model of arithmetic ability assessment using Wilson's (2004) four building blocks of constructing measures and (b) providing empirical evidence for the validation study. The instrument consists of 21 multiple-choice questions that hierarchically evaluate arithmetic intended learning outcomes (ILOs) on arithmetic ability, hierarchically, based on Bloom's cognitive taxonomy for 138 primary three grade students. The theoretical model describes students' arithmetic ability on three distinct levels: solid, developing, and basic. At each level, the model describes the characteristics of the tasks that the students can answer correctly. The analysis shows that the difficulty of the items followed the expected order in the theoretical construct map, where the difficulty of each designed item aligned with the cognitive level of the student, the item difficulty distribution aligned with the structure of the person construct map, and word problems required higher cognitive abilities than the calculation problems did. The findings, however, pointed out that more difficult items can be added to better differentiate students with different ability levels, and an item should be revised to enhance the reliability and validity of the research. We conclude that the conceptualizations of such formative assessments provide meaningful information for teachers to support learning and tailoring instruction.

1. INTRODUCTION

The central purpose of classroom assessment is to provide feedback to improve student learning and teachers' pedagogies (Black & Wiliam, 2010; Dixson & Worrell, 2016; Shepard, 2006; Stiggins, 1994; Wiggins, 1998). However, grading continues to dominate pedagogical practices, even in the context of formative assessment, diluting its effectiveness. Consequently, crucial questions concerning the essential features for collecting, formatting, and acting on evidence of learning through formative assessment remain unanswered.

The situation has not changed much today, as classroom formative assessment practices continue to be counterproductive and incoherently disconnected from each other and from high-stake accountability assessments (Wilson, 2004; NRC, 2006; Gorin & Mislevy, 2013). Contrary

[*] CONTACT: Emily Pey-Tee Oon ✉ peyteeoon@um.edu.mo University of Macau, Faculty of Education, Macau, China

e-ISSN: 2148-7456 /© **IJATE** 2023

to expectations based on the central role of feedback in effective classroom practices (Hattie, 2008; Duckor & Holmberg, 2017), teachers often do not review student responses to formative assessment questions or reflect upon the content measured, such as whether items are ambiguously worded or a scoring criterion is unclear (Black & Wiliam, 1998; Guskey, 2003; Popham, 2009, 2010).

When student scores and grades are the primary focus of assessment and the full value of formative feedback is not obtained, outcome quality is inevitably undercut, even though formative practices have repeatedly been shown to be the most effective means to improve student achievement (Black & Wiliam, 2003; Hattie, 2008). Instructionally-relevant feedback is essential to providing the leverage needed for advancing toward the desired learning outcome (Brookhart et al., 2010).

The purpose of formative assessment is to help teachers identify difficulties obscuring students' conceptual understanding, charting a path forward along a learning progression (Bell & Cowie, 2001; Black et al., 2011; Cardinet, 1989). Formative practices encompass a broad range of qualitative and quantitative assessments of as and learning in the classroom, and are not limited to feedback from formally scored assessments (Baird, et al., 2017; Duckor & Holmberg, 2017; Fisher, 2013). However, without an essential feedback mechanism, formative practices of any kind will fail to produce the desired effect. Properly conceived, designed, and implemented, formative assessment is integrated with instruction, and should be a key tool for monitoring learning progressions and supporting the attainment of learning outcomes (Clark, 2012; Gorin & Mislevy, 2013; Popham, 2009, 2010).

Wilson (2004) proposed constructing measurement instruments using four building blocks: construct map, item design, outcome space, and measurement model, to properly conceive, design, and implement formative practices in classrooms. The blocks ascertained the validity of the formative assessment information produced from the test items. Building blocks are a reference that aid in the assessment design cycle. Each building block focuses on one of the four principles: developmental theory perspective (construct map), a match between instruction and assessment (item design), management by teachers (outcome space), and evidence of high quality (measurement model). When this cycle is reiterative, block coherence is improved because each block's information can optimize other blocks. This model tests construct consistency for objective proof of knowledge, skills, and attitude measurements.

The Rasch measurement model is applied for shortcomings that plagued the classical test theory that it is sample dependent and item dependent (Embretson & Reise, 2000; French, 2001; Hambleton et al., 1991; Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985) that limit the generalizability of research findings (Wright & Master, 1982). Furthermore, Rasch measurement model is applied in the current study to ascertain validity of the test. A number of related studies have reported on scale validity based on content validity, as judged by experts in the relevant domain in science education (e.g. Adillah et al., 2022; Beck, 2020; Hidayati et al., 2019; Luque-Vara et al., 2020; Nasir et al., 2022; Wole et al., 2021). Content validity based solely on professional judgment is insufficient to establish validity (Messick, 1981; 1989). At the most, testing validity merely on content validity is insufficient (Lee & Fisher, 2005) because validity refers to "*the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"* (Messick, 1989, *p*. 13). As quoted by Fisher (1997), "*The conventional focus on content validity has misled us about what is important in educational measurement"*.

Arithmetic skills that involve basic operations of addition, subtraction, multiplication, and division are the foundation for advanced mathematical concepts (Björklund, 2021; Hong Kong Education Bureau, 2000, 2018; Parviainen, 2019; Sievert et al., 2021; Vlassis et al., 2022) such as algebra and geometry (Vlassis et al., 2022). It helps students develop logical thinking and problem-solving abilities (Björklund, 2021). Thus, arithmetic is an essential component of the primary mathematics curriculum (Engvall et al., 2020). Baroody and Dowker (2003), Dowker (2005), Geary (1993), Goldman and Hasselbring (1997), Hiebert and Lefevre (1986), and Kilpatrick et al. (2001) carried out research on how students attain arithmetic proficiency. However, they did not address the validity issue of the instrument used to measure students' arithmetic abilities.

In the present study, we aimed to provide evidence of the validity of a classroom assessment evaluating the arithmetic ability of primary students based on Wilson 's (2004) four building blocks. We then provide empirical evidence for the validation of the hypothesized model based on Rasch's (1960) measurement model.

1.1. Theoretical Framework

1.1.1. *Construct mapping*

The first building block was a constructed map. This is a diagrammatic representation of the construct specifications. It operationalizes constructs in the successive stages of understanding or abilities. It shows how students' understanding evolves, and how their responses to items might change or develop. The construct assessed in this research was arithmetic ability, defined as accurately solving addition, subtraction, multiplication, and division problems mentally. This includes choosing the correct arithmetic operation and calculating the solution (Millians, 2011).

Students' arithmetic ability can be represented using a person-construct map (Wilson, 2004). Theoretically, it is suggested that students' arithmetic abilities are developed in stages where they start from the basic skills of solving addition and subtraction problems, and then move on to more complex operations of multiplication and division (Hong Kong Education Bureau, 2018). According to the Cognitive Development Theory, students initially use concrete objects and counting strategies to solve arithmetic problems. As they progress through these stages, they develop more abstract and efficient strategies for solving problems. The construct map developed in this study was considered within these contexts.

