International Journal of Assessment Tools in Education
2021, Vol. 8, No. 2, 357–375
https://doi.org/10.21449/ijate.909316
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Developing a Two-Tier Proportional Reasoning Skill Test: Validity and Reliability Studies

Kubra Acikgul [1,*]

[1]İnönü University, Faculty of Education, Department of Mathematics Education, Malatya, Turkey

ARTICLE HISTORY
Received: Dec. 08, 2020
Revised: Mar. 02, 2021
Accepted: Mar. 16, 2021

Keywords:
Proportional reasoning,
Two-tier test,
Middle school.

Abstract: The main aim of this study is to develop a useful, valid, and reliable two-tier proportional reasoning skill test for middle school 7th and 8th-grade students. The research was carried out using the sequential explanatory mixed method. The study group of this research comprised of 391 ($n_{7th\text{-}grade}$= 223, $n_{8th\text{-}grade}$= 168) students. With validity and reliability studies, the content, face, construct, discriminant validity, and reliability coefficient of the test were examined. As a result, the two-tier proportional reasoning skill test with 12 items under 3 factors (qualitative prediction and comparison, missing value, numerical comparison) valid and reliable for adequate values specified in the literature.

1. INTRODUCTION

The concept of proportional reasoning has wide usability in mathematics education. According to the National Council of Teachers of Mathematics (NCTM) (2000), proportional reasoning is a unifying concept since most of the many important concepts of elementary school mathematics are related to it. The fact that proportional reasoning is an important factor in establishing relations between concepts makes it to be accepted as one of the basic thoughts that form the core of the mathematics curriculum (Lesh et al., 1988). For example, according to the Common Core State Standards for School Mathematics, proportional reasoning is one of the basic math skills in the USA. Similarly, proportional reasoning also has an important place in the Middle School Mathematics Curriculum in Turkey (Ministry of National Education [MoNE], 2018). Further to that, proportional reasoning skills are the basis for gaining higher mathematical reasoning beyond elementary school mathematics (Lesh et al., 1988). Many subjects of geometry, analysis, and algebra require students to have proportional reasoning skills (Allain, 2000). Thus, it is called the capstone of elementary concepts (e.g., arithmetic, measurement) and the cornerstone of higher-level mathematics (Lesh et al., 1988; Post et al., 1988). Despite the mentioned importance, interest, and emphasis in the curriculums and literature, proportional reasoning skill is mentioned complex, difficult to teach, and cognitively challenging for students (Alfieri et al., 2015; Lamon, 2007). So, it is considered important to examine students' proportional reasoning skills to determine the difficulties experienced by

*CONTACT: Kübra AÇIKGÜL kubra.acikgul@inonu.edu.tr İnönü University, Faculty of Education, Department of Mathematics Education, Malatya, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

students and teachers. Hilton et al. (2013) stated that developing a diagnostic tool to measure the proportional reasoning skills of the students could be valuable for teachers to determine the teaching activities that are suitable for the specific needs of the students in proportional reasoning education.

This study aimed to develop a useful, valid, and reliable two-tier Proportional Reasoning Skill Test (PRST) for middle school 7th and 8th-grade students. In this study, the content validity, face validity, construct validity, discriminant validity, and Cronbach alpha, and composite reliability coefficients of the PRST were examined. The ability of students to distinguish non-proportional problems from proportional problems shows that they have proportional reasoning (Lim, 2009). So, in the current study, to determine the discriminant validity of the PRST, the relationship between PRST scores and Non-Proportional Reasoning Skill Test (N-PRST) scores was examined. Therefore, this study also aimed to develop a useful, valid, and reliable two-tier N-PRST.

1.1. Proportional Reasoning

Proportional reasoning is fundamental to understand many situations in mathematics and science education (e.g., density, speed) and everyday life problems (Cramer & Post, 1993). Therefore, the importance of proportional reasoning is mentioned and defined in many studies. Tourniaire and Pulos (1985, p. 181) define proportion as "a statement of the equality of two ratios i.e. a/b=c/d". According to Behr et al. (1988, p. 92), proportional reasoning is "a form of mathematical reasoning that involves a sense of covariation and multiple comparisons, and the ability to store and process several pieces of information". Lamon (2007, p. 647) refers to proportional reasoning as "detecting, expressing, analyzing, explaining, and providing evidence in support of assertions about proportional relationships". Also, many researchers (e.g., Behr et al., 1992; Cramer et al., 1993; Lesh et al., 1988) explains proportional reasoning as an understanding of the comparisons between quantities embedded in proportional situations. Proportional reasoning is also mentioned as an ability to distinguish between proportional and non-proportional situations (Cramer et al., 1993; Lesh et al., 1988; Lim, 2009). Considering these definitions, in this study, proportional reasoning is defined as an ability that includes multiplicative comparisons between quantities and requires distinguishing between proportional and non-proportional situations.

According to Weinberg (2002), most students have difficulty with proportional reasoning as that they do not understand the proportional situation or the solution strategy to be used. Various researches reveal that (e.g., Arıcan, 2019; Cramer et al., 1993; Dinç-Artut & Pelen, 2015; Mersin, 2018; Pelen & Dinç-Artut, 2015; Singh, 2000; Van Dooren et al., 2010), students could not distinguish proportional and non-proportional situations, and they use additive reasoning instead of multiplicative reasoning while solving proportional problems. Moreover, students use many faulty solution strategies such as not being able to determine when to use proportional reasoning (Van De Walle et al., 2013) and ignoring some of the data given in the problems (Özgün-Koca & Altay, 2009).

Students also have difficulty solving different types of proportional reasoning problems. For example, Lawton (1993) determined that the problem type is a factor that influences proportional reasoning. Dinç-Artut and Pelen (2015) found that the problem types affect the strategies used by students. According to Soyak and Işıksal (2017), to overcome the students' difficulties about proportional situations, students should be exposed to different proportional reasoning problem types. Therefore, it was decided to include different types of problems in the PRST.

Cramer and Post (1993) classify proportional reasoning problems in three categories; missing-value problems, numerical comparison problems, and qualitative prediction and comparison

problems. In missing value problems, three values of four numerical values are given and the other value is asked. In numerical comparison problems, two rates are given and the rates are compared. Qualitative prediction and comparison problems require comparisons that are not dependent on specific numerical values. The test developed in this study includes these three problem types presented by Cramer and Post (1993).

Non-proportional reasoning problems are also used in determining proportional reasoning in this study. Van Dooren et al. (2005) classify non-proportional problems as additive, constant, and linear. The linear problems are in the form of $f(x) = ax + b$ ($b \neq 0$). The additive problems are expressed as a constant difference between two variables. In the constant problems, there is no relation between the given variables. In this study, Van Dooren et al. (2005) classification was used in developing the N-PRST.

1.2. Significance of the Study

Many important topics of the elementary school curriculum are linked to proportional reasoning skills (NCTM, 2000). Especially middle school (5-8th grades) is considered as a critical period to create meanings about proportion reasoning (NCTM, 2000; Van Dooren et al., 2010). Therefore, proportional reasoning is one of the essential mathematical skills that middle school students should have. According to Ayan and Isiksal-Bostan (2019), middle school years are the best period for forming new understandings about proportions and developing proportional reasoning, so it is important to examine the middle school students' proportional reasoning skills. Thus, in this study developing a PRST for middle school students is considered crucial.

Open-ended and multiple-choice problems are problem types measuring students' learnings and understandings in education and research fields (Ozuru et al., 2013). Similarly, when proportional reasoning tests for middle school students are examined, it is seen that most tests consist of open-ended or multiple-choice problems. Also, the number of tests consisting of open-ended questions is significantly greater than the number of multiple-choice tests. For example, Lawton (1993) developed an 8-item written test for 6th-grade students consisting of conventional ratio and proportion problems. Allain (2000) developed a valid and reliable proportional reasoning instrument for girls (6-8th-grades) studying at a middle school in North Carolina. The instrument consists of 10 open-ended problems (part-part-whole, associated sets, comparison, missing value, mixture, graphing, and scale problems). Duatepe et al. (2005) prepared a proportional reasoning test consisting of 10 open-ended items (missing value, quantitative comparison, qualitative comparison, non-proportional type relation, and inverse relation) by using the problems in the literature. Akkus and Duatepe-Paksu (2006) developed a measurement tool and rubrics for measuring and evaluating proportional reasoning skills. The test consists of 15 open-ended problems (missing value, quantitative comparison, qualitative comparison, non-proportional type relation, and inverse relation) applied to the 7th-grade and 8th-grade. Pelen and Dinç-Artut (2015) developed a proportional reasoning test consisting of 24 open-ended missing value word problems. Although most of the proportional reasoning tests consist of open-ended problems, there are a few multiple-choice tests to measure middle school students' proportional reasoning skills. According to Bright et al. (2003), different methods are likely to reveal different information about students' proportional reasoning. Therefore, it is also important to use multiple assessment methods, such as multiple-choice and constructed-response. Their test consisted of 4 multiple-choice problems and 1 constructed problem for 8th-grade and 9th-grade students. Arıcan (2019) developed a proportional reasoning test for middle school students consisting of 22 multiple-choice problems.

As mentioned above, to determine the proportional reasoning skills of middle school students open-ended questions are often used, however, researchers have mentioned some disadvantages of using open-ended problems. According to Hilton et al. (2013), open-ended problems are powerful methods for determining students' understanding, but their use may not be practical

in cases when the number of students is high. Reja et al. (2003) stated that with open-ended problems practitioners have the opportunity to discover the answers that students give spontaneously, but open-ended problems have disadvantages such as needing extensive coding when compared with closed-ended problems. Similarly, Hyman and Sierra (2016) emphasized that, in closed-ended problems such as multiple-choice questions, data is coded and analyzed quickly. According to Burton et al. (1991), multiple-choice problems have applicability in measuring higher-order targets such as understanding, application, and analysis. Despite these features, Hyman and Sierra (2016) pointed out that in multiple-choice problems, in-depth information cannot be obtained as students read only a few of the options and choose the option that best represents their views or behaviors. Similarly, Burton et al. (1991) has indicated that multiple-choice tests do not allow students to determine certain learning outcomes such as produce original ideas, organized personal thoughts.

Considering the aforementioned advantages and disadvantages, two-stage diagnostic tests, were developed in the 1980, have the positive aspects of the multiple-choice tests and minimize their disadvantages (e.g., Haslam & Treagust, 1987; Peterson et al., 1986). In two-tier tests, students have two tasks. The first tier includes multiple-choice problems and asks a student to make a choice, and the second tier asks the student justifications for choices made in the first tier (Haja & Clarke, 2011; Tsui & Treagust, 2010). These tests require more time, more effort, and higher-order skills such as reading, thinking, interpreting, understanding skills (Afnia & Istiyono, 2020; Haja & Clarke, 2011). Also, they give students a chance to justify their choice, thus it shifts the focus to the mathematical reasoning process rather than only answering (Haja & Clarke, 2011).

According to Tamir (1990), there are two important reasons /advantages of justifying multiple-choice problems. First, the students who are asked to justify their choices in the multiple-choice section must explain the reasons for their choice by considering all the options. Secondly, the student, who is aware that his/her choice will be justified, tries to learn the topics in-depth and will be ready to write a complete and adequate justification. Thus, a two-tier test is an effective and sensitive way to evaluate meaningful learning (Tamir, 1989, 1990). Despite the mentioned advantages, Haja and Clarke (2011) pointed out that two-tier tests are not widely used in mathematics education. Indeed, when the proportional reasoning tests for middle school students are examined, two-tier tests are found in just a few studies (Haja & Clarke, 2011; Hilton et al., 2013; Mersin, 2018).

Haja and Clarke (2011) aimed to evaluate middle school students' justification skills with two-tier proportional reasoning tasks. The tasks were "select answer" tasks and "marked answer" tasks. The select answer tasks had two types: 1. The student selects the answer, 2. The student selects the answer and writes a justification. Similarly, marked answer tasks had two types: 1. The student selects justification for the marked answer, 2. The student writes a justification for the marked answer. As a result of the study, it is stated that the two-tier tasks give more information about the students' alternative conceptions and reveal the students' reasoning.

Hilton et al. (2013) draw attention to the importance of justifying students' answers to understand the students' proportional reasoning and developed a two-tier multiple-choice diagnostic test for middle school students. The test consists of 12 items related to non-proportional (constant/additive), one or two-dimensional scale, missing value, familiar rate, rate, translation of representations, relative thinking, inverse proportion. Mersin (2018) mentioned that proportional reasoning measuring tools used in Turkey do not allow students to justify their answers and so she translated the two-tier proportional reasoning diagnostic test developed by Hilton et al. (2013) into Turkish. Since the two-tier PRST in the literature is a limited number, and there isn't any two-tier PRST developed in Turkish, it is considered important to develop a two-tier PRST that can allow to justify their answers.

According to Haja and Clarke (2011), in two-tier problems, if students are asked to construct their justification, given tasks become more intellectually demanding and require students to have more sophisticated expression skills. For this reason, the first tier of the test developed in this study is multiple-choice, and the second tier is prepared in an open-ended format in which students can explain their justifications verbally using their expressions. It is believed that determining the proportional reasoning skills of students with a two-tier PRST can benefit researchers, curriculum planners, and teachers. By using PRST, they can determine and understand the proportional reasoning skill levels of students more accurately. Also, they can determine whether students can differentiate between proportional and non-proportional situations.

2. METHOD

2.1. Design

In this research, sequential exploratory mixed-method research was used to develop PRST. The sequential exploratory mixed method is a two-phase model. In the first phase, the researcher studies the subject qualitatively; and in the second phase, s/he continues his/her study quantitatively (Creswell & Plano Clark, 2011). In the qualitative phase of the study, a problem pool consisting of two-tier problems was prepared and experts' opinions were taken for face validity. The content validity was evaluated both qualitative (experts' opinions) and quantitative (Davis's (1992) method) methods. Also, in the quantitative stage, construct validity, discriminant validity, and reliability were tested.

2.2. Study Group

The study group of this research comprised of 391 middle school students studying in two public schools in the south of Turkey. The convenience sampling method was used to determine the study group. Students were studying in schools where the researcher could easily reach in terms of time and place (Cohen et al., 2013). The aim of the study was explained to the students and volunteer students who participated in the study. The PRST comprised ratio and proportion subjects of the 7th-grade mathematics curriculum (MoNE, 2018). So, the research was conducted with 7th (n_{female}=126, n_{male}=97), and 8th (n_{female}=96, n_{male}=72) grade students.

2.3. Procedure

The following stages were followed in the test development process.

2.3.1. *Determining the purpose of the test*

First, the purpose of the test was determined. The purpose of the test is to determine the proportional reasoning skills of 7th and 8th-grade students as valid and reliable.

2.3.2. *Determining the scope and properties of the test and item writing*

For determining the scope of the test, qualitative methods (document review, literature review, expert opinions) were used. Firstly, the Mathematics Curriculum (MoNE, 2018) and mathematics textbooks of middle school were examined. It was determined that the subject of ratio-proportion was taught more in 7th-grade. For this reason, the test items were prepared within the context of 7th-grade learning outcomes. The learning outcomes related to the ratio-proportion at the 7th-grade in the Mathematics Curriculum (MoNE, 2018) are presented below.

Learning Outcome 1: If one of the quantities is 1, it determines the value of the other.
Learning Outcome 2: Given one of the two quantities whose ratio is given to each other, it finds the other.
Learning Outcome 3: Decides whether the two quantities are proportional by examining real-life situations.
Learning Outcome 4: Expresses the relationship between two direct proportional quantities.

Learning Outcome 5: Determines and interprets the proportionality constant of two direct proportional quantities.
Learning Outcome 6: Decides whether two quantities are inverse proportional by examining real-life situations.
Learning Outcome 7: Solves problems related to direct and inverse proportion.

Then the literature was reviewed and it was seen that there were different types of problems in determining proportional reasoning skills. Cramer and Post (1993) categorized proportional reasoning problems into three categories: missing-value problems, numerical comparison problems, and qualitative prediction and comparison problems. Considering the Cramer and Post's (1993) category and learning outcomes related to the ratio-proportion at the 7th-grade, the problem pool consisted of 15 problems, 5 of which was qualitative prediction and comparison, 5 of which was missing value, and 5 of which was numerical comparison. Problems included real-life contexts. The problems were prepared to have two-tier answer options. The first tier consisted of a multiple-choice answer, with four choices. The second tier was the open-ended answer part, which includes justifying the answer given in the multiple-choice section. The PRST is presented in the Appendix.

To determine the discriminant validity of the PRST, its correlation with N-PRST was examined. In this context, a two-tier N-PRST was developed in this study. The N-PRST problem pool consisted of 6 problems including 2 problems on additive situations, 2 problems on linear situations, and 2 problems on constant situations (Van Dooren et al., 2005). Problems included real-life contexts and had two-tier answer options: first-tier multiple-choice answers, second-tier open-ended answers.

Additive: Both Harmankaya Family and Orçan Family have two children. The sum of the Harmankaya Family's ages is 50, while the sum of the Orçan Family's ages is 60. Accordingly, what will be the sum of the Orçan Family's ages when the sum of the Harmankaya Family's ages will be 100?

A) 2 times the sum of the Harmankaya Family's ages.
B) 10+ of the sum of the Harmankaya Family's current ages.
C) 2 times the sum of the Orçan Family's current ages.
D) 50+ of the sum of the Orçan Family's current ages.

Justification:

Linear: Yusuf, who has TL40, starts to save a fixed amount of money every week to buy the computer game he wants. If Yusuf has a total of TL120 at the end of 4 weeks, how much will he has after 8 weeks?

A) 8 times of his initial money
B) 2 times of the money he has at the end of 4 weeks
C) TL240 more than his initial money
D) TL80 more than the money he has at the end of 4 weeks

Justification:

Constant: Baker Mehmet, who has an oven with a maximum capacity of 60 loaves of bread at a time, bakes bread as the number of the orders he receives each time. He started to receive orders as soon as he opens his bakery. First Ahmet orders 20 loaves of bread for his döner shop, then Yusuf wants 40 loaves of bread for his restaurant. Baker Mehmet bakes the first 20 loaves of bread in 12 minutes. In how many minutes does Uncle Mehmet bake the remaining 40 loaves of bread?

A) 24 minutes
B) 12 minutes

C) 36 minutes
D) The information provided is inadequate.

Justification:

2.3.3. *Content and face validity*

The draft PRST and N-PRST were submitted to three experts (one assessment and evaluation expert and two mathematics education experts) to check the face and content validity using an expert opinion form. First, the experts were informed about the content of the PRST (qualitative prediction and comparison, missing value, numerical comparison problems) and N-PRST (additive, constant, and linear problems). The experts assessed the test items in terms of clarity, suitability to the purpose, suitability to the level of 7th-grade and 8th-grade students. Their opinions were evaluated using Davis's (1992) method. According to Davis's (1992) method, each item was evaluated choosing one of these: “a) Appropriate”, “b) Item should be slightly revised”, “c) Item should be reviewed”, “d) Item is not appropriate”. The Content Validity Index (CVI) was calculated for each item using the formula $a+b/n$ (a= number of experts who ticked the “Appropriate”, b= number of experts who ticked the “Item should be slightly revised”, n = number of experts). All items had CVI values of .80 and above, so all were included in the draft test (Davis, 1992). The opinions of 2 master students who were mathematics teachers were also taken about in terms of clarity, suitability to the purpose, suitability to the level of 7th-grade and 8th-grade students. Next, the draft tests were applied to 7th-grade (n= 19) students as a pre-application to determine the comprehensibility and language suitability. Considering the opinions of the experts and students, the necessary arrangements were made.

2.3.4. *Pilot study and scoring the answers*

In the pilot study, draft tests were applied to 391 middle school students and then, the student answers were scored.

2.3.4.1. Scoring the Answers of PRST. In the multiple-choice answer tier, the correct answer is 1 point and the wrong answer is 0 point. For scoring the open-ended tier, the rubrics developed by Akkus and Duatepe-Paksu (2006) were edited and used. Since there are different types of problems in the PRST, two rubrics (a rubric for the items of qualitative prediction and comparison, a rubric for the items of missing value and numeric comparison) are used. For open-ended answers, the lowest score is 0 points and the highest score is 3 points. Consequently, in the two-tier PRST, the lowest score is 0 points and the highest score is 4 points. The rubrics items and points are explained in Table 1.

Table 1. *Rubrics items.*

Problem Type	Point	Items
Missing value/Numerical Comparison Problems	0	No answer. There is no clue that proportional reasoning exists. The proportion is established between the wrong variables. There is an additive comparison of data. There is a random use of numbers and operations.
	1	Only the correct answer is given, there is not any mathematical operation. There are some clues that proportional reasoning exists.
	2	There is proportional reasoning among the expected variables, but there is a calculation error or the correct answer isn't provided.
	3	There is proportional reasoning to solve the problem correctly, and the correct answer is given.
Qualitative Prediction and Comparison Problems	0	No answer There is no clue that proportional reasoning exists.
	1	There are some clues that proportional reasoning exists.
	2	There is proportional reasoning to solve the problem correctly. The description is done using the root of the problem.
	3	There is proportional reasoning to solve the problem correctly. The correct answer is given with original sentences, and the explanations are enriched with methods such as forming shapes, drawing, giving examples.

2.3.4.2. Scoring the Answers of N-PRST. N-PRST consists of two-tier answer options. In multiple-choice answer options, the correct answer is 1 point and the wrong answer is 0 point. The rubric is used to score open-ended answers. Rubric items and points are explained in Table 2.

Table 2. *Rubrics items.*

Problem Type	Point	Items
Non-Proportional Reasoning Skills Problems	0	No answer There is proportional reasoning among the variables. Non-proportional situations are indistinguishable. Multiplication strategy is applied to a constant, additive, or linear situation.
	1	There are clues that non-proportional situations are distinguished from proportional situations. There is an additive comparison of data.
	2	Non-proportional situations are distinguished from proportional situations but the explanation is insufficient or made by using the problem root.
	3	Non-proportional situations are distinguished from proportional situations. The correct answer is given by using original sentences, and the explanations are enriched with methods such as forming shapes, drawing, or giving examples.

The open-ended answers of PRTS and N-PRST were scored by the researcher using the rubrics mentioned above. To ensure the interrater reliability, randomly selected 50 students' answers were scored by a second-rater. Spearman rho correlation coefficient was used to determine the relationship between the scores of the two raters. As a result of the analysis, it was seen that the

Spearman rho correlation of items coefficients ranged between .984 and .999. Also, the Wilcoxon Sign Test was used to determine whether there was a significant difference between the points given by the raters. Wilcoxon Sign Test results showed that there was no significant difference between the raters' scorings ($p > .05$).

2.3.5. *Construct validity, discriminant validity and reliability studies*

2.3.5.1. Proportional Reasoning Skill Test. In this study, to determine the construct validity of the PRST, Confirmatory Factor Analysis (CFA) was performed. Also, for the construct validity, corrected item-total correlations were calculated and the discrimination of the items was investigated by examining the differences between the lower 27% and upper 27% groups. To determine the reliability of the test, Cronbach alpha and composite reliability coefficients were calculated. Before the data analyses, the data set for 391 cases checked to ensure the normal distribution assumption. For all items, kurtosis and skewness values were calculated. For the items (except for items 8, 13, and 15), kurtosis and skewness values were found to be between ± 2. These values indicated that the items had a normal distribution (Cameron, 2004). However, it was determined that item 8 (skewness= 3.061, kurtosis= 10.889) about numerical comparison and item 13 (skewness= 2.806, kurtosis=9.233) and item 15 (skewness= 2.583, kurtosis= 5.829) about missing value didn't have the normal distribution. For this reason, the data related to these items were excluded from the data set and not included in the analysis. For 12 items, z scores were between ± 3.29 showed the test had univariate normality. According to Mahalanobis distance values, there were no multivariate outlier values and the data set had multivariate normality ($p<.001$ for the χ^2) (Tabachnick & Fidell, 2013). The correlation matrix for all items was examined and coefficients were found between .30 and .90 for all cases. These values showed that there were not singularity and multicollinearity problems. While anti-image correlation coefficients for each item (r=.863 to .929) were adequate for sampling adequacy of individual items (Field, 2009; Tabachnick & Fidell, 2013), results of KMO statistics (KMO=.897) and Bartlett Sphericity Test (χ^2 = 1290.527, df = 66, p= .000 <.05) proved the sampling adequacy of data set. Also, when comparing the scores of the lower 27% and upper 27% groups (n= 106), the normality of each item (n= 12) was examined in terms of the group variable. As a result, for all items, it was determined that the skewness and kurtosis values were outside the ± 2 range. Therefore, differences between groups were investigated by using the Mann-Whitney U test which is one of the nonparametric tests.

To test the discriminant validity of PRST, the relationship between the scores obtained from the PRST and the scores obtained from the N-PRST was examined. For N-PRST scores, skewness was calculated as = 1.334 and kurtosis as= 1.403 while skewness was calculated as = .545, and kurtosis as =-. 675 for the PRST scores. The skewness and kurtosis values showed that data sets were close to the normal distribution (Cameron, 2004). Accordingly, the Pearson Correlation Coefficient was calculated to examine the relationship between the scores. After determining the construct validity of the two-tier test with CFA, the item statistics of the first multiple-choice tier were calculated. Item analysis for the multiple-choice tier was done with the Test Analysis Program (TAP). Item discrimination index, item difficulty index, and item-total correlation were calculated for the construct validity of the multiple-choice tier.

2.3.5.2. Non-Proportional Reasoning Skill Test. In this study, to determine the construct validity of the test, the CFA was performed. Also, for the construct validity, corrected item-total correlations were calculated and the discrimination of the items was studied by examining the differences between the lower 27% and upper 27% groups. To determine the reliability of the test, Cronbach alpha internal consistency coefficients were calculated. All items (n= 6) have a normal distribution (kurtosis and skewness values found to be between ±2 (Cameron, 2004)). z scores (between ± 3.29) and Mahalanobis distance values ($p<.001$ for the χ2) showed that the test had univariate and multivariate normality (Tabachnick & Fidell, 2013).

Correlation coefficients for all items were between .30 and .90. Thus, singularity and multicollinearity problems weren't determined. Results of KMO statistics (KMO=.705) and Bartlett Sphericity Test (χ^2= 287.427; *df*= 15; *p*= .00 <.05) showed the sampling adequacy of data set, and anti-image correlation coefficients for each item (*r*= .677 to .757) were adequate for sampling adequacy of individual items (Field, 2009; Tabachnick & Fidell, 2013).

3. RESULT / FINDINGS

3.1. Results of Non-Proportional Reasoning Skill Test

To determine the construct validity of the N-PRST, CFA was performed. The data set (*n*=391) was transferred to the LISREL program and a covariance matrix was prepared. It was determined that t values were between 5.71 and 9.49 and significant (*p*<.01) for all values. Then the goodness of fit values and modification suggestions were examined. According to these suggestions, error variances of item1-item2 and item2-item5 were linked. The goodness of fit values for pre-modification and post-modification are shown in Table 3.

Table 3. *The goodness of fit values.*

The goodness of fit values	Pre-modification	Post-modification
p	.00*	.00*
χ^2/df	52.15/9=5.79	26.06/7=3.72
RMSEA	.111	.08
SRMR	.065	.047
GFI	.96	.98
AGFI	.90	.93
CFI	.85	.93
NFI	.83	.91

*p<.05

As seen in Table 3, after modification goodness of fit values were χ^2/df = 3.72 (26.06/7), RMSEA = .08, SRMR = .047, GFI = .98, AGFI =. 93, CFI = .93, NFI = .91. Figure 1 shows the path diagram for the final model. According to Figure 1, the standardized factor loadings vary between .35 and .59.

Figure 1. *Path diagram of the model for non-proportional reasoning two-tier test.*

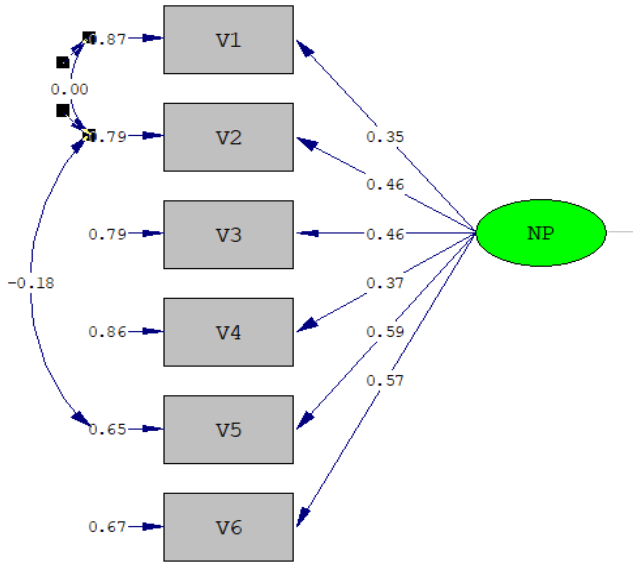


Table 4. *Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlations.*

Items	Group	Mean Rank	Sum of Ranks	*U*	*p*	Corrected Item-Total Cor.
Item 1	Lower	81.50	10758.00	1980.00	.00*	.563
	Upper	166.82	17683.00			
Item 2	Lower	81.33	10735.00	1957.00	.00*	.555
	Upper	167.04	17706.00			
Item 3	Lower	80.38	10609.50	1831.50	.00*	.543
	Upper	168.22	17831.50			
Item 4	Lower	101.42	13387.50	4609.50	.00*	.381
	Upper	142.01	15053.50			
Item 5	Lower	86.97	11480.00	2702.00	.00*	.503
	Upper	160.01	16961.00			
Item 6	Lower	93.50	12342.00	3564.00	.00*	.434
	Upper	151.88	16099.00			

*p<.05

As seen in Table 4, it was determined that there were statistically significant differences between the upper 27% and lower 27% groups for all items, and corrected item-total correlations ranged from .381 and .563. Also, as a result of the reliability analysis, the Cronbach alpha reliability and composite reliability coefficients of the test were .643 and .845.

3.2. Results of Two-Tier Proportional Reasoning Skill Test

3.2.1. *Construct validity*

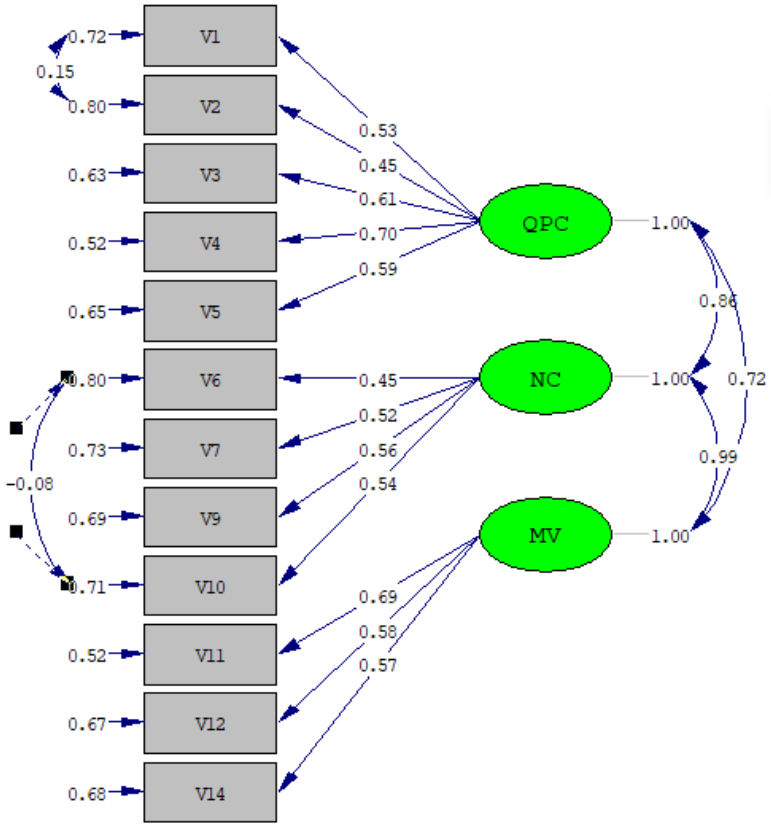
In this study, the CFA was performed to test the 3-factor structure of the PRST. Firstly the data set obtained from 391 students was transferred to the Lisrel program, and a covariance matrix was prepared. For the model, t values were between 8.07 and 13.92 and statistically significant (p<.01). The CFA analysis computed a significant p-value (χ^2 = 96.23, p= .00013 <.05) for the model. So the goodness of fit values and modification suggestions were examined. According to these suggestions, error variances of item1-item2 and item6-item10 were linked. The goodness of fit values for pre-modification and post-modification are shown in Table 5 and Figure 2 shows the path diagram for the model.

Table 5. *The goodness of fit values.*

The goodness of fit values	Pre-modification	Post-modification
p	.00013*	.00366*
χ^2/df	96.23/51=1.87	79.67/49=1.63
RMSEA	.048	.040
SRMR	.045	.040
GFI	.96	.97
AGFI	.94	.95
CFI	.98	.98
NFI	.96	.96

*p<.05

Figure 2. *Path diagram of the model for proportional reasoning two-tier test.*



As seen in Figure 2, the standardized factor loadings for qualitative prediction and comparison factor varied between .45 and .70, for numerical comparison factor between .45 and .56, and for missing value factor the standardized loadings between .57 and .69. So other goodness of fit values were also examined. As seen in Table 5, goodness of fit values in CFA after modification were χ^2 / *df* = 1.63, RMSEA = .040, SRMR = .040, GFI = .97, AGFI =. 95, CFI = .98, NFI = .96.

Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlation coefficients are presented in Table 6. As seen in Table 6, the corrected-item total correlations ranged from .446 and .592. Also, it was determined that there were statistically significant differences between the upper 27% and lower 27% groups for all items.

Table 6. *Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlations.*

Items	Group	Mean Rank	Sum of Ranks	*U*	*p*	Corrected Item-Total Correlation
Item 1	Lower	65.58	7082.50	1196.500	.00*	.504
	Upper	152.62	16788.50			
Item 2	Lower	68.48	7395.50	1509.500	.00*	.446
	Upper	149.78	16475.50			
Item 3	Lower	64.25	6939.50	1053.500	.00*	.569
	Upper	153.92	16931.50			
Item 4	Lower	60.65	6550.00	664.000	.00*	.568
	Upper	157.46	17321.00			
Item 5	Lower	65.91	7118.00	1232.000	.00*	.518
	Upper	152.30	16753.00			
Item 6	Lower	76.15	8224.50	2338.500	.00*	.455
	Upper	142.24	15646.50			
Item 7	Lower	76.34	8244.50	2358.500	.00*	.518
	Upper	142.06	15626.50			
Item 9	Lower	61.71	6664.50	778.500	.00*	.528
	Upper	156.42	17206.50			
Item 10	Lower	79.52	8588.00	2702.000	.00*	.543
	Upper	138.94	15283.00			
Item 11	Lower	73.57	7945.50	2059.500	.00*	.493
	Upper	144.78	15925.50			
Item 12	Lower	62.19	6717.00	831.000	.00*	.592
	Upper	155.95	17154.00			
Item 14	Lower	74.28	8022.00	2136.000	.00*	.526
	Upper	144.08	15849.00			

*$p<.05$

3.2.2. *Discriminant validity*

To test the discriminant validity of PRST, the relationship between the PRST scores and N-PRST scores was examined. As a result of the Pearson Correlation Test, it was determined that the relationship coefficient was r=.683 and statistically significant (p= .00 < 0.05).

3.2.3. *Reliability analysis*

According to reliability analysis, Cronbach alpha values were α = 0.748 for the qualitative prediction and comparison factor, α = 0.631 for numerical comparison factor, and α = 0.651 for missing value factor. For the total, the Cronbach alpha reliability coefficient was calculated as α = 0.849. Also, the composite reliability coefficient of the test was 0.656.

3.2.4. *Test Statistics*

The test statistics of PRST are shown in Table 7.

Table 7. *Test statistics of PRST.*

Factor	Mean	Standard Deviation
QPC	3.03	1.83
NV	1.47	1.35
MV	1.58	1.62
Total	2.15	1.40

As seen in Table 7, it could be said the students' PRST level was medium.

3.3. Construct Validity and Reliability of Multiple-Choice Test

Item analysis and test statistics results of the multiple-choice test are shown in Table 8 and Table 9.

Table 8. *Item analysis results.*

Item No	Item Discrimination Index (*r*)	Item Difficulty Index (*p*)	Point Biserial
01	.64	.44	.56
02	.64	.35	.57
03	.71	.37	.62
04	.77	.45	.63
05	.55	.33	.55
06	.37	.15	.53
07	.48	.19	.57
09	.62	.25	.53
10	.62	.30	.62
11	.45	.16	.62
12	.35	.13	.58
14	.31	.13	.50

Table 9. *Test statistics.*

Statistics	Value
Mean	2.951
Standard Deviation	2.951
Mean Item Difficulty	.278
Mean Discrimination Index	.545
KR-20	.810
Mean Point Biserial	.573

The item discrimination indices ranged between .31 and .77, the item difficulty indices ranged between .13 and .45, and the point biserial coefficients ranged between .50 and .63. According to test statistics, test discrimination was very good (Ebel, 1965; Wells & Wollack, 2003), but the test was difficult (Crocker & Algina, 2008). Also, the KR-20 value showed the multiple-choice test had good reliability (Kline, 2011; Rudner & Schafer, 2002; Wells & Wollack, 2003).

4. DISCUSSION and CONCLUSION

This research aims to develop a useful, valid, and reliable two-tier PRST for middle school 7th and 8th-grade students. The test includes problems that measure qualitative prediction and comparison, missing value, and numerical comparison (Cramer & Post, 1993). The study was designed with the sequential exploratory mixed-method approach, in which qualitative and quantitative research methods were used. Firstly, a problem pool consisting of 15 problems (5 of which was qualitative prediction and comparison, 5 of which was missing value, and 5 of which was numerical comparison) was prepared. The first tier consisted of a multiple-choice problem, with four choices. The second tier was the open-ended answer part, which included explaining and justifying the answer given multiple-choice tier. The two rubrics (a rubric for the qualitative prediction and comparison problems, and a rubric for the missing value and numeric comparison problems) were used for scoring the open-ended tier. The rubrics items

were adapted from the study of Akkus and Duatepe-Paksu (2006). In the two-tier test, the lowest score was 0 and the highest score was 4 points for each problem.

In this study, the face, content, construct, discriminant validity of the PRST were tested. The content and face validity of the test were provided with assessment and evaluation (n=1) and mathematics education (n=2) experts' opinions (Gable, 1986). For the construct validity studies of the test, firstly CFA was performed. According to Mueller and Hancock (2001), the main advantage of CFA is that it enables researchers to bridge between theory and observation. CFA facilitates testing the relationship between the latent constructs (QPC, NC, MV) and observed variables (Suhr, 2006).

In this study, CFA was carried out to test a 3-factor structure with the data set consisting of 12 problems which ensured the univariate and multivariate normal distribution assumptions. In the factor analysis, the standardized factor loadings values of .40 and above are accepted as meaningful load values (Gable, 1986; Hatcher, 1994). Hair et al. (2014) expressed that factor loadings values of .30 and above are accepted practically significant for sample sizes of 350 or greater. Based on CFA results, it could be said that the standardized factor loadings were meaningful. According to commonly agreed goodness of fit values criteria (e.g., Brown, 2006; Hair et al., 2014; Hu & Bentler, 1999; Tabachnick & Fidell, 2013), χ^2 / df <2, RMSEA, SRMR <.05, GFI, AGFI, CFI, NFI > .90 values are acceptable values, χ^2 / df <5, RMSEA, SRMR <.08, GFI, AGFI, CFI, NFI > .95 values are excellent values. Accordingly, when the values calculated as a result of the analysis are taken into consideration, it could be said that the χ^2 / df, RMSEA, SRMR, GFI, AGFI, CFI, NFI are excellent values. Also, Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups scores and corrected item-total correlations showed that the problems tended to measure the same skill and discrimination indexes were high (Büyüköztürk, 2010). According to these results, it could be said that the two-tier PRST has construct validity. Also, item analysis results show, multiple-choice tier has very good discrimination (Ebel, 1965; Wells & Wollack, 2003) and good reliability (Kline, 2011; Rudner & Schafer, 2002; Wells & Wollack, 2003). But according to mean item difficulty, the multiple-choice test is difficult (Crocker & Algina, 2008).

To determine the discrimination of the test, the Pearson Correlation Coefficient was calculated between the PRST scores and the N-PRST score. The N-PRST was developed in this study to determine the PRST's discriminant validity. The two-tier N-PRST consisted 6 problems: additive situations (n=2), linear situations (n=2), and constant situations (n=2). The validity and reliability studies showed that N-PRST was useful, valid (according to results of CFA, Mann-Whitney U test results for the comparison of the upper 27% and lower 27% groups and corrected item-total correlations), and reliable. Cohen (1988) interpreted the strength of relationship as; .10-.29 "small", .30-.49 "medium" and .50-1.0 "large". As a result of the analysis, it was determined that there was a statistically significant large relationship between proportional and non-proportional test scores. The positive relationship could be considered as evidence that students can distinguish between proportional and non-proportional situations, and the discrimination of the PRST is high. According to these results, it could be said that the discriminant validity of the test was ensured (Fornell & Larcker, 1981).

Kline (2011) states that the reliability coefficient is excellent if it is around .90, very good around .80, sufficient around .70, and insufficient under .50. Rudner and Schafer (2002) stated that the 0.50 or 0.60 reliability coefficient for the tests performed in the classroom can be seen as sufficient. Ebel and Frisbie (1991) stressed that when it comes to ratings for a group of people, such as the classroom, the reliability coefficient should be 0.65. The Cronbach Alpha coefficient of the test is very good and the composite reliability coefficient is acceptable.

As a conclusion, it can be said that the two-tier PRST is useful, valid, and reliable to measure middle school students' proportional reasoning skills. The psychometric properties of the test

developed in this study are tested using the data obtained from the middle school 7th and 8th-grade students. Researchers can study further the psychometric properties of the PRST for students who graduated from middle school. It is expected that this valid and reliable PRST will encourage other researchers to study the relationships of proportional reasoning skills with different variables (e.g., academic achievement or attitude on ratio-proportion). This research was conducted with students selected by convenience sampling. Although the test is determined as valid and reliable, to increase the generalizability of the test, it is recommended to examine the psychometric properties of the test with a group determined by random sampling.

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).

Authorship contribution statement

Kübra Açıkgül: Investigation, Resources, Introduction, Methodology, Software, Analysis, Findings, Discussion, Supervision, and Validation, Writing original draft.

ORCID

Kubra ACIKGUL https://orcid.org/0000-0003-2656-8916

5. REFERENCES

Afnia, P.N., & Istiyono, E. (2020, February). The development of two-tier multiple choice instruments to measure higher order thinking skills bloomian. In *3rd International Conference on Learning Innovation and Quality Education (ICLIQE 2019)* (pp. 1038-1045). Atlantis Press. https://doi.org/10.2991/assehr.k.200129.128

Akkus, O., & Duatepe-Paksu, A. (2006). Construction of a proportional reasoning test and its rubrics. *Eurasian Journal of Educational Research, 25*, 1-10.

Alfieri, L., Higashi, R., Shoop, R., & Schunn, C. D. (2015). Case studies of a robot-based game to shape interests and hone proportional reasoning skills. *International Journal of STEM Education*, *2*(4), 1-13. https://doi.org/10.1186/s40594-015-0017-9

Allain, A. (2000). *Development of an instrument to measure proportional reasoning among fast-track middle school students*. [Master's thesis]. University of North Carolina State.

Arıcan, M. (2019). A diagnostic assessment to middle school students' proportional reasoning. *Turkish Journal of Education*, *8*(4), 237-257. https://doi.org/10.19128/turje.522839

Ayan, R., & Isiksal-Bostan, M. (2019). Middle school students' proportional reasoning in real life contexts in the domain of geometry and measurement. *International Journal of Mathematical Education in Science and Technology*, *50*(1), 65-81. https://doi.org/10.1080/0020739X.2018.1468042

Behr, M., Lesh, R., & Post, T. (1988). Proportional reasoning, In M. Behr and J. Hiebert (Eds.), *Number concepts and operations in the middle grades.* Lawrence Erlbaum Associates.

Behr, M., Harel, G., Post, T., & Lesh, R. (1992). Rational number, ratio and proportion. In D. Grouws (Eds.), *Handbook on research of teaching and learning* (pp. 296-333). McMillan.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. In David A. Kenny (Eds.), *Methodology in the Social Sciences.* The Guilford Press.

Bright, G. W., Joyner, J. M., & Wallis, C. (2003). Assessing proportional thinking. *Mathematics Teaching in the Middle School, 9*(3), 166-172. https://www.jstor.org/stable/41181882

Burton, S. J., Sudweeks, R. E., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University

Testing Services and The Department of Instructional Science. https://testing.byu.edu/handbooks/betteritems.pdf

Büyüköztürk, S. (2010). *Sosyal bilimler için veri analizi el kitabı* [Data analysis handbook for social sciences]. Pegem Akademi.

Cameron, A. (2004). Kurtosis. In M. Lewis-Beck, A. Bryman and T. Liao (Eds.). *Encyclopedia of social science research methods.* (pp. 544-545). SAGE Publications, Inc.

Common Core State Standards Initiative (US). Common core state standards for mathematics. http://www.corestandards.org/Math/

Cramer, K., & Post, T. (1993). Proportional reasoning. *The Mathematics Teacher*, *86*(5), 404-407. https://www.jstor.org/stable/27968390

Cramer, K., Post, T., & Currier, S. (1993). Learning and teaching ratio and proportion: research implications. In D. Owens (Eds.), *Research ideas for the classroom* (pp. 159-178). Macmillan Publishing Company.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education*. Routledge.

Creswell, J .W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Sage Publications.

Davis, L. L. (1992). Instrument review: getting the most from a panel of experts. *Applied Nursing Research*, *5*, 194-197. https://doi.org/10.1016/S0897-1897(05)80008-4

Dinç-Artut, P., & Pelen, M. S. (2015). 6th grade students' solution strategies on proportional reasoning problems. *Procedia-Social and Behavioral Sciences*, *197*, 113-119. https://doi.org/10.1016/j.sbspro.2015.07.066

Duatepe, A., Akkus-Cıkla, O., & Kayhan, M. (2005). An investigation on students' solution strategies for different proportional reasoning items. *Hacettepe Journal of Education Faculty*, *28*, 73-81. https://dergipark.org.tr/en/pub/hunefd/issue/7808/102422

Ebel, R. L. (1965). Measuring educational achievement. Prentice Hall.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice Hall.

Field, A. (2009). *Discovering statistics using SPSS*. Sage Publication.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research, 18*(3), 328–388. https://doi.org/10.1177/002224378101800313

Gable, R. K. (1986). *Instrument development in the affective domain.* Kluwer-Nijhoff Publishing.

Hair, J. F., Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2014). *Multivariate data analysis.* Pearson New International Edition.

Haja, S., & Clarke, D. (2011). Middle school students' responses to two-tier tasks. *Mathematics Education Research Journal*, *23*(1), 67-76. https://doi.org/10.1007/s13394-011-0004-5

Haslam, F., & Treagust, D. F. (1987). Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. *Journal of Biological Education*, *21*(3), 203-211. https://doi.org/10.1080/00219266.1987.9654897

Hatcher, L. (1994). A *step-by-step approach to using the SAS® system for Factor Analysis and Structural Equation Modeling*. SAS Institutte, Inc.

Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle - years students' proportional reasoning. *Mathematics Education Research Journal*, *25*(4), 523-545. https://doi.org/10.1007/s13394-013-0083-6

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journ*al, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hyman, M. R., & Sierra, J. J. (2016). Open-versus close-ended survey problems. *Business Outlook, 14*(2), 1-5.

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

Lamon, S. J. (2007). Rational numbers and proportional reasoning: Toward a theoretical framework for research. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 629–668). Information Age Publishing.

Lawton, C. A. (1993). Contextual factors affecting errors in proportional reasoning. *Journal for Research in Mathematics Education, 24*(5), 460-466. https://doi.org/10.2307/749154

Lesh, R., Post, T., & Behr, M. (1988). Proportional reasoning. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 93-118). Lawrence Erlbaum & National Council of Teachers of Mathematics.

Lim, K. (2009). Burning the candle at just one end: Using nonproportional examples helps students determine when proportional strategies apply. *Mathematics Teaching in the Middle School, 14*(8), 492–500. https://doi.org/10.5951/MTMS.14.8.0492

Mersin, N. (2018). An evaluation of proportional reasoning of middle school 5th, 6th and 7th grade students according to a two-tier diagnostic test. *Cumhuriyet International Journal of Education, 7*(4), 319–348. https://doi.org/10.30703/cije.4266271

Ministry of National Education [MoNE], (2018). Matematik dersi öğretim programı. (İlkokul ve Ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. Sınıflar) [Mathematics curriculum. (Primary and Middle Schools 1, 2, 3, 4, 5, 6, 7 and 8th Grades) [Middle school mathematics curricula for grades 5, 6, 7, and 8)]. https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=329

Mueller, R. O., & Hancock, G. R. (2001). Factor analysis and latent structure: Confirmatory factor analysis. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 5239-5244). Pergamon.

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and Standards for School Mathematics*. National Council of Teachers of Mathematics.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended problems. *Canadian Journal of Experimental Psychology, 67*(3), 215-227. https://doi.org/10.1037/a0032918

Özgün-Koca, S. A., & Altay, M. K. (2009). An investigation of proportional reasoning skills of middle school students. *Investigations in Mathematics Learning, 2*(1), 26-48. https://doi.org/10.1080/24727466.2009.11790289

Pelen, M. S., & Dinç-Artut, P. (2015). 7th grade students'problem solving success rates on proportional reasoning problems. *The Eurasia Proceedings of Educational and Social Sciences, 2*, 96-100. https://doi.org/10.21890/ijres.71245

Peterson, R. F., Treagust, D. F., & Garnett, P. (1986). Identification of secondary students' misconceptions of covalent bonding and structure concepts using a diagnostic test instrument. *Research in Science Education, 16*, 40-48. https://doi.org/10.1007/BF02356816

Post, T. R., Behr, M. J., & Lesh, R. (1988). Proportionality and the development of pre-algebra understandings. In A. Coxford & A. Shulte (Eds.), *The ideas of algebra, K-12* (pp. 78-90). National Council of Teachers of Mathematics.

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. In A. Ferligoj & A. Mrvar (Eds.), *Developments in applied statistics* (pp. 159–177). FDV.

Rudner, L. M., & Shafer, W. D. (2002). *What teachers need to know about assessment.* National Education Association.

Singh, P. (2000). Understanding the concepts of proportion and ratio constructed by two grade six students. *Educational Studies in Mathematics, 43,* 271-292. https://doi.org/10.1023/A:1011976904850

Soyak, O., & Isiksal, M. (2017, February). Middle school students' difficulties in proportional reasoning. Paper Presented at the *CERME 10,* Dublin, Ireland.

Suhr, D. (2006). Exploratory or confirmatory factor analysis. *SAS Users Group International Conference* (pp. 1- 17). SAS Institute, Inc.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Pearson.

Tamir, P. (1989). Some issues related to the use of justifications to multiple-choice answers. *Journal of Biological Education*, *23*(4), 285-292. https://doi.org/10.1080/00219266.1989.9655083

Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education*, *12*(5), 563-573. https://doi.org/0.1080/0950069900120508

Tourniaire, F., & Pulos, S. (1985). Proportional reasoning: A review of the literature. *Educational Studies in Mathematics*, *16*(2), 181-204. https://doi.org/10.1007/BF02400937

Tsui, C. Y., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, *32*(8), 1073-1098. https://doi.org/10.1080/09500690902951429

Van De Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2013). *Elementary and middle school mathematics: teaching developmentally*. Pearson Education, Inc.

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Not everything is proportional: Effects of age and problem type on propensities for over generalization. *Cognition and Instruction*, *23*(1), 57-86. https://doi.org/10.1207/s1532690xci2301_3

Van Dooren, W., De Bock, D., & Verschaffel, L. (2010). From addition to multiplication and back: The development of students' additive and multiplicative reasoning skills. *Cognition and Instruction, 28*, 360–381. https://doi.org/10.1080/07370008.2010.488306

Weinberg, S. L. (2002). Proportional reasoning: One problem, many solutions! In B. Litwiller (Eds.), *Making sense of fractions, ratios, and proportions: 2002 year book* (pp. 138-144). National Council of Teachers of Mathematics.

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability. Testing & Evaluation Services*. University of Wisconsin. http://testing.wisc.edu/Reliability.pdf.