IJETS@R

(ISSN: 2587-0238)

Gürkan, A, K. & Ulu, M. (2023). The Valıdıty And Relıabılıty Study Of Elementary Fourth Grade Number Sense Scale, *International Journal of Education Technology and Scientific Researches, 8(22), 883-909.*

DOI: http://dx.doi.org/10.35826/ijetsar.594

Article Type: Research Article

THE VALIDITY AND RELIABILITY STUDY OF ELEMENTARY FOURTH GRADE NUMBER SENSE SCALE[1]

Ayşe Kübra GÜRKAN
Teacher, Ministry of Education, İstanbul, Turkey, kgurkan9515@gmail.com
ORCID: 0000-0002-5972-0008

Mustafa ULU
Assoc. Prof. Dr, Kütahya Dumlupinar University, Kütahya, Turkey, mustafa.ulu@dpu.edu.tr
ORCID: 0000-0002-3961-1533

Received: 14.01.2023 *Accepted: 20.05.2023* *Published: 01.06.2023*

ABSTRACT

In this study, it was aimed to develop a valid and reliable number sense scale for elementary school students. Construct validity studies were conducted with 299 elementary school 4th grade students studying in 2 public schools in Eskişehir province selected by cluster sampling method. Criterion validity studies were conducted with 312 elementary school 4th grade students studying in 4 public schools in Eskişehir. As a result of the literature review and expert opinions, it was decided that the theoretical structure of the scale would consist of 6 factors and 31 items, namely the meaning and size of numbers, decomposing and combining numbers, determining reference points, the effect of operations on numbers, the flexible use of numbers, and the sensibility of numbers. Item analysis studies were first conducted on the draft scale and it was decided to remove 6 items with low discrimination from the test. Following the item analysis studies, construct validity studies were conducted. Within the scope of construct validity, it was aimed to test the theoretical structure created by the researchers. In this context, confirmatory factor analysis (CFA) technique was thought to be more appropriate for the purpose. As a result of CFA, it was determined that the theoretical structure of the scale produced a high fit with its 6-factor 25-item structure. It was determined that the reliability coefficients of the whole scale and each factor were sufficient. Finally, within the scope of criterion validity, the relationship of the number sense scale with the problem solving achievement test developed by Ulu (2017) was examined. As a result, it was seen that all factors of the number sense scale had a moderately significant relationship with problem solving success, and this finding allowed us to see that the criterion validity of the number sense scale was sufficient.

Keywords: Elementary, number sense scale, validity and reliability.

[1] This article is produced by first author's master thesis, under supervision of second author

INTRODUCTION

Numbers are one of the most important skills of mathematics. Individuals are introduced with numbers at an early age. Children are first introduced to numbers at the age of two by using ordinal counting skills. At the age of four, they begin to solve problems by gaining the cardinal value of numbers (Olkun, Fidan & Babacan-Özer, 2013; Griffin ve diğerleri, 1994; Griffin & Case 1997). The use of numbers starts at an early age and continues throughout life (Reys, Reys, Nohda & Emori, 1995, Yazgan, Bintaş & Altun, 2002, Gülbağcı-Dede & Şengül, 2016, Çavuş-Erdem ve Duran, 2015). Numbers are a necessary component of other areas of mathematics such as geometry, measurement, algebra, data analysis (Birgin ve Peker, 2022; Li ve Yang, 2010; Çekirdekçi, Şengül ve Doğan, 2017, Jordan, Glutting ve Ramineni, 2010). For this reason, mathematics education programs emphasize the development of individuals who use numbers effectively both in daily life and in mathemetic learning areas (CNIM Circulum, 2001; National Council of Teachers of Mathematics, 2000; Australian Education Council, 1991, Milli Eğitim Bakanlığı, 2005). The effective use of numbers is defined as number sense (Griffin, Case & Siegler, 1994; Griffin & Case 1997, NCTM, 2000; Nickerson & Whitacre; 2008, Russell, 2000).