The person construct map assumes that arithmetic ability is a unidimensional latent variable that extends from low to high (Table 1). Students were categorized into three groups according to their levels of arithmetic ability. At the basic level, students with low ability can remember and understand algorithms (four operations) to solve simple operation problems. At a developing level, mid-range students can analyse information and apply the underlying algorithms to solve less challenging or less complex operational problems. At a solid level, high-ability students can evaluate the information given in words, work out the best operation to solve complex problems in words, and justify their solutions (Table 1).

Table 1. *Person construct map of predicted arithmetic ability levels of primary school students.*

Level	Respondents	Description
3	Solid	➢ Students can evaluate the information given in words and work out addition and subtraction operations to solve mixed operation problems. They can correctly identify a mixed operation to solve two-step problems when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can evaluate the information given in words and work out the best multiplication operation to solve complex problems and justify their solutions. They can accurately identify multiplication to solve and explain two-step problems by multiplying a one-digit number by a one-digit number. (Multiplication construct) ➢ Students can evaluate the information given in words and work out the best division operation to solve complex problems and justify their solutions. They can correctly identify division to solve one-step and two-step problems of the quotient of a one-digit number and explain the reasons. (Division construct)
2	Developing	➢ Students can analyse the question and apply the best strategy to solve addition and subtraction. They can precisely calculate addition with carrying (carrying once and carrying twice) and subtraction with borrowing (borrowing once, borrowing twice and borrowing twice for ones, when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can analyse the information given in numbers and apply the necessary addition and subtraction to solve conceptually less challenging mixed operation problems. Individually, they can calculate mixed operation (addition and subtraction) with and without carrying and borrowing (when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can analyse the information given in numbers and apply the best strategy to solve conceptually less challenging multiplication operation problems. They understand multiplication and can correctly solve problems by multiplying a one-digit number by a one-digit number. (Multiplication construct) ➢ Students can analyse the information given in numbers and apply the best strategy to solve conceptually less challenging division operation problems. They can correctly calculate division with a remainder and the quotient of a one-digit number. (Division construct)
1	Basic	➢ Students can remember and understand addition and subtraction algorithms to solve simple operation problems. They can accurately calculate addition and subtraction with no carrying and borrowing (when the answer is smaller than 1,000). (Mixed addition and subtraction operation construct) ➢ Students can memorise and understand the multiplication table and use what they recall to solve simple operation problems. Precisely, they can calculate the multiplication of a one-digit number by a one-digit number and use it to solve one-step problems without explanation in words. (Multiplication construct) ➢ Students can memorise and understand division algorithms to solve simple operation problems. They can calculate division without a remainder and the quotient of a one-digit number. (Division construct)

1.1.2. *Item design*

The item design encapsulates the types of items used to provide evidence of students' knowledge and understanding embodied in the theoretical construct map. It guides how the learning outcomes will be measured and aligns the curriculum and assessment using standard conditions. This enabled assessment of each level defined in the construct map (Table 1). A total of 21 multiple choice question (MCQ) items were designed based on the three cognitive knowledge levels (Basic, Developing, and Solid) from the person construct map (Table 2). MCQ tests can save time and reduce grading costs (Alderson, 1990; Liu et al., 2008), can test multiple knowledge domains within the same test (Çataloğlu & Robinett, 2002), enabling more objective grading to ensure the fairness of the test and facilitate item and test analysis, which can improve teaching and students' learning (Gurel et al., 2015). Therefore, we used MCQs to measure students' attainment of learning outcomes.

Table 2. *Item design.*

Level	Respondents Level	Items	Operation	Problem type	Option type	Bloom's cognitive level
3	Solid	Q13, Q15	Division	Word	Algorithm	Evaluate
		Q4, Q9	Multiplication	Word	Algorithm	Evaluate
		Q21	Mixed	Word	Algorithm	Evaluate
2	Developing	Q5, Q10	Division	Word	Number	Analyse
		Q6, Q12	Multiplication	Word	Number	Analyse
		Q2	Division	Calculation	Number	Apply
		Q19	Mixed	Calculation	Number	Apply
		Q7, Q11, Q16	Subtraction	Calculation	Number	Apply
		Q3, Q17, Q20	Addition	Calculation	Number	Apply
1	Basic	Q1	Division	Calculation	Number	Understand
		Q14Q8	Subtraction	Calculation	Number	Understand
		Q18	Addition	Calculation	Number	Understand
			Multiplication	Calculation	Number	Remember

1.1.3. *Outcome space*

The outcome space encapsulates different student response levels for items correlated with the construct level. It guides the assessment of students' responses to items relative to the construct map. Specifically, it can be used as a scoring guide to ensure that student answers align with the constructed map. Teachers assigned scores to an item designed for a particular knowledge level based on the construct map in the outcome space. When the item design is completed, teachers then decide which factors may affect the item response, and classify and score these factors to ensure meaningful student responses. In this study, MCQs were used to design items that were scored dichotomously (Incorrect = 0 and Correct = 1; Wilson, 2004).

1.1.4. *Measurement model*

The measurement model is the framework by which assessors equate student scores on items from specific construct levels and apply the scored responses to the constructs. The assumption is that student scores on individual items align with the knowledge construct map. The resulting model is a measurement or interpretation model (Wilson, 2004). This helps teachers understand and evaluate student responses to the items. The Rasch measurement model was used in the current study. The model transforms the scores into the locations of items in the construct map. It is an objective measurement suitable for various random, hierarchical, and classified data analyses (Linacre, 2000). It was thus used in this study to relate data to assessment targets and construct maps. The output is a Wright map that displays student performance on elements of the construct map and enables comparisons between students. In addition, it places students and items on the same scale (with an arbitrary scale representing a student's chances of a positive response at that position). This, in turn, documents the measurement system and assesses the construct validity (Wilson, 2004, p. 156-157).

1.1.5. *Validation study for the hypothesized model*

We sought to evaluate whether the theoretical person constructs a map of arithmetic ability following the four building blocks (Wilson, 2004) aligns with the statistical results of the Rasch measurement model. Specifically, we examined whether the following were true.

1. The difficulty of each designed item aligns with the cognitive ability of students in the following order: Level 1 ability is less than Level 2 ability, which is less than Level 3 ability (Table 1).
2. Basic students can solve Level 1 items, developing students can solve Level 1 and 2 items, and solid students can solve items from all levels.
3. Arithmetic calculation problems are more accessible to students' ability levels than arithmetic word problems (at Level 3).

2. METHOD

2.1. Participants

A sample of 138 Primary 3 students from a single-gender school in Macao participated in this study. The students were from four different classes taught by three teachers. One teacher taught two classes and the other two teachers taught one class each. The students had completed their Primary 2 mathematics course and had just entered Primary 3.

