Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas

Development and validation of a scale to assess the frequency of using classroom assessment tools and methods for mathematics teachers

Osman Birgin
Usak University, Faculty of Education, Türkiye
osman.birgin@usak.edu.tr

Murat Yılmaz
100.Yıl Mehmetcik Middle School, Denizli, Türkiye
2043053001@ogr.usak.edu.tr

RESUMEN
Este estudio tiene como objetivo el desarrollo de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula por parte del profesorado de matemáticas. En el estudio participaron 320 profesores de matemáticas turcos. Los datos fueron analizados utilizando los paquetes de software estadístico SPSS 17.0 y LISREL 8.8. Para evaluar la validez del constructo se usó el AFE y el AFC, las correlaciones ítem-total y las validaciones convergente y discriminante. El AFE y el análisis de ítems se realizaron con 170 profesores de matemáticas. Tras el AFE, se identificó una escala de cuatro factores con 25 ítems, que explicaban el 52,38 % de la varianza. El AFC fue llevado a cabo con 150 profesores de matemáticas. Los resultados del AFC de primer y segundo orden indicaron que el modelo de la escala presentaba adecuados índices de ajuste, y se confirmó la validez convergente y discriminante de la escala. Los coeficientes alfa de Cronbach y de fiabilidad compuesta (RC) fueron satisfactorios. Los resultados de este estudio mostraron que la escala es un instrumento fiable y válido para evaluar la frecuencia con que el profesorado de matemáticas utiliza los instrumentos y métodos de evaluación en el aula.

PALABRAS CLAVE
Profesorado de matemáticas; Evaluación en el aula; Herramientas y métodos de evaluación; Elaboración de escalas.

ABSTRACT
The purpose of this study was to develop a scale to assess the frequency of using classroom assessment tools and methods for mathematics teachers. This study
was conducted on 320 Turkish mathematics teachers. The data were analysed using the SPSS 17.0 and LISREL 8.8 statistical software packages. EFA and CFA, item-total correlations, and convergent and discriminant validities were used to assess the construct validity. EFA and item analysis were performed with 170 mathematics teachers. After EFA, a four-factor scale with 25 items was identified, explaining 52.38% of the variance. The CFA was carried out with 150 mathematics teachers. The first and second-order CFA results indicated that the scale model had good fit indices, and the convergent and discriminant validity of the scale was confirmed. The Cronbach’s alpha and composite reliability (CR) coefficients were satisfactory. The results of this study showed that the scale is a reliable and valid instrument for assessing the frequency of mathematics teachers’ use of classroom assessment tools and methods.

KEYWORDS
Mathematics teacher; Classroom assessment; Assessment tools and methods; Scale development.

1. INTRODUCTION

In recent years, curriculum development studies have shifted from teacher-centred teaching approaches to student-centred teaching approaches such as collaborative learning, project-based learning, and social constructivist learning (National Council of Teachers of Mathematics [NCTM], 2020; Ministry of National Education [MoNE], 2018). In addition to procedural and conceptual learning, higher-order thinking abilities such as problem-solving, reasoning, critiquing, analysing, synthesizing, and creating are highlighted (Krathwohl, 2002; Mullis et al., 2016). In the student-centred approach to education, the connection of concepts to everyday life, active participation in the teaching process, and the construction of students’ knowledge are crucial (Fosnot, 1996). Furthermore, assessment is an integral part of the teaching process, focusing on both the products and the process of learning (Shepherd, 2000). As a result, classroom assessment requires the use of alternative assessment tools and methods that enable students to determine their performance in the learning process, identify their strengths and areas for development, and measure high-level knowledge and skills (Birgin, 2010).

The Turkish mathematics curriculum (MoNE, 2018), which has been updated in parallel with contemporary education approaches, emphasises the use of various alternative assessment methods rather than a single assessment method to determine student achievement. As a result, in addition to traditional and objective assessment tools (e.g., written examinations, multiple-choice tests), teachers are expected to use alternative assessment tools in the learning process, such as performance tasks, portfolios, demonstrations, interviews, observations, self- and peer-assessment, rubrics, and concept maps. Implementation of the curriculum is largely possible with teachers who have pedagogical knowledge and experience of the curriculum’s teaching and assessment approaches. Research findings that reveal teachers’ inadequacies and lack of knowledge about alternative assessment tools in Türkiye (e.g., Aras, 2020; Birgin & Baki, 2009; Gelbal & Kelecioglu, 2007; Gök & Şahin, 2009; Özenç & Çakır, 2015) raise concerns about the quality of classroom assessment practices.

In the literature, the frequency of use of teachers’ assessment techniques has been investigated using instruments such as questionnaires, scales, observation, and interviews. It was found that scale development studies to assess the frequency of use of teachers’ assessment tools and methods, on the other hand, were limited (e.g., Adams & Hsu, 1998; Bol et al., 1998; Cizek et al., 1995; McMillan et al., 2002; Nazlıçicek & Akarsu, 2008; Yıldırım Ekinci & Köksal, 2011). While exploratory or confirmatory factor analysis was conducted in some studies to assess the use of teachers’ assessment techniques (e.g., Bol et al., 1998; McMillan et al., 2002; Yıldırım et al., 2011), some
instruments were developed based on expert opinion in other studies. Moreover, the scale items and the sub-factors, which were previously developed for determining teachers’ assessment tools and methods, differed according to curricula, grades, and course types. In addition, some studies showed that the preferred assessment techniques of mathematics teachers differed from those of teachers in other disciplines (e.g., Bol et al., 1998; Nazlıççek & Akarsu, 2008; McMillan et al., 2002). As a result, there is a need for an instrument to assess the use of classroom assessment tools and methods that are compatible with the mathematics curriculum in Türkiye. This study aims to develop a scale to assess the frequency of use of classroom assessment tools and methods for mathematics teachers. This study contributes to filling the gap in the literature.

1.1. Importance of Classroom Assessment

Assessment serves several purposes, including determining student achievement, placement in the curriculum, determining the effectiveness of the curriculum, teaching methods and strategies, identifying students’ learning strengths and weaknesses, and monitoring their development (MoNE, 2018). However, the classroom assessment to support teaching and learning is fundamentally different from large-scale testing designed to monitor trends, hold schools to account, evaluate teachers and programmes, or inform the selection and placement decisions (Shepard, 2020). As a result, classroom assessment is typically used for diagnostic, formative, and summative assessment in the teaching process (Black & William, 1998; Stiggins, 1997). Diagnostic assessment is used at the beginning of the teaching process to gather detailed information about students’ prior knowledge, misconceptions, interests, abilities, and other characteristics. It is a specific form of formative assessment. This assessment allows for the selection, organisation and planning of content, teaching methods and strategies based on the needs of the students. Formative assessment is used to monitor students’ learning processes throughout the instruction and design the learning environments to optimise learning (William, 2011). Formative assessment aims to identify and compensate for errors and deficiencies in learning, rather than grading. This assessment is seen as a way of determining student progress, providing feedback, and determining future instructional activities (McMillan, 2013). As a result, the formative assessment focuses on the students’ learning and development and helps to eliminate deficiencies and structure knowledge (Birgin & Baki, 2012). Another type of assessment is a summative assessment, which is usually conducted at the end of the teaching period to assess learning, certification, and accountability. This assessment is used to determine whether students have achieved the objectives and for grading. It is usually carried out with examinations at the end of the unit or term (Shepard, 2000).

In recent years, classroom assessment has been described using the concepts of assessment of learning (Aol), assessment for learning (Afl), and assessment as learning (Aaf) (Assessment Reform Group [ARG], 1999; Black & Wiliam, 2018; Earl, 2012; McMillan, 2013; Sadeghi & Rahmati, 2017). The concepts of Aaf, Afl and Aol reflect different approaches to assessment. In general, Aaf describes the active participation of students in assessment and learning, Afl describes the identification of learning throughout the assessment, and Aol describes the measurement of learning through assessment (Schellekens et al., 2021). Aol is a formal assessment of what a learner has learned at the end of a chapter, module, or course. The purpose of the Aol is to identify and label students’ successes or failures as a result of the teaching process (Schellekens et al., 2021). It is generally used to assess performance and measure outcomes following a formal learning activity. It often takes the form of a test or examination. It also functions as a summative assessment (Earl, 2012). Afl is an approach to teaching and learning that generates feedback that is used to improve the performance of students (ARG, 1999). Afl requires students to become more engaged in their learning. The purpose of the AFL approach is to motivate students, provide feedback, and identify areas for further learning. Afl has been linked to formative assessment because it helps to structure student learning through practices such as questioning and providing feedback (McMillan, 2013; Schellekens et al., 2021). Aaf, like the Afl, has a strong focus on the metacognitive learning processes and describes the role of students in monitoring and directing their learning (Black & Wiliam, 2018). Aaf involves active student participation in self-assessment.
and self-directed learning as a distinct function to enhance the learning process (Earl, 2012). AaL promotes student independence and flexibility to improve learning. According to this approach, the student is seen as an active evaluator to support the development of metacognitive and self-regulated learning skills (Lam, 2016).

In the AoL approach, traditional and objective assessment methods such as multiple-choice, fill-in-the-blank, short-answer items, and written tests are generally used for a period of time. However, these assessment tools are not sufficient to ensure student development and address learning deficits (Birgin, 2010; Stiggins, 1997). AfL and AaL approaches, on the other hand, impose new roles on teachers and students (ARG, 1999; Lam, 2016). In these approaches, the teachers are expected not only to assess the learning product but also to organise the learning environment and guide students to facilitate learning, while students are expected to actively participate in the learning process and assess themselves and their peers.

According to Klenowski (2009), AfL is conceived as being at the centre of classroom practice, thus representing the close relationship between the teacher and the student in the assessment process. Similarly, AaL is a formative assessment strategy that is part of ongoing classroom practice and aims to improve the quality of students’ learning processes (Clark, 2012). Numerous studies have provided evidence to support formative assessment (AfL and AaL implementation), which has been shown to increase achievement and conceptual learning (e.g., Andersson & Palm, 2017; Black & Wiliam, 1998; Heritage & Wylie, 2018; Kingston & Nash, 2011; Kyaruzi et al., 2019), allow students to take more responsibility for their learning, and motivate students to learn (e.g., Arnold, 2022; Black & Wiliam, 2018; Deluca et al., 2018; Wiliam, 2011). In addition, shared student success criteria, peer support, and teacher feedback through AfL and AaL practices helped students participate in their learning and were given opportunities to improve their work (e.g., Baas et al., 2020; Deluca et al., 2018; McDowell et al., 2011; Schellekens et al., 2021). Furthermore, Lam (2016) found that AfL, when advocated in classroom-based portfolios, can promote students’ motivation for learning and facilitate the development of students’ language awareness, self-assessment skills, and self-reflection. Therefore, AaL, AfL, and AoL should be balanced in the design and implementation of the lesson plans, courses, and curricula to gain the benefits of each assessment and to promote student learning (Schellekens et al., 2021). In this context, in addition to traditional and objective assessment tools, a variety of alternative assessment methods (e.g., performance task, project, portfolio, self- and peer-assessment, interview, demonstration, student journal, concept map) should be used to focus on the learning process, to monitor students’ development and to provide more feedback.

1.2. The Classroom Assessment Approach in the Turkish Mathematics Curriculum

After 2005, the primary and secondary education curricula in Türkiye were updated based on contemporary pedagogical approaches. The revised mathematics curriculum (MoNE, 2013; 2018) emphasised conceptual learning and high-level knowledge and skills. The objectives of the mathematics curriculum include applying learned knowledge and skills to everyday problems, developing mathematical reasoning and problem-solving skills, making connections between mathematical concepts and other disciplines, having positive attitudes and confidence in mathematics, acquiring self-regulation skills, using mathematics as a communication tool, and developing both aesthetic and mathematical skills (MoNE, 2018). The approach of the revised curriculum has also fundamentally changed assessment practices. According to the updated mathematics curriculum (MoNE, 2018), assessment should be used to measure students’ performance in the learning process, detect shortcomings, understand the effectiveness of teaching techniques, and show the weak and strong parts of teaching. Assessment should be used to monitor student progress and to support the learning process. The use of performance-based methods and a variety of assessment tools is also recommended to identify students’ learning levels and to support the development of different knowledge and skills.
In order to be aware of learning deficits and develop higher-order thinking skills, the mathematics curriculum (MoNE, 2018) recommends the use of alternative assessment methods such as portfolios, projects, performance tasks, rubrics, class discussions, presentations, exhibitions, observations, interviews, self- and peer-assessment, and concept maps, in addition to limited examinations. This approach implies that assessment is an integral part of the teaching process in the revised curriculum, that both the learning product and the learning process are considered in determining student performance, and that diagnostic and formative assessment are important as well as summative assessment.

According to the Ministry of National Education’s Preschool Education and Primary Education Institutions Regulation (MoNE, 2014), the success of middle school students is determined based on exam results, participation in lesson activities, and, if any, project tasks (p. 7). In the regulation, participation in lesson activities is defined as “the activities that students do in the classroom or school that enable them to use and develop their cognitive, affective, and psychomotor skills, such as critical thinking, problem-solving, reading comprehension, and research” (p. 1). In this regard, it is recommended to use product and process-oriented assessments, to actively engage students in the assessment, to consider individual and group performance, cognitive, affective, and psychomotor development, and participation in classroom activities. This curriculum approach suggests that AoL, AfL, and AaL should be balanced in the teaching process.

1.3. Previous Research on Teachers’ Use of Assessment Tools and Methods

Teachers’ knowledge and experience of classroom assessment are crucial to assessment practices (Zhang & Burry–Stock, 2003). Many studies have been conducted in different countries to determine teachers’ classroom assessment practices and their use of assessment tools (e.g., Adams & Hsu, 1998; Birgin & Baki, 2012; Brookhart, 1993; Cheng & Sun, 2015; Duncan & Nooman, 2007; Gök & Şahin, 2009; McMillan, 2001; McMillan et al., 2002; Randall & Engelhard, 2010; Zhang & Burry–Stock, 2003). For example, in a study conducted with 893 teachers, Bol et al. (1998) found that primary school teachers were significantly more likely to use alternative assessment strategies than high school teachers. Mathematics teachers also preferred traditional assessment methods much less than teachers of other subjects. In their study of 470 middle school teachers, Yazıcı and Sözbilir (2014) investigated teachers’ frequency of use and levels of proficiency in assessment methods. The results showed that teachers preferred traditional assessment methods over alternative approaches. It was found that there was a positive and significant relationship between the frequency of use and adequacy level of assessment methods. It was also found that the frequency of use and the level of adequacy of assessment methods differed significantly according to the year of teaching, gender, and subject. On the other hand, previous studies have revealed that many factors influence teachers’ classroom assessment practices, and these factors vary according to the curriculum applied in the county and grade (e.g., Cizek et al., 1995; Duncan & Noonan, 2007), teacher’s level of knowledge and proficiency in assessment (e.g., Birgin, 2010; Zhang & Burry–Stock, 2003), and their beliefs about assessment tools (e.g., Sun & Cheng, 2014).

The results of the previous studies showed that most of the teachers in Türkiye considered themselves more competent in using traditional and objective assessment tools, but they lacked sufficient knowledge and experience in using alternative assessment tools and methods and had difficulties (e.g., Duran et al., 2013; Gelbal & Kelecioğlu, 2007; Nazlıççek & Akarsu, 2008). Teachers used more summative assessment practices, adapted traditional classroom assessment practices to the updated curriculum (Birgin & Baki, 2012), and preferred alternative assessment tools less (e.g., Aras, 2019; Gök & Şahin, 2009; Peker & Gülle, 2011; Yıldızlı, 2020). Teachers mostly used mixed question types (e.g., written answer required, multiple choice, and short answer) in examinations, and they preferred less alternative assessment methods other than the compulsory project and performance task (Acar–Erdol & Yıldızlı, 2018; Birgin, 2010).

Some studies have investigated the frequency of teachers’ use of assessment techniques using an instrument such as a questionnaire, scale, observation, and interview. However, there are few scale development studies to determine the assessment tools and methods used by teachers.
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas

Osman Birgin, Murat Yılmaz


2. METHODS

2.1. Participants

The participants in the study were selected through voluntary and purposive sampling. The location of the school, the type of school (secondary school, high school), and the seniority of the teachers were used as criteria in the selection of purposive sampling. The study group consists of a total of 320 mathematics teachers in a province in the south-western region of Türkiye. This study was carried out with two sample groups selected through a convenience sampling method. Item analysis and exploratory factor analysis (EFA) of the scale were carried out with the first study group consisting of 170 mathematics teachers. Confirmatory factor analysis (CFA) of the scale was conducted with the second study group consisting of 150 mathematics teachers.

As can be seen in Table 1, 180 of the participants worked in middle schools (grades 5–8) and 140 of them were in high schools (grades 9–12). 49.7% (n=159) of the participants were female and 50.3% (n=161) were male mathematics teachers. Regarding the location of the school, 50% (n = 160) of the participants were located in the city centre, 31.6% (n = 101) of them were located in the provincial centre, and 18.4% (n = 59) of them were located in the countryside. In terms of professional seniority, 15.3% of participants had 1–5 years of seniority; 20.6% of them had 6–10 years of seniority; 15.6% of them had 11–15 years of seniority; 16.6% of them had 16–20 years of seniority; and 31.9% of them had more than 20 years of seniority.
2.2. Scale Development Process and Data Analysis

The first step was to develop scale items for the classroom assessment tools and methods used by mathematics teachers. For this purpose, previous studies were reviewed (e.g., Adams & Hsu, 1998; Aras, 2019; Bol et al., 1998; Cheng & Sun, 2015; Cizek et al., 1995; Duncan & Noonan, 2007; Gelbal & Kelecioglu, 2007; McMillan, 2001; McMillan et al., 2002; Peker & Gule, 2011; Yildirim Ektinci & Koksal, 2011). In addition, the Turkish mathematics curriculum for middle and high school, which was updated in 2018, was reviewed. Following the literature review, 33 draft items were created by considering the types of questions (open-ended, multiple-choice, short-answer, fill-in-the-blank, matching, true-false, and mixed), traditional and alternative assessment tools and methods that can be used in mathematics (unit follow-up test, in-class observation, homework, performance task, project, self-assessment, peer-assessment, group assessment, demonstration, oral presentation, rubric, portfolio, class discussion, student journal, interview, concept map, diagnostic branching tree, structured grid, mathematics attitude scale, worksheet, poster, crossword puzzle, drama, word association test, semantic feature analysis, knowledge competition).

The draft scale was reviewed for content, face, and language validity by two experts in mathematics education, one expert in measurement and assessment education, two experts in educational science, and eight experienced mathematics teachers. Following the opinions and suggestions of the experts and mathematics teachers, it was decided to remove four items (poster, drama, semantic feature analysis, and knowledge competition) because they were less used in mathematics lessons. The items were scored on a 5-point Likert scale (1 = not at all, 5 = always) for the frequency of using classroom assessment tools.

In this study, ethical approval and institutional approval were obtained before the instrument application. Informed consent was obtained from the participants and the research was conducted on a voluntary participation basis. The validity and reliability procedures of the scale...
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas
Osman Birgin, Murat Yılmaz

Fueron realizados en dos etapas. En la primera etapa, la escala se administró a 170 maestros de matemáticas para determinar el factor estructura de la escala. Cada ítem se calificó de 1 (no en absoluto) a 5 (siempre). El paquete de software estadístico SPSS 17.0 se utilizó para el análisis de datos. Antes del análisis de datos, se realizaron análisis de datos, normalidad, linealidad y análisis de valores extremos (excepto los valores -3 y +3) para verificar la adecuación de los análisis y las suposiciones. Los ítems que fueron dejan sin llenar (datos faltantes) no se incluyeron en el cálculo de la puntuación de la escala. Los datos con valores extremos fueron excluidos del análisis. Se encontró que el kurtosis (-.110) y la skewness (-.305) de los datos para el primer grupo de estudio variaban entre -1 y +1 y tenían una distribución normal. Para examinar la estructura factorial de la escala, se realizó una análisis factorial exploratorio (EFA) utilizando la principal componente y la rotación Promax para permitir correlaciones entre los factores. Se utilizó el índice de correlación de ítem y el índice de discriminación de ítem para analizar los ítems.

En la segunda etapa, la análisis factorial confirmatorio (CFA), validez convergente y discriminante se realizaron con 150 maestros de matemáticas para confirmar la estructura factorial de la escala. Los coeficientes de alfa de Cronbach y la consistencia compuesta (CR) se calcularon para medir la consistencia interna de la escala y sus sub-factores. Los kurtosis y skewness valores de los datos para el CFA fueron encontrados a ser -.335 y -.230, respectivamente, y los datos tenían una distribución normal. Se realizó el CFA usando el software estadístico LISREL 8.8.

3. ENCUENTROS

3.1. Resultados del análisis factorial exploratorio (EFA)

Para determinar la estructura factorial de la escala, se realizó un EFA. El test de eficiencia de Kaiser-Meyer-Olkin (KMO) y el test de esfericidad de Bartlett se utilizaron para determinar la adecuación del data para el análisis factorial. El resultado del test de KMO se encontró en el .767, lo que indicó una buena adecuación de la muestra. Además, el test de esfericidad de Bartlett fue significativo (\(\chi^2 = 1753.17, df = 325, p < .001\)), lo que mostró que los ítems correlacionaban y un análisis factorial podía ser aplicado (Tabachnick & Fidel, 2007). Factores latentes fueron extraídos utilizando la Análisis Principal Componente (PCA) y la rotación Promax. El Promax rotation method is an oblique rotation that allows for correlations between the factors (Tabachnick & Fidel, 2007). Se recomendó que cada factor contenga al menos tres ítems con cargas mayores a .40 y sin cross-loadings (Kline, 2016). El análisis factorial inicial extrajo siete factores con valores de eigenvalues mayor a 1.00, explicando 64.07 % de la varianza (Figura 1). Como tres factores tenían tres o menos ítems, el análisis factorial fue restringido a cuatro factores.
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas
Osman Birgin, Murat Yılmaz

Figure 1. Scree plot graph of the scale

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor Common Variance</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>C22. Student journal</td>
<td>.628</td>
<td>.797</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C25. Diagnostic branching tree</td>
<td>.658</td>
<td>.769</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C26. Structure grid</td>
<td>.531</td>
<td>.726</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C23. Concept map</td>
<td>.538</td>
<td>.714</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C11. Self-assessment</td>
<td>.626</td>
<td>.658</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C10. Math attitude scale</td>
<td>.567</td>
<td>.656</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C24. Interview</td>
<td>.384</td>
<td>.509</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5. Matching item</td>
<td>.669</td>
<td>.848</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C6. True-False item</td>
<td>.675</td>
<td>.835</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It was found that one item loading (C14. Oral presentation) was below .30, and three items (C27. Worksheet, C28. Crossword puzzle, C29. Word association test) cross-loaded on more than one factor. Therefore, four items were removed and factor analysis was repeated on the remaining 25 items. Repeated factor analysis revealed a four-factor scale with eigenvalues greater than 1.5, explaining 52.38% of the variance, and factor loadings ranging from .513 to .848 (Table 2). Item common factor variances ranged from .40 to .67.
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas
Osman Birgin, Murat Yılmaz

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor Common Variance</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4. Fill-in-the blank item</td>
<td>.676</td>
<td>.825</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3. Short answer item</td>
<td>.429</td>
<td>.693</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C7. Mixed items</td>
<td>.497</td>
<td>.653</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2. Multiple-choice item</td>
<td>.432</td>
<td>.607</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C16. Project assessment</td>
<td>.579</td>
<td>.734</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C21. Portfolio</td>
<td>.532</td>
<td>.735</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C19. Class discussion</td>
<td>.558</td>
<td>.699</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C18. Rubric</td>
<td>.467</td>
<td>.609</td>
<td>-.316</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C13. Demonstration</td>
<td>.456</td>
<td>.598</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C17. Performance task</td>
<td>.440</td>
<td>.591</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C20. Group assessment</td>
<td>.406</td>
<td>.514</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C8. Unit follow-up test</td>
<td>.563</td>
<td>.767</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C9. In-class observation</td>
<td>.444</td>
<td>.698</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1. Open-ended question</td>
<td>.418</td>
<td>.622</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C15. Homework</td>
<td>.402</td>
<td>.513</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As can be seen in Table 2, the factor loadings of the first sub-factor, consisting of eight items, ranged from .509 to .797 and explained 21.36% of the variance, the factor loadings of the second sub-factor, consisting of six items, ranged from .607 to .848 and explained 14.09% of the variance. The factor loadings of the third factor, consisting of seven items, ranged from .514 to .734 and explained 9.02% of the variance, and the factor loadings of the fourth sub-factor, consisting of four items, ranged from .513 to .767 and explained 7.91% of the variance. The internal consistency of the sub-factors was assessed using Cronbach’s alpha coefficient, all of which ranged from .84 to .95 and were greater than .70, indicating good consistency (Table 2).

The sub-factors of the scale were named according to the characteristics of the items they contained. The items in the first factor (student diary, diagnostic branching tree, structured grid, peer assessment, self-assessment, interview, concept map, attitude scale) were named “alternative assessment methods” because they were mostly related to alternative assessment tools, the items in the second factor (multiple-choice, short-answer, fill-in-the-blank, true/false, matched, mixed) were named “objective assessment methods” because they were mostly related to objective assessment tools. The items in the third factor (portfolio, project, performance task, rubric, group assessment, demonstration, class discussion) were named “performance assessment methods” because they were performance assessment tools, and the items in the fourth factor (unit follow-up test, class observation, open-ended question, homework) was na-
med as "traditional assessment methods" because they were the assessment tools traditionally used by teachers.

3.2. Results of the Item Analysis

For the item analysis, the item-total correlation was first calculated. It is emphasised that an item-total correlation greater than.30 has good discriminating power, an item-total correlation between.20 and.30 can be included in the scale if it is considered necessary, and an item-total correlation less than.20 should be eliminated from the scale (Field, 2009).

<table>
<thead>
<tr>
<th>Item</th>
<th>Item-Total Correlation</th>
<th>t-value for item (upper - lower group)</th>
<th>Item</th>
<th>Item-Total Correlation</th>
<th>t-value for item (upper - lower group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C22</td>
<td>.528**</td>
<td>10.118**</td>
<td>C2</td>
<td>.326**</td>
<td>3.148*</td>
</tr>
<tr>
<td>C25</td>
<td>.571**</td>
<td>8.006**</td>
<td>C21</td>
<td>.617**</td>
<td>5.246**</td>
</tr>
<tr>
<td>C26</td>
<td>.654**</td>
<td>8.384**</td>
<td>C16</td>
<td>.520**</td>
<td>4.270**</td>
</tr>
<tr>
<td>C12</td>
<td>.642**</td>
<td>7.506**</td>
<td>C19</td>
<td>.611**</td>
<td>6.916**</td>
</tr>
<tr>
<td>C23</td>
<td>.489**</td>
<td>9.294**</td>
<td>C18</td>
<td>.596**</td>
<td>9.294**</td>
</tr>
<tr>
<td>C10</td>
<td>.632**</td>
<td>7.267**</td>
<td>C13</td>
<td>.635**</td>
<td>4.714**</td>
</tr>
<tr>
<td>C11</td>
<td>.605**</td>
<td>8.777**</td>
<td>C17</td>
<td>.560**</td>
<td>4.662**</td>
</tr>
<tr>
<td>C24</td>
<td>.635**</td>
<td>5.782**</td>
<td>C20</td>
<td>.612**</td>
<td>5.246**</td>
</tr>
<tr>
<td>C5</td>
<td>.448**</td>
<td>3.651**</td>
<td>C8</td>
<td>.406**</td>
<td>3.653**</td>
</tr>
<tr>
<td>C6</td>
<td>.344**</td>
<td>3.291**</td>
<td>C9</td>
<td>.365**</td>
<td>3.185*</td>
</tr>
<tr>
<td>C4</td>
<td>.308**</td>
<td>3.253**</td>
<td>C1</td>
<td>.347**</td>
<td>3.738**</td>
</tr>
<tr>
<td>C7</td>
<td>.317**</td>
<td>3.867**</td>
<td>C15</td>
<td>.420**</td>
<td>3.482**</td>
</tr>
<tr>
<td>C3</td>
<td>.320**</td>
<td>3.229**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *p <.05 **p <.01

Table 3 shows that the item-total correlations ranged from.308 to.654 and were significant (p <.01). Another way to determine the distinctiveness of the scale items was to compare the scores of the 27% upper and lower group scores using the t-test. As can be seen in Table 3, the results of the t-test for the 27% upper and lower groups of the scale items ranged from 3.148 to 10.188 and were significant (p <.01). These results showed that all items on the scale were discriminative.

3.3. Results of Confirmatory Factor Analysis (CFA)

Another way to test the construct validity of the scale is to use CFA (Tabachnick & Fidell 2007). Many fit indices are used in the CFA. Among the fit indices used in this study are the ratio of chi-square/degrees of freedom (RMSEA (Root Mean Square Error of Approximation), IFI (Incremental Fit Index), GFI (Goodness of Fit Index), AGFI (Adjusted Goodness of Fit Index), CFI (Comparative Fit Index) and NFI (Normed Fit Index), S-RMR (Standardised Root Mean Square Residual). For model fit, the ratio of should be less than 3, but less than 5 is also acceptable (Çokluk et al., 2012; Kline,
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas

Osman Birgin, Murat Yılmaz

2001). It is suggested that SRMR and RMSEA should be ≤.05 to obtain the best fit and ≤.08 to represent a good fit; IFI, CFI, and NNFI should be ≥.90 for a good fit (Çokluk et al., 2012; Kline, 2011), GFI and AGFI should be ≥.80 to obtain an acceptable fit (Schermelleh-Engel et al., 2003; Tabachnick & Fidell 2007).

In order to confirm the four factors of the scale on 150 mathematics teachers who were not part of the first sample group, a first-order CFA was conducted using LISREL 8.8. As a result of this analysis, the chi-square test (for goodness of fit was calculated as 567.71 ($df = 269, p < .001$), and the ratio of was found to be 2.11, which was less than 3. This result showed that the model had a good model fit. In addition, modifications were suggested for four items (C3 and C2, C17 and C16). Following the necessary modifications based on the results of the analysis, a path diagram with 25 items consisting of this four-factor structure was constructed (Figure 2). The standardised factor loadings of the items ranged from 0.57 to 0.94, and the t-test results for the items varied from 7.24 to 15.39, which were significant ($p < .01$).

As can be seen in Table 4, the first-order CFA fit index after structural modifications was found to be $= 1.35 (= 359.55, df = 267, p < .001$), which was less than 3. This result indicated that the model had a good fit (Kline, 2016). Other fit indices for the final model were found to be RMSEA =.048, NNFI =.98, IFI =.98, GFI =.84, CFI =.98, AGFI =.80, and SRMR =.060, respectively. Some fit indices (RMSEA, NNFI, IFI, CFI) indicated that the scale model had a good fit (Tabachnick & Fidell 2007), while the SRMR, AGFI, and GFI fit indices indicated that the model had an acceptable fit (Kline, 2016; Schermelleh-Engel et al., 2003). These results confirm the factor structure of the scale.

A second-order CFA was conducted to determine the structural relationship between the scale and the sub-factors. The results in Table 5 indicate that the standardised factor loadings of the first-order latent variables in the model ranged from.22 to.59, and the t-test results were significant ($p < .01$). All of the fit indices for the 2nd-order after structural modification in Table 6 ($= 373.25, df = 269, = 1.39, RMSEA =.051, NNFI =.98, IFI =.98, GFI =.83, CFI =.98, AGFI =.80, and SRMR =.084$) confirmed that the model had a good and acceptable fit (Çokluk et al., 2012; Kline, 2016).
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas

Osman Birgin, Murat Yılmaz

Figure 2. Path diagram related to the first-order CFA (standardised factor loads)

Chi-Square=359.57, df=267, P-value=0.00013, RMSEA=0.048
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas

Osman Birgin, Murat Yılmaz

Table 4. Fit indices related to first- and second-order CFA of the scale.

<table>
<thead>
<tr>
<th>Fit Index</th>
<th>$\chi^2$/df</th>
<th>RMSEA</th>
<th>IFI</th>
<th>GFI</th>
<th>AGFI</th>
<th>CFI</th>
<th>NNFI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Fit Criteria</td>
<td>≤.3</td>
<td>≤.05</td>
<td>≥.95</td>
<td>≥.95</td>
<td>≥.90</td>
<td>≥.95</td>
<td>≥.95</td>
<td>≤.05</td>
</tr>
<tr>
<td>Acceptable Fit Criteria</td>
<td>≤.5</td>
<td>≤.08</td>
<td>≥.90</td>
<td>≥.80</td>
<td>≥.80</td>
<td>≥.90</td>
<td>≥.90</td>
<td>≤.08</td>
</tr>
<tr>
<td>First-order CFA result</td>
<td>1.35</td>
<td>.048</td>
<td>.98</td>
<td>.84</td>
<td>.80</td>
<td>.98</td>
<td>.98</td>
<td>.060</td>
</tr>
<tr>
<td>2nd-order CFA result</td>
<td>1.39</td>
<td>.051</td>
<td>.98</td>
<td>.83</td>
<td>.80</td>
<td>.98</td>
<td>.98</td>
<td>.084</td>
</tr>
</tbody>
</table>

Table 5. Factor loadings in the 2nd-order CFA for the scale model.

<table>
<thead>
<tr>
<th>2nd-Order Variable</th>
<th>First-Order Variable</th>
<th>Factor Load</th>
<th>Error Variance</th>
<th>t-value</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment Methods</td>
<td>F1. Alternative assessment methods</td>
<td>.59</td>
<td>.65</td>
<td>3.79</td>
<td>.350</td>
</tr>
<tr>
<td></td>
<td>F2. Objective assessment methods</td>
<td>.28</td>
<td>.36</td>
<td>2.29</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>F3. Performance assessment methods</td>
<td>.66</td>
<td>.57</td>
<td>3.93</td>
<td>.430</td>
</tr>
<tr>
<td></td>
<td>F4. Traditional assessment methods</td>
<td>.22</td>
<td>.95</td>
<td>2.00</td>
<td>.050</td>
</tr>
</tbody>
</table>

3.4. Results of Convergent and Discriminant Validity

Convergent and discriminant validity are two fundamental aspects of construct validity. Convergent validity describes the degree to which the new scale is related to other variables and other measures of the same construct (Hair et al., 2010). The construct should correlate not only with similar, related factors but also with dissimilar, unrelated factors. Convergent validity was assessed using the average variance extracted (AVE). AVE indicates the average percentage of variation explained (variance extracted) by the items of a construct (Hair et al., 2014). According to Fornell and Larcker (1981), convergent validity is established when the AVE is greater than or equal to .50. The results in Table 6 showed that the AVE values for sub-factors were .511 (F1), .798 (F2), .712 (F3) and .745 (F4) respectively, confirming the convergent validity.

Table 6. Reliability, convergent, and discriminant validity results.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cronbach’s Alpha</th>
<th>AVE</th>
<th>CR</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1. Alternative assessment methods</td>
<td>.958</td>
<td>.745</td>
<td>.958</td>
<td>.863$^a$</td>
<td>.12</td>
<td>.41$^{**}$</td>
<td>.09</td>
</tr>
<tr>
<td>F2. Objective assessment methods</td>
<td>.867</td>
<td>.511</td>
<td>.859</td>
<td>.715$^a$</td>
<td>.15$^*$</td>
<td>.38$^{**}$</td>
<td></td>
</tr>
<tr>
<td>F3. Performance assessment methods</td>
<td>.946</td>
<td>.712</td>
<td>.944</td>
<td>.844$^a$</td>
<td>.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F4. Traditional assessment methods</td>
<td>.940</td>
<td>.798</td>
<td>.940</td>
<td>.893$^a$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: AVE = Average Variance Extracted, CR = Composite Reliability, $^a$Diagonal elements (in bold) represent the square root of AVE, $^*p < .05$, $^{**}p < .01$
Discriminant validity was established to determine the distinctiveness of the constructs in the scale. Discriminant validity indicates that construct measures that theoretically should not be highly correlated, are not related to each other (Hair et al., 2014). Discriminant validity was assessed by comparing the square root of the AVE with the correlations between the constructs. If the AVE of both constructs is greater than the squared correlation, then discriminant validity is confirmed (Fornell & Larcker, 1981). Table 6 shows that there were no high inter-correlations between the constructs of the scale, which ranged from 0.09 to 0.41. In addition, the square root of the AVE values for each construct, which ranged from 0.715 to 0.893, was found to be higher than the inter-constructs correlation. These results confirmed the discriminant validity of the sub-factors.

3.5. Results of Reliability Analysis

In this study, Cronbach’s alpha reliability and Composite reliability (CR) coefficients were calculated to determine the internal consistency of the scale. According to Fornell and Larcker (1981), CR is an indicator of the shared variance among the observed variables used as indicators of a latent construct. It is suggested that the reliability coefficient for the scales is higher than 0.70 (Field, 2005). As can be seen in Table 6, Cronbach’s alpha coefficient for the sub-factors was calculated to be 0.958 (F1), 0.867 (F2), 0.946 (F3), and 0.940 (F4), respectively. The Cronbach’s alpha coefficient of the total scale was found to be 0.91. The CR values for the sub-factors were found to be 0.958, 0.859, 0.944, and 0.940 respectively. These results indicated that the scale has a high internal consistency reliability.

4. DISCUSSION AND CONCLUSION

The purpose of this study was to develop a scale to assess the frequency of mathematics teachers’ use of classroom assessment tools and methods. A draft scale of 29 items was developed after reviewing previous studies and taking into account the opinions of experts in the field and teachers. In the first phase of this study, EFA was conducted with 170 teachers to assess the construct validity of the scale. The Kaiser–Meyer–Olkin (KMO) value for the suitability of EFA was 0.767, and the chi-square value for Bartlett’s sphericity test was 1753.17 ($p < 0.001$). These results demonstrated the feasibility of the EFA (Tabachnick & Fidel, 2007). Factor analysis was carried out using principle component analysis and the promax rotation method. The sub-factors were determined by ensuring that the factor loading of the items was at least 0.40 (Çokluk et al., 2012). The results of the first-factor analysis revealed a seven-factor scale with eigenvalues greater than one. Four items were removed from the analysis because their factor loadings were less than 0.40 or had cross-loadings. Repeating the factor analysis on the remaining 25 items extracted a four-factor scale with an eigenvalue greater than 1.5, whose factor loadings ranged from 0.51 to 0.85, and explained 52.38 % of the variance. In the social sciences, the variance explained in multifactor scales is accepted to be between 40 % and 60 % (Tabachnick & Fidel, 2007). These results indicate that the sub-factors of the scale and its related items are sufficient to explain the frequency of use of classroom assessment tools and methods for mathematics teachers.

For the item analysis, the item-total correlation was calculated and the t-test was used to determine the item discrimination for the 27 % upper and lower groups. For scale development studies, it is recommended that item–total correlation values should be greater than 0.30 and that items with values less than 0.20 should be removed from the scale (Çokluk et al., 2012). The item–total correlation values of the scale in the current study ranged from 0.31 to 0.65, and the t-test results for the upper (27%) and lower (27%) groups were significant ($p < 0.01$). These results showed that the discrimination of the scale items was within an acceptable range (Field, 2005).

The sub-factors of the scale that emerged from the result of EFA were named “traditional assessment methods”, “objective assessment methods”, “performance assessment methods” and “alternative assessment methods” based on the characteristics of the items. The Turkish mathematics curriculum suggests that assessment practices should be carried out for Assessment of Learning (AoL), Assessment for Learning (AfL), and Assessment as Learning (AaL) and that alter-
native assessment techniques should be used in the learning process (MoNE, 2018). Considering this aspect, it can be stated that the scale factors and items developed are appropriate for the mathematics curriculum. Furthermore, the sub-factors of this scale developed for mathematics teachers are similar to some previous studies (e.g., Bol et al., 1998; Cizek et al., 1995; Gelbal & Keleşciolu, 2007; McMillan et al., 2002; Nazlıçicek & Akarsu, 2008). For example, a three-factor scale developed by Nazlıçicek and Akarsu (2008) includes “traditional assessment methods” (e.g., written exam, test, quiz, homework, open-ended question) and “alternative assessment methods” (e.g., portfolio, project, observation, concept map, laboratory reports, self-assessment). Peker and Gülle (2011) developed a three-factor scale with “homework type assessment tools”, “classical assessment tools” and “alternative assessment tools”. The scale developed by Çalışkan and Kaşıkçı (2010) for social studies teachers has sub-factors including “objective assessment tool” (e.g., open-ended test, short-answer, matching item, oral exam) and “alternative assessment tool” (e.g., performance task, portfolio, portfolio, word association). Unlike previous studies, some sub-factors and items that emerged in this study differed from those in previous studies (e.g., Bol et al., 1998; Büyüktokatlı & Bayraktar, 2014; McMillan, 2001; McMillan et al., 2002; Nazlıçicek and Akarsu, 2008; Peker & Gülle, 2011; Yıldırım Ekinç and Köksal, 2011). This result can be explained by the differences in grade level and pedagogical approach of the curriculum adopted in the countries, as well as the assessment tools preferred by mathematics teachers and other subject teachers (e.g., Turkish, social studies, English, and science). In this respect, the scale in the current study contributes to the existing literature.

In the second stage of this study, first- and second-order CFAs were conducted with 150 mathematics teachers to assess the construct validity of the scale. After the proposed modification, the results of the first-order CFA (= 1.35, RMSEA =.048, IFI =.98, GFI =.84, CFI =.98, NNFI =.98) indicated that the model’s fit indices were good or acceptable (Çokluk et al., 2012; Kline, 2016; Tabachnick & Fidell, 2007). The fit indices of the model were found to be acceptable as a result of the second-order CFA (= 1.39, RMSEA =.051, IFI =.98, GFI =.83, CFI =.98, NNFI =.98). These results supported the four-factor structure of the scale. Convergent and discriminant validity was also established using average variance extracted (AVE) and composite reliability (CR) coefficients (Fornell & Larcker, 1981). Discriminant validity was assessed by comparing the square root of the AVE with the inter-construct correlations. The AVE values for the sub-factors, which ranged from.51 to.79 and were greater than.50, provided evidence of convergent validity. Discriminant validity was supported by the square root of the AVE values for sub-factors, which ranged from.715 to.893 and were higher than the inter-construct correlation.

The Cronbach’s alpha and the CR values were calculated to determine the internal consistency. It should be noted that Cronbach’s alpha and the CR values were greater than.70 (Çokluk et al., 2012). The Cronbach’s alpha values were determined to be.91 (total scale),.958 (F1),.867 (F2),.946 (F3), and.940 (F4), respectively. The CR values were determined to be.958 (F1),.859 (F2),.944 (F3), and.940 (F4). These results showed that the total scale and the sub-factors had satisfactory internal consistency.

In this current study, a scale of 25 items and 4 sub-factors was developed. According to the results of this study, the scale is valid and reliable for measuring the frequency of using classroom assessment tools and methods for mathematics teachers. The sample of this study is limited to Turkish middle and high school mathematics teachers. Therefore, the validity of this scale can be verified by administering it to teachers in countries other than Türkiye. In future research, in addition to determining the frequency of mathematics teachers’ use of classroom assessment methods, the factors that influence the frequency of use of assessment methods can be investigated using quantitative statistical approaches. Furthermore, the reasons for the frequency of teachers’ use of assessment tools and methods can be explored in detail by using qualitative research techniques.
ACKNOWLEDGMENTS
This study is derived from the second author’s Master’s thesis.

ETHICAL STANDARD STATEMENT
Research ethics approval was obtained from the Ethics Committee for Scientific Research and Publication of Usak University (date: 08.03.2022, decision no: 2022-09).

DISCLOSURE STATEMENT
No potential conflict of interest was reported by the authors

5. REFERENCES


Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas

Osman Birgin, Murat Yılmaz


Lam, R. (2016). Assessment as learning: Examining a cycle of teaching, learning, and assessment of writing
Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas

Osman Birgin, Murat Yılmaz


Ministry of National Education (MoNE) (2013). Ortaokul matematik dersi (5, 6, 7 ve 8. Sınıflar) öğretim Programı [Middle school mathematics curriculum (5th, 6th, 7th and 8th grades)]. Ankara: MoNE.


Osman Birgin, Murat Yılmaz

Desarrollo y validación de una escala para evaluar la frecuencia de uso de herramientas y métodos de evaluación en el aula para profesores de matemáticas