It is seen that there is no common definition of number sense in litarature. Greeno (1991) defined number sense as flexible thinking ability with numbers, using mental estimation strategies and reasoning about numerical quantities. Reys et al. (1999) define number sense as the ability to develop appropriate strategies to the situation and to use numbers and operations flexibly. Kalchman, Moss, and Case (2001) defined number sense as estimating the given quantity, recognizing inconsistency in the results, flexible calculation, and making connections between different representations of the number. Yang (2019) defines number sense as the ability to deal with problems encountered in daily life by developing flexible and effective strategies. Based on the definitions, it can be said that individuals with high number sense use numbers and strategies flexibly.

Different studies results confirm the definitions of number sense (Alsawaie, 2011; Harç, 2010, Şengül & Gülbağcı-Dede, 2012; Çekirdekçi, Şengül & Doğan, 2016; Jordan, Glutting & Ramineni, 2009, Mohamed & Johnny, 2010; Çontay & İymen, 2011). In these studies, it has been observed that individuals with high number sense can use numbers more flexibly and develop different strategies in the solution process compared to individuals with low number sense. In some studies, it was found that there was a positive relations between number sense and mathematics achievement (Akkaya, 2016; Çekirdekçi, Şengül, & Doğan, 2016; Harç, 2010; Kayhan Altay, 2010; Mohamed & Johnny, 2010; Tunalı, 2018), while in some studies, number sense positively affected mathematics achievement (Olkun, Mutlu, & Sarı, 2017; Reys & Yang, 1998). Another remarkable result of the early researches are number sense skills which require strategic thinking are used much less than rule-based solutions (Alsawaie, 2011; Harç, 2010, Şengül & Gülbağcı-Dede, 2012; Çekirdekçi, Şengül & Doğan, 2016; Jordan, Glutting & Ramineni, 2009, Mohamed & Johnny, 2010; Çontay & İymen, 2011). Different studies indicate that rule-based solutions, even if they provide the correct answer, prevent the development of higher level mathematical skills such as questioning and reasoning (Baki & Kartal; 2004, Anderson, 2010; Brynes & Wasik 1991).

In the literature, like in its definition, there is no common consensus on the components of number sense. Greeno (1991) stated that number sense skills consist of three dimensions: flexible calculation, numerical estimation and numerical reasoning. McIntosh et al. (1992) stated that the dimensions of number sense skills consist of three dimensions: the concept of numbers, operations with numbers, and simple calculation methods with numbers and operations. Yang (1995) stated that number sense consists of five dimensions: comprehension of the meaning of numbers, decomposition and recombination, size of numbers, comparison and use of reference points, and flexibility in calculation. Reys and Yang (1998) stated that number sense consists of six dimensions: comprehending the meaning and size of numbers, using similar representations of numbers, effects and meanings of operations, using reference points, and flexibility in mental and written calculation.

In Turkey, scales have been developed at different levels to measure number sense skills. One of the scale was developed by Dede and Şengül (2016) to determine the number sense of pre-service mathematics teachers. The structure of the scale was formed based on the model developed by Yang (1995). The scale consisted of 31 items in 4 factors: the meaning of numbers, the magnitude of numbers, flexible operation and judging the sensibility of the result, and estimation. The factor structure of the scale was decided based on expert opinion. The validity and reliability of the scale was determined by KR-20 coefficient and item analysis was conducted.

As a result of the literature review, four scales were developed to measure the number sense skills of secondary school students in Turkey (Harç, 2010; Kayhan Altay and Umay, 2010; Birgin and Peker, 2022; Alkaş Ulusoy and Şahiner, 2016). One of these scales was developed by Harç (2010). The scale consists of 5 factors: meaning and size of numbers, equivalent representation of numbers, meaning and effect of operations, mental calculation, and measurement references. The factor structure of the scale was developed based on the number sense components by Reys and Yang (1998) and Yang (1995). The factor structure of the scale was decided based on expert opinion. The validity and reliability of the scale was determined by calculating the KR-20 coefficient and conducting item analysis. The other was developed by Kayhan Altay and Umay (2010). The scale was developed based on Yang's (1995) number sense components. The construct validity of the scale was established using exploratory factor analysis. As a results of the exploratory factor analysis 17-item scale was developed consisting of 3 factors: flexibility, conceptual thinking in fractions, and the use of a reference point. Also the scales reliability was determined by KR-20 coefficient and item analysis was conducted. Alkaş Ulusoy and Şahiner (2016) developed a number sense self-efficacy scale for secondary school students based on the five number sense components created by Yang (1995). However, as a result of the EFA, a four-factor structure emerged: self-efficacy for understanding the meaning and size of numbers, self-efficacy for flexibility in calculation, self-efficacy for flexibility in application, and self-efficacy for mental calculation-estimation. Another scale was developed by Birgin and Peker (2022) for secondary school students. The factor structure of the scale was reached by combining the models developed by Reys and Yang (1995) and Greeno (1991). The content validity and factor structure of the draft scale were decided based on expert opinion. Within the

results of item analysis, non-discriminative questions were removed from the test. Within the scope of construct validity studies with the remaining questions, first EFA and then CFA were conducted. The reliability of the scale croncbach was reached. As a result of the study, the theoretical structure of the scale, which was consisted of 36 items and 6 factors (number knowledge, quantitative reasoning and inference, equivalent representation of numbers, effect of operations, use of reference points in measurement, and mental calculation), was confirmed.

Two scales were developed for the measurement of number sense skills of elementary school students. The first was developed by Can (2012). The scale is based on the five-factor theoretical structure developed by Yang (1995). As a result of the EFA, it was determined that a three-factor structure was formed as utilization of the comparison point, flexibility in calculation and comprehension of number sizes. The fit of the structure formed as a result of EFA was tested by CFA, and it was seen that the structure produced a high fit. The second scale was developed by Çekirdekçi, Şengül, and Doğan (2016). The structure of the scale was based on the 6-factor theoretical structure created by Reys and Yang (1998). However, as a result of the EFA, it was seen that a 3-factor 11-item structure was formed as knowing the equivalents of numbers and quantitative reasoning-drawing, calculating the effects of operations using reference points, knowing the meaning of numbers, and flexible thinking.

The number sense scales developed in Turkey are based on the theoretical models developed by Reys and Yang (1998) and Yang (1995). In the scale developed by Birgin and Peker (2022), the quantitative reasoning and inference component developed by Greeno (1991) was added to the model developed by Reys and Yang (1995). In the number sense scales of Dede and Şengül (2016) and Harç (2010), factor structures were formed based on expert opinion, and the adequacy of the questions in the scale was decided by item analysis. Kayhan Altay and Umay (2010) and Alkaş Ulusoy and Şahiner (2016) designed their scales according to the 5 number sense components created by Yang (1995), but as a result of EFA, it was seen that Kayhan Altay and Umay (2010) reached a 3-factor structure and Alkaş Ulusoy and Şahiner (2016) reached a 4-factor structure different from the theoretical structure. In the study conducted by Birgin and Peker (2022), it was seen that the 6-factor theoretical structure was conformed with the EFA results and confirmed by the CFA.

The study group consisted of fourth grade elementary school students in this research. In Turkey, there are two scales developed by Çekirdekçi et al. (2016) and Can (2012) to measure number sense at the elementary school level. Çekirdekçi et al. (2016) designed their scale based on the 6-factor structure developed by Reys and Yang (1998) and Can (2012) designed his scale based on the 5-factor theoretical approach developed by Yang (1995). However, both of them reached a 3-factor scale different from the theoretical structure as a result of EFA. It was thought that defining the elementary school number sense scale with more factors could contribute more to the determination of students' number skills. In this context, it was decided to test the 6-factor theoretical model created as a result of synthesizing the theoretical structures by Reys and Yang (1998), Yang (1995) and

Greeno (1991). In this context, the 6-factor theoretical model formed by synthesizing the theoretical constructs of Reys and Yang (1998), Yang (1995) and Greeno (1991) was tested. The hypothesis model is given in Figure 1.

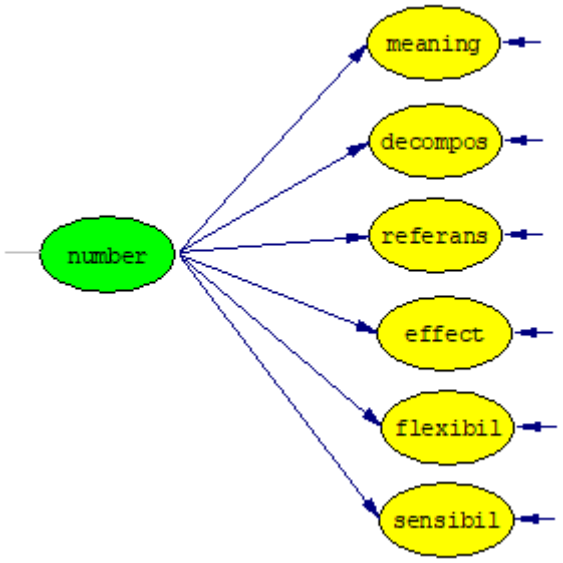

Figure 1. Hypothesis Model of the Number Sense Scale Test

METHOD

Research model

This study, in which the validity and reliability study of the elementary school number sense test was conducted, used the correlational survey model. The correlational survey model is sometimes used to reveal whether there is a change between two or more variables and sometimes to reveal whether the reasons for the change in a variable can be explained by other variables (Karasar, 2002). In this study, the relationship between the items and factors, factors and factors, factors and general structure of the number sense test and problem solving skill, which requires strategic thinking skills just like number sense skill, was examined. Since the study focuses on the correlation between variables, it can be said that the relational survey model was used.

Study group

It is difficult to reach the whole sample related to the problem situation to be solved due to weakened control and economic difficulties. It is therefore more feasible to select smaller samples representing a limited part of the population for the variables to be measured (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz ve Demirel, 2008). If the members of the study group are selected individually, member sampling method is used; if they are selected in groups, cluster sampling method is used (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz ve Demirel, 2008). Determining the study group with the element sampling method will create a divided structure

in the allocation of students to schools. For this reason, the study group was selected by cluster sampling method since it would provide greater convenience in the data collection process.

When determining the study group with the cluster sampling method, the number of people in the population is first determined. Then, the sample to be reached is calculated based on the number of people in the population. In the next stage, information about the clusters in the universe and the number of elements in each cluster is collected. Finally, enough clusters to form the sample are selected from the list by random methods (Özen ve Gül, 2005; Baştürk ve Taştepe, 2013). In order to obtain a reliable factor structure in validity and reliability studies, a minimum sample size of 200 is recommended (Schumacker ve Lomax, 1996; Gerbig ve Hamilton, 1996). A different view on sample size is that 5 times the number of questions should be reached, and 10 times the number of individuals should be reached for more reliable results (Hair, Anderson, Tatham ve Black, 1998; Brown, 2006; Kelloway, 1995). There are 31 questions in the research. Using the criterion of number of questions x 5, 165 people should be reached. Since the number 165 did not meet the minimum sample size criterion of 200 set by Schumacker and Lomax, 1996, Gerbig and Hamilton, 1996, the sample size of the study was reached by using the number of questions x 10 criterion. According to the determined criteria, it was aimed to reach 31x10=310 people. In this context, it was seen that a total of 328 students in 2 schools selected by random methods met the sample size. However, 17 students did not come to school and the solutions of 12 students were not accepted for evaluation due