2.2. Test Instrument

One month before the school year, researchers (first and second authors) reviewed Primary 2 and Primary 3 mathematics curricula, textbooks, and workbooks used by Macau schools. A test instrument was constructed in consultation with the head of the Elementary Mathematics Department.

It consisted of 21 MCQs encompassing four operations to measure the arithmetic ability of elementary students. The elementary mathematics department head with 16 years of teaching experience and three mathematics teachers with teaching experience ranging from 1 to 3 years validated the test instrument. Four MCQs assessed addition, four assessed subtraction, six assessed division, five assessed multiplication, and two assessed mixed addition and subtraction (Table 2). In addition to the classification by operation, the 21 items were also categorized according to Bloom's cognitive levels: remembering, understanding, applying, analysing, and evaluating (Table 2). Furthermore, the 21 items were divided into arithmetic calculation problems (requiring students to apply one or more operations) and arithmetic word problems. There were 12 calculation problems and nine-word problems (Table 2).

Each item has four options, with three distractors and one correct option. The 21 items were grouped into two groups based on the option types. While 16 items required students to pick the correct numerical figure, five required students to select the correct algorithm (Table 2). All 21 MCQs were ordered randomly during the test. Prior to data collection, item difficulty was validated by 16 preservice science and mathematics teachers.

2.3. Data Collection Procedure

The test was administered at the beginning of the school year. It covered the arithmetic that students should have already learned based on The Curriculum Framework for Formal Education of Local Education and The Macao Requirements of Basic Academic Attainments of Local Education System (BAA). In the first week of September 2019, the teachers informed the students of the assessment date and coverage of this assessment. The teachers distributed a test instrument containing 21 items and a scantron answer sheet on the assessment date. The students recorded their answers by shading the box of the chosen option for each MCQ with a pencil. The students completed the test for 20 minutes.

The teachers collected scantron sheets at the end of the designated time. The sheets were provided to the first author, who used an optical mark reader to scan the answer sheets for data input. Each correct answer was recorded as 1, and each incorrect answer was recorded as 0. Thus, the maximum raw score was 21 and the minimum was 0. All student data (class number, item number, and student's answers) and test scores for each item were entered into a spreadsheet.

2.4. Calibration of Item Difficulties and Person Abilities

The Rasch model uses logarithmic transformation to calibrate a person and items on the same single-dimensional ruler (Wright & Masters, 1982). Based on their respective positions on this single-dimensional continuum, comparisons can be made between person and person, item and item, and person and item, yielding objective and linear measures of person's ability and item difficulty. Data were analyzed using Winsteps 4.4.5 and the dichotomous Rasch model (Rasch, 1960) as each MCQ only had a correct or incorrect answer.

$$P_{ni1} = \frac{\exp\left(\theta_n - \delta_i\right)}{1 + \exp\left(\theta_n - \delta_i\right)}$$

The input of this logistic function is the difference between the person's ability θ_n and item difficulty δ_i. The output is the probability that person n correctly answers item i, P_ni1. Therefore, item difficulty and person ability were measured on a standard interval scale. The items had a mean difficulty of 100.00, a minimum of 81.61 (item 8), and a maximum of 119.50 (item 4). The students had a mean ability of 109.27, minimum of 89.99, and maximum of 130.33.

3. RESULTS

3.1. Quality of the Data

Data quality was evaluated by assessing data fit, reliability, and fairness using the Rasch Model of Winsteps (Lincare, 2012) prior to reporting the validation results.

3.1.1. *Fit diagnosis*

In the Rasch model, the expectation is that students with higher ability will have a higher probability of answering more difficult items correctly than students with lower ability who will have a higher probability of correctly answering easier items than difficult items (Wright & Stone, 1999, p. 48). Infit and Outfit evaluate how well the data fit the structure of this expectation.

An item with Zstd greater than 2.00 distorts the model fit because it is a poor fit due to unexpected, unrelated irregularities (Linacre, 2012). Of the 21 items, only Q9, the third most challenging item, had an Infit Zstd >2.00 and an Outfit Zstd >2.00. All other item Infit and Outfit values lie between -1.75 and +1.40, signifying a good fit to the Rasch model.

Items with MnSq from 0.50 to 1.50 are considered productive for measurement (Linacre, 2012). The item with the lowest Outfit MnSq was Q14 (0.65), and the item with the highest Outfit MnSq was Q1 (1.40). The item with the lowest Infit MnSq was Q20 (0.88) and the item with the highest Infit MnSq was Q9 (1.22). Therefore, all items were within the expected range of 0.50 to 1.50 and they usefully fit the Rasch model (Linacre, 2012). In other words, all items were retained for the subsequent analysis.

Q9 had the highest Infit Zstd and Outfit Zstd of the 21 items. This may imply that its wording is ambiguous, its options are misleading, or both.

3.1.2. *Reliability of data*

Item reliability and Pearson's reliability were used to indicate reliability. From easy to difficult, item locations operationally define the latent variable (Wright & Stone, 1999, p. 151). Thus, the items should be appropriately located and separated along a line to assess relative item difficulties and item redundancy gaps.

Item reliability shows how this sample of students separated the 21 items in the assessment. A value closes to 1.00 indicates a higher precision (Wright & Stone, 1999, p. 151). The item reliability in this assessment was 0.96, indicating that the sample size was sufficient to support the instrument's construct validity (Lincare, 2012).

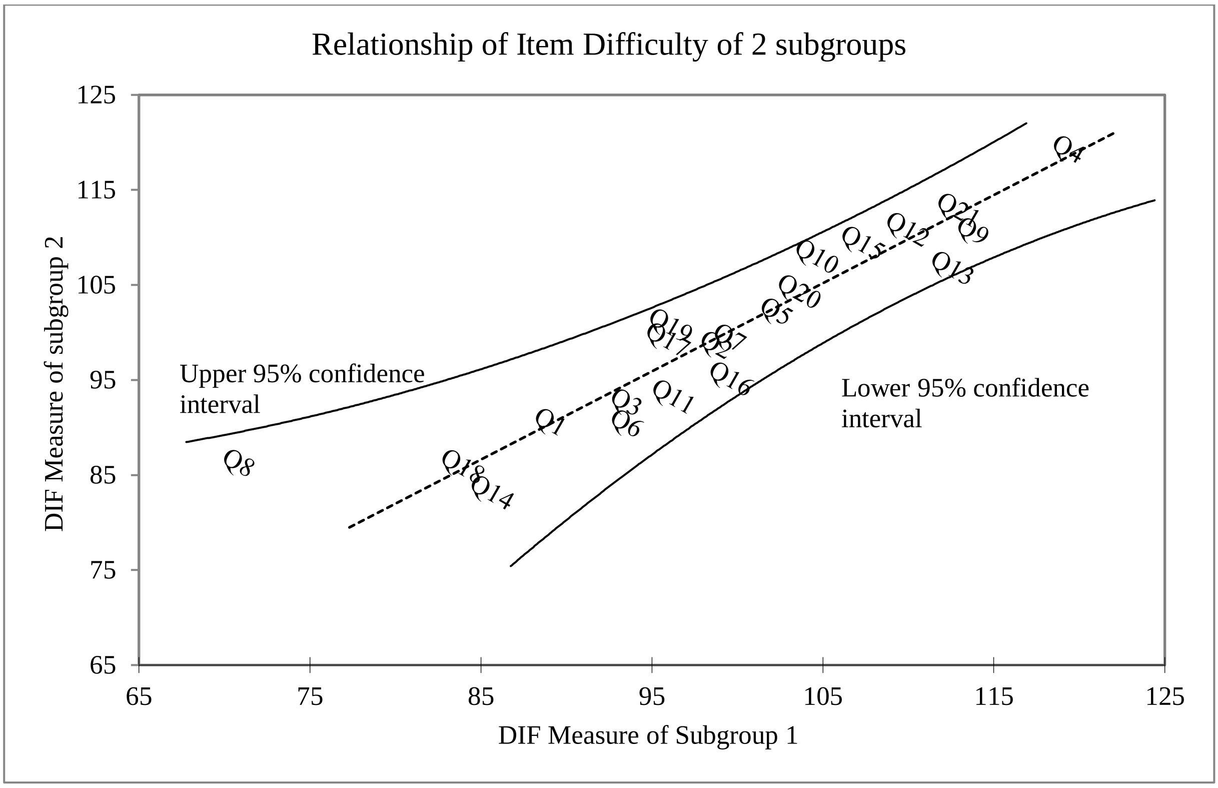
The Pearson reliability is the same as traditional test reliability (Lincare, 2012). This shows how this set of 21 items separated the sample of 138 students. It ranges from 0.00 to 1.00; a value close to 1.00 indicates higher precision (Wright & Stone, 1999, p. 151). The Pearson's reliability for this assessment was 0.60, indicating that the items were not sufficient to classify students into different ability levels. This means that the sample of students had similar ability levels or too few items (Lincare, 2012) to evaluate the latent variable.

3.1.3. *Fairness of data: Differential item functioning*

All items should behave similarly to students with the same abilities. If an item functions differently across different subgroups of students, the validity of this instrument may be questioned (Wilson, 2004). The sample students were randomly assigned to two subgroups and were expected to function the same across these subgroups. In other words, differential item functioning (DIF) should not occur (Bond & Fox, 2015, p. 281-282).

If an item had a p-value greater than 0.05, DIF did not occur significantly across the two subgroups (Linacre, 2009). As the USCALE of this assessment was 8.52, one logit equals 8.52 units (J.M. Linacre, personal communication, 22nd September 2020). Therefore, a DIF may exist if the DIF contrast is greater than 5.43. Q8 and Q13 had DIF contrasts of -15 and 6, respectively. However, because they are still within the 95% confidence interval (Figure 1), their DIF is not considered significant, which means that they behave similarly for different subgroups. The results indicated that the items were fair; therefore, item validity was upheld.

Figure 1. *DIF measures for the two subgroups under 95% confidence interval.*



3.2. Is Item Difficulty for Each Level Aligned with Student Ability in Order?

The Wright map represents item difficulty versus person ability (Figure 2a). The student ability has a range of 90.00 logits to 130.33 logits. The item difficulty ranges from 81.61 logits to 119.50 logits (Figure 2a). Items Q8, Q18, and Q14, which all students could answer, were more than one standard deviation below the mean item difficulty. Item Q4 was the most difficult and lay almost two standard deviations above the mean item difficulty. Although it is the most difficult item, ten students (7.24% of the sample size) were able to solve it and all the other items. In other words, the 21 designed items were not sufficient to assess the upper limit of the arithmetic ability of the ten students. This indicates that items that are more difficult than Q4 should be included in the instrument. In addition, there is a significant gap between Q4 and Q21. To better assess what students with an ability between 113 and 120 logits can do, more items of intermediate difficulty should be added to fill this gap.

In alignment with Bloom's taxonomy, the four easiest items (Q1, Q14, Q18, and Q8) mainly require students to remember and understand. In contrast, the most difficult items (Q4, Q21, Q9, Q13, and Q15) mainly required students to evaluate. Items with a medium difficulty level mainly require students to analyse and apply (Table 2 and Figure 2b). Observed item difficulty based on Rasch analysis and predicted item difficulty from the person construct map were compared (Table 3) to examine item difficulty alignment. The findings show that item difficulty hierarchy generally aligns with Bloom's taxonomy.

Table 3. *Comparison of alignment between observed items based on Rasch analysis and predicted construct map based on item development.*

Level	Observed item cognitive levels based on Rasch analysis	Predicted construct map based on item development in person construct map	Bloom's cognitive level
3 Solid	Q4, Q9, Q13, Q15, Q21	Q4, Q9, Q13, Q15, Q21	Evaluate
2 Developing	Q2, Q3, Q5, Q6, Q7, Q10, Q11, Q12, Q16, Q17, Q19, Q20	Q2, Q3, Q5, Q6, Q7, Q10, Q11, Q12, Q16, Q17, Q19, Q20	Apply and Analyse
1 Basic	Q1, Q8, Q14, Q18	Q1, Q8, Q14, Q18	Remember and understand

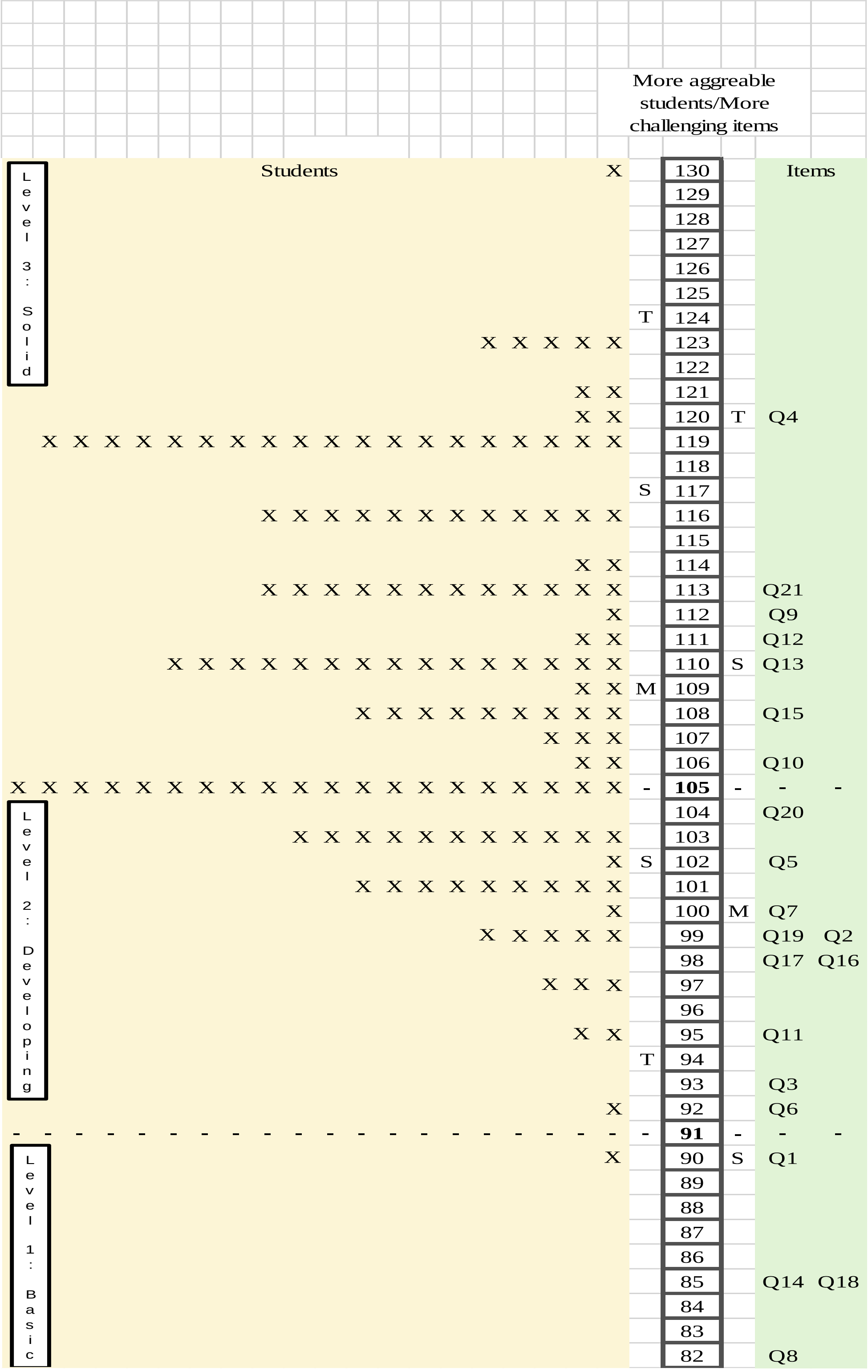
Note: Predicted cognitive items are from the person construct map (Table 2) and the observed items are based on Rasch analysis (Figure 2a).

To lay out the findings more precisely, items were further grouped according to Bloom's cognitive abilities (i.e., remember, understand, apply, analyse, and evaluate), as shown in Figure 2b. In general, the structure of items follows the expected order of the person construct map; that is, Level 1 (Basic) items are the easiest group of items among the three levels, Level 2 (Developing) items are more difficult than Level 1 but are easier than Level 3, and Level 3 (Solid) items are the most difficult among the three levels. A similar pattern is also observed in Figure 2c.

All nine word problems (mean difficulty = 119.50) were in the top half of the scale, and the remaining 12 arithmetic calculation problems (mean difficulty = 104.07) were in the bottom half. Apparently, word problems were more difficult than word problems.

However, the two-word problems (Q5 and Q6) shared the same difficulty level as the arithmetic calculation problems (Figure 2c). Hence, further investigation is required for Q5 and Q6.

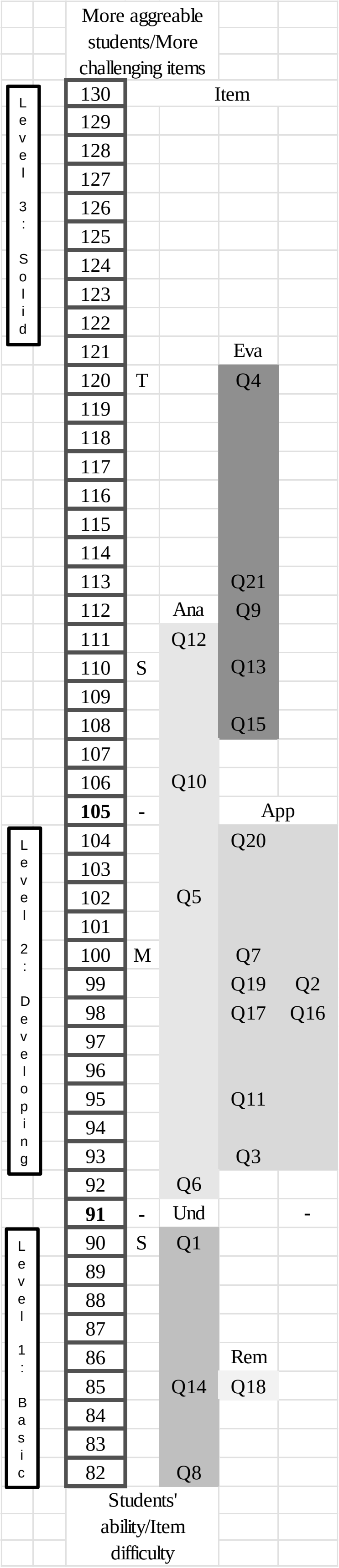
Figure 2a. *Wright map distribution of students' ability and item difficulty.*



Note:1. The observations on the left (in yellow) show the distribution of measured student abilities. The students showed the lowest ability at the bottom and highest ability at the top. The observations on the right in green show the item difficulty distribution, with the least challenging items at the bottom and the most challenging items at the top. M, on the left, indicates the mean student ability, S one standard deviation point, and T two standard deviation points of student ability, respectively. Similarly, on the right, M shows the mean item difficulty, S is one standard deviation point, and T is two standard deviation points of item difficulty.

Note: 2. "X" = 1 student; Mean person ability = 109.27 (Standard deviation = 7.44); Mean item difficulty = 100 (Standard deviation = 10.00)

Figure 2b. *Wright map – Cognitive ability.*



	More aggreable students/More challenging items				
Level 3: Solid	130		Item		
	129				
	128				
	127				
	126				
	125				
	124				
	123				
	122				
	121			Eva	
	120	T		Q4	
	119				
	118				
	117				
	116				
	115				
	114				
	113			Q21	
	112		Ana	Q9	
	111		Q12		
	110	S		Q13	
	109				
	108			Q15	
	107				
	106		Q10		
	105	-		App	
Level 2: Developing	104			Q20	
	103				
	102		Q5		
	101				
	100	M		Q7	
	99			Q19	Q2
	98			Q17	Q16
	97				
	96				
	95			Q11	
	94				
	93			Q3	
	92		Q6		
	91	-	Und		-
Level 1: Basic	90	S	Q1		
	89				
	88				
	87				
	86			Rem	
	85		Q14	Q18	
	84				
	83				
	82		Q8		
	Students' ability/Item difficulty				

Note: "Rem"=Remember; "Und"=Understand; "App"=Apply; "Ana"=Analyse, "Eva"=Evaluate.

Figure 2c. *Wright map – Arithmetic calculation and word problem.*

	More aggreable students/More challenging items				
Level 3: Solid	130			Item	
	129				
	128				
	127				
	126				
	125				
	124				
	123				
	122				
	121				WP
	120	T			Q4
	119				
	118				
	117				
	116				
	115				
	114				
	113				Q21
	112				Q9
	111				Q12
	110	S			Q13
	109				
	108				Q15
	107				
	106				Q10
	105	-	Cal		-
Level 2: Developing	104		Q20		
	103				
	102				Q5
	101				
	100	M	Q7		
	99		Q19	Q2	
	98		Q17	Q16	
	97				
	96				
	95		Q11		
	94				
	93		Q3		
	92				Q6
	91	-			-
Level 1: Basic	90	S	Q1		
	89				
	88				
	87				
	86				
	85		Q14	Q18	
	84				
	83				
	82		Q8		
	Students' ability/Item difficulty				

Note: "Cal" = arithmetic calculation problem (dark gray shading); "WP" = arithmetic word problem (light gray shading).

3.3. Are Respondent Categories (Basic, Developing, Solid) Aligned with the Person Construct Map?

Guided by the item design (Table 2), students were separated into three levels: Level 1 (basic students), Level 2 (developing students), and Level 3 (solid students). The students were assumed to be able to solve Level 1 items. Developing students were assumed to be able to solve Level 1 and 2 items. Lastly, the Solid students were assumed to be able to solve Levels 1, 2, and 3 items.

As the items were designed according to what the primary three students had learned in the first two, and the assessment was conducted in the first month of this school year, most of the students could solve all Level 1 items. Hence, only one student was at Level 1 (Figure 2a). Therefore, even students with the lowest ability could solve Level 1 items, given that their ability is higher than the difficulty of Level 1 items. In other words, the appropriate difficulty level for Level 1 items was overestimated based on the person construct map (Table 1) and item design of the building blocks (Table 2).

Developing students should be able to solve levels 2 and 1. There were 53 students (38.41% of the sample) in the developing student category, and the developing students were aligned with Level 2 items and above Level 1 items (Figure 2a).

Solid students should be able to solve Level 3, 2, and 1 items. There were 84 students (60.87%) in the solid student category. We noted that some solid students' abilities were higher than the difficulty of all items (Figure 2a). This suggests that the Level 3 item difficulty was overestimated.

This non-normal distribution of students for the Basic, Developing, and Solid levels (1, 53, and 84 students, respectively) suggests that more difficult items should be added to better assess arithmetic ability. Overall, the distribution of students (Figure 2a) followed the structure of the predicted construct map based on item development. Students at the Basic level (corresponding to Level 1 in Table 1) could solve the least difficult problems with the lowest cognitive level in Table 2. Students at the Developing level (Level 2) solved more difficult problems. Finally, students at the solid level (Level 3) solved the most difficult problems with a higher Bloom’s cognitive level.

3.4. Are Arithmetic Calculation Problems (Levels 1 and 2) More Accessible than Arithmetic Word Problems (Level 3)?

Figure 2c shows the operational grouping of arithmetic word problems and calculations. Items Q5 (“64 books) were packed in boxes of 7. How many books are left?’) and Q6 (“There are 14 students. None of the patients had coins. How many coins do they have altogether?”) combined arithmetic word problems with an arithmetic calculation problem operation (Figure 2c) because of the 2-step operation. Theoretically, they are expected to cluster with other problems. However, Rasch analysis of student responses placed them as difficult as arithmetic calculation items (Figure 2c).

When we compared these two items with other word problems, we noted the following: the mean number of words in the Level 3 word problems was approximately 24 words. In contrast, the mean numbers of words in Q5 and Q6 were 16. Therefore, one explanation for their position in Figure 2c is that students find these questions easier to read, comprehend, and solve because of the lower word number and complexity than other Level 3, more complex items. Furthermore, the operation that students require for solving the arithmetic word problems is directly given in the question stems of Q5 and Q6 (e.g., they are asked to find out “How many … altogether” and “How many … are left”). In other words, students immediately signalled that Q5 and Q6 were addition and subtraction problems, and they only needed to determine the correct numerical answer from the four numerical options (Table 2). As such, Q5 and Q6 require a lower level of language proficiency, one-step calculation, and lower thinking skills and may explain why they fall below other arithmetic word problems and into the Level 2 category.

In contrast, all other Level 3 items demanded higher-order thinking. For example, when students read these questions (e.g., "Does she need extra money?' If yes, how much?"; "Mr Lee spent three times as much as Mr Chan"; "At least how many …"; "At most …"), they cannot simply compute the required answer; instead, these require a higher level of language ability, analysis and at least two operation steps to arrive at an answer. In other Level 3 arithmetic word problems, students were required to select the correct algorithm from the four algorithms (Table 2). In other words, all other Level 3 word problems require higher-order thinking skills.

However, aside from items Q5 and Q6, the distribution of the other items fits the structure of the predicted construct map based on item development (Figure 2c). Thus, basic-level students can solve Level 1 items, developing-level students can solve Level 1 and Level 2 items, and solid-level students can solve Level 1, Level 2, and Level 3 items. It is worth noting that Levels 1 and 2 are primarily arithmetic calculation problems. At the same time, Level 3 items were solely arithmetic word problems. This implies that word problems can distinguish solid students from Developing and Basic students, which is consistent with the contention that Level 3 questions require a higher cognitive level. In comparison, Level 1 and 2 questions were accessible to students with lower cognitive levels.

However, to test the premise that word problems are more difficult and require a higher cognitive level than problems involving only calculation, Rasch's Principal Components Analysis (PCA) of residuals was performed on all questions to test unexpectedness (Linacre, 2012). PCA's standardized residual (loadings) was analysed after extracting the primary Rasch dimension. Higher factor loadings indicate substantial unexplained variance (Bond & Fox, 2015). In other words, the residuals of these items are not the result of random noise. This analysis tests whether the common factor can explain variance (Linacre, 1998, p. 636).

The items were clustered into two groups: the items that have a positive loading (from +0.01 to + 0.57) make up cluster 1, and items with a negative loading (from -0.42 to -0.10) make up cluster 2. Thus, items represent two strands of the same latent variable. One strand comprises arithmetic calculation problems, and the other comprises arithmetic word problems. Items with higher positive loadings are primarily single-step calculation problems. By comparison, those with higher negative loadings were primarily word problems. Thus, these findings support the contention of two strands: arithmetic calculation ability and the ability to solve arithmetic word problems.

4. DISCUSSION and CONCLUSION

This work aims to validate a formative assessment based on Wilson's (2004) four building blocks, which can be used to meaningfully measure students' understanding. The current study illustrates a measure of primary student arithmetic ability. The data for the hypothesized model show evidence of reliability and validity based on the Rasch framework, with the exception of item Q9 (*Mr Chan spent 3 dollars. Mr Lee spent 3 times as much as Mr Chan. Ms Fong spent 3 times as much as Mr Lee. How much did Ms Fong spend?*) where it reports an Infit Zstd of 2.95 and an Outfit Zstd of 2.90. We noted that the structure of each sentence in Q9 was simple, but the relationship was complex. This suggests that students might comprehend individual sentences, but not the overall context. Thus, they must select the best algorithm instead of the correct numerical answer, demanding higher-order thinking, metacognition, and language proficiency versus the one-step mathematical operations needed for other items. We retained this information in the assessment analysis of the current study; however, further investigation is required for future research. It also points out the need to systematically analyze individual test items that students may perceive differently than the teacher's original intention. Low person reliability indicates that there are insufficient items to evaluate the latent variable (Lincare, 2012). Hence, the research findings should be viewed within this limitation. Future research should consider adding more items to the test to enhance its reliability.

The hypothesized model, however, shows sufficient evidence of validity, as follows:

a) The difficulty distribution of the items follows the expected order for the construct map, where the difficulty for each designed item aligns with the student's cognitive level in that order.
b) The students' performance distribution followed the structure of the person-construct map. Students at the Solid level could answer items that require students to evaluate, students at the developing level could answer items that require them to apply and analyse, and students at the basic level could answer items that require them to memorize and remember. The results indicate that internal validity is upheld when the construct is in order, as expected (Wilson, 2004, p. 157–158). However, as only one student fell into the basic level, more challenging Level 1 items may need to be added to the instrument to improve internal validity. Future research may need to redefine the three-level categorization of students. However, the performance of this cohort of students was unexpectedly higher than anticipated.
c) Solving arithmetic calculation problems is the foundation of solving arithmetic word problems. However, the latter requires additional skills such as reading comprehension and analysis, pattern recognition, semantic relations, and problem-model strategies (Chiang & Chen, 2019; Cummins, 1991; Prakitipong & Nakamura, 2006; Riley et al., 1983; Simon, 1978; Weitheimer, 1959). Pertinent literature supports this contention. Given the above discussions, we found that word problems can be categorized into simpler and more difficult items. Therefore, future research is required to study the factors affecting the difficulty of word problems.

The credibility and interpretation of the assessment information are not dependent only on the item content. To be instructionally useful, the items must define a meaningful hierarchy of increasing difficulty in which easier items assess the conceptual understanding needed in the solution of more difficult items (Alonzo & Steedle, 2009; Black et al., 2011; Fisher, 2013). In addition, item content must be aligned with local learning objectives if the goal of coherence is to be realized (Baird et al., 2017). The formative assessment in the current study provides this type of assessment information. Specifically, the conceptualizations of understanding about the topic following the four building blocks (Wilson, 2004) in the current study provide information on where a student stands relative to intended learning outcomes in a person construct map. Classroom assessment of this kind sets up information sources for teachers to formulate valuable and timely feedback for students about '*what might be useful to do next* (Black & William, 1998; Mislevy et al., 2003). This improves coherence in documenting learning, enhances classroom feedback, and shifts focus away from grades to more authentically serve classroom assessment purposes in facilitating learning at the individual level.

Acknowledgments

We want to thank the mathematics teachers in this study. This research would not have been possible without their willingness to share their experiences.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: University of Macau, SSHRE19-APP065-FED.

Authorship Contribution Statement

Wai Kei Chan: Design of the research project, design of assessment, data collection and analysis, and drafting the manuscript. **Li Zhang**: Design of the research project, design of assessment, and drafting the manuscript. **Emily Pey-Tee Oon**: Design of the research project, critical feedback on the manuscript, and writing and editing of the manuscript.

Orcid

Wai Kei Chan https://orcid.org/0000-0003-2023-5141
Li Zhang https://orcid.org/0000-0003-1091-5979
Emily Pey-Tee Oon https://orcid.org/0000-0002-1732-7953

REFERENCES

Adillah, G., Ridwan, A., & Rahayu, W. (2022). Content validation through expert judgement of an instrument on the self-assessment of mathematics education student competency. *International Journal of Multicultural and Multireligious Understanding*, *9*(3), 780-790. http://dx.doi.org/10.18415/ijmmu.v9i3.3738

Alderson, J.C. (1990). Testing reading comprehension skills. *Reading in a Foreign Language*, *6*(2), 425-438.

Alonzo, A.C., & Steedle, J.T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389-421.

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

Baroody, A.J., & Dowker, A. (Eds.). (2003). *The development of arithmetic concepts and skills: Constructing adaptive expertise.* Lawrence Erlbaum Associates Publishers.

Beck, K. (2020). Ensuring content validity of psychological and educational tests – the role of experts. *Frontline Learning Research*, *8*(6), 1-37. https://doi.org/10.14786/flr.v8i6.517

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*(5), 536-553.

Björklund, C., Marton, F., & Kullberyg, A. (2021). What is to be learnt? Critical aspects of elementary skills. *Educational Studies in Mathematics*, *107*, 261-284. https://doi.org/10.1007/s10649-021-10045-0

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Black, P., & Wiliam, D. (2003). In praise of educational research': Formative assessment. *British Educational Research Journal*, *29*(5), 623-637.

Black, P., & Wiliam, D. (2010). *Inside the black box: Raising standards through classroom assessment*. *Phi Delta Kappan*, *92*(1), 81-90. https://doi.org/10.1177/003172171009200119

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspectives*, *9*, 1-52.

Bloom, B.S. (1956). *Taxonomy of educational objectives. The classification of educational goals*. Handbook 1: Cognitive domain. David McKay.

Bond, T., & Fox, C.M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Brookhart, S.M., Moss, C.M., & Long, B.A. (2010). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 41-58.

Cardinet, J. (1989). Evaluer sans juger. *Revue Française de Pédagogie*, *88*, 41-52.

Cataloglu, E., & Robinett, R.W. (2002). Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career. American Journal of Physics, *70*(3), 238-251. https://doi.org/10.1119/1.1405509

Chiang, T., & Chen, Y. (2019). Semantically-aligned equation generation for solving and reasoning math word problems. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1272

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205-249.

Cummins, J. (1991). *Interdependence of first- and second-language proficiency in bilingual children.* Language Processing in Bilingual Children, pp. 70-89. Cambridge University Press. https://doi.org/10.1017/cbo9780511620652.006

Dixson, D.D., & Worrell, F.C. (2016). Formative and summative assessment in the classroom. *Theory into Practice*, *55*(2), 153-159. https://doi.org/10.1080/00405841.2016.1148989

Dowker, A. (2005) Early Identification and Intervention for Students with Mathematics Difficulties. *Journal of Learning Disabilities*, *38*(4), 324. http://dx.doi.org/10.1177/00222194050380040801

Duckor, B., & Holmberg, C. (2017). *Mastering formative assessment moves: 7 high-leverage practices to advance student learning*. ASCD Press.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.

Engvall, M., Samuelsson, J., & Östergren, R. (2020). The effect on students' arithmetic skills of teaching two differently structured calculation methods. *Problems of Education in the 21st Century*, *78*(2), 167-195. https://doi.org/10.33225/pec/20.78.167

Fisher, W.P., Jr. (1997). Is content validity valid? *Rasch Measurement Transactions*, *11*, 548.

Fisher, W.P., Jr. (2013). Imagining education tailored to assessment as, for, and of learning: Theory, standards, and quality improvement. *Assessment and Learning*, *2*, 6-22.

Geary, D.C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, *114*(2), 345–362. https://doi.org/10.1037/0033-2909.114.2.345

Goldman, S.R., & Hasselbring, T.S. (1997). Achieving meaningful mathematics literacy for students with learning disabilities. *Journal of Learning Disabilities*, *30*(2), 198–208.

Gorin, J.S., & Mislevy, R.J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment (K-12 Center at Educational Testing Service No. Invitational Research Symposium on Science Assessment)*. ETS.

Gurel, D.K., Eryilmaz, A., & McDermott, L.C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *EURASIA Journal of Mathematics, Science and Technology Education*, *11*(5). https://doi.org/10.12973/eurasia.2015.1369a

Guskey, T.R. (2003). How classroom assessments improve learning. *Educational Leadership*, *60*(5), 6-11.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluwer.Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hidayati, K., Budiyono, & Sugiman. (2019). Using alignment index and Polytomous item response theory on statistics essay test. *Eurasian Journal of Educational Research*, *79,* 115-132.

Hiebert, J., & Lefevre, P. (1986). *Conceptual and procedural knowledge in mathematics: An introductory analysis*. In J. Hiebert (Ed.), Conceptual and procedural knowledge: The case of mathematics (pp. 1–27). Lawrence Erlbaum Associates, Inc.

Hong Kong Education Bureau (2018). *Explanatory Notes to Primary Mathematics Curriculum (Key Stage 1).*

Hong Kong Education Bureau (2000). *Mathematics education key learning area.*

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. National Academy Press.

Lee, N.P., & Fisher, W.P., Jr. (2005). Evaluation of the diabetes self-care scale. *Journal of Applied Measurement*, *6*(4), 366-381.

Linacre, J.M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, *12*(2), 636.

Linacre, J.M. (2000). Computer-adaptive testing: A methodology whose time has come. *MESA Memorandum*, *69*, 1991-2000.

Linacre, J.M. (2009). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of Applied Measurement, 10*(2), 157-169.

Linacre, J.M. (2012). *A user's guide to Winsteps. Ministeps. Rasch-model computer programs. Program manual 3.74.0*. https://www.winsteps.com/a/Winsteps-Manual.pdf

Liu, O.L., Lee, H.S., Hofstetter, C., & Linn, M.C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*(1), 33-35.

Luque-Vara, T., Linares-Manrique, M., Fernández-Gómez, E., Martín-Salvador, A., Sánchez-Ojeda, M.A., & Enrique-Mirón, C. (2020). Content validation of an instrument for the assessment of school teachers' levels of knowledge of diabetes through expert judgment. *International Journal of Environmental Research and Public Health*, *17*(22), 8605. https://doi.org/10.3390/ijerph17228605

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, *10*(9), 9-20.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). Macmillan Publishing.

Millians, M. (2011). Computational skills. In S. Goldsteing & J. A. Naglieri (Eds.), *Encyclopedia of child behavior and development*. Springer. https://doi.org/10.1007/978-0-387-79061-9_645

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3-62. https://doi.org/10.1207/s15366359mea0101_02

National Research Council (NRC), (2006). *Systems for state science assessment. Committee on Test Design for K-12 Science Achievement*. M. R. Wilson and M. W. Bertenthal (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. The National Academies Press.

Nasir, N.A.M., Singh, P., Narayanan, G., Mohd Habali, A.H., & Rasid, N.S. (2022). Development of mathematical thinking test: Content validity process. *ESTEEM Journal of Social Sciences and Humanities*, *6*(2), 18-29.

Parviainen, P. (2019). The development of early mathematical skills - A theoretical framework for a holistic model. *Journal of Early Childhood Education Research*, *8*(1), 162-191.

Popham, W.J. (2009). Our failure to use formative assessment: *Immoral Omission. Leadership*, *1*(2), 1-6.

Popham, W. (2010). Wanted: A formative assessment starter kit. *Assessment Matters*, *2*, 182.

Prakitipong, N., & Nakamura, S. (2006). Analysis of mathematics performance of grade five students in Thailand using Newman procedure. *Journal of International Cooperation in Education*, *9*(1), 111-122.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Mesa Press.

Riley, M.S., Greeno, J.G., & Heller, J.I. (1983). Development of children's problem-solving ability in arithmetic. In H. Ginsburg (Ed.), The development of mathematical thinking (pp. 153-196). Academic Press.

Shepard, L.A. (2006). *Classroom assessment*. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 623-646). Rowman & Littlefield.

Sievert, H., van den Ham, A.-K., & Heinze, A. (2021). Are first graders' arithmetic skills related to the quality of mathematics textbooks? A study on students' use of arithmetic principles. *Learning and Instruction*, *73*. https://doi.org/10.1016/j.learninstruc.2020.101401

Simon, H.A. (1978). *Information-processing theory of human problem solving*. In W.K. Estes (Ed.), *Handbook of learning and cognitive processes (Volume 5): Human information processing* (pp. 271-295). Psychology Press.

Stiggins, R.J. (1994). *Student-centered classroom assessment*. Merrill.

Vlassis, J., Baye, A., Auqui ère, A., de Chambrier, A.-F., Dierendonck, C., Giauque, N., Kerger, S., Luxembourger, C., Poncelet, D., Tinnes-Vigne, M., Tazouti, Y. & Fagnant, A. (2022). Developing arithmetic skills in kindergarten through a game-based approach: a major issue for learners and a challenge for teachers. *International Journal of Early Years Education*. https://doi.org/10.1080/09669760.2022.2138740

Wertheimer, M. (1959). *Productive thinking*. Enlarged Edition. Harper and Brothers.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* Jossey-Bass.

Wilson, M. (2004). *Constructing measures: An item response modeling approach.* Routledge.

Wole, G.A., Fufa, S., & Seyoum, Y. (2021). Evaluating the Content Validity of Grade 10 Mathematics Model Examinations in Oromia National Regional State, Ethiopia. *Mathematics Education Research Journal*. https://doi.org/10.1155/2021/5837931

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Mesa Press.

Wright, B.D., & Stone, M.H. (1999). *Measurement essentials*. Wide Range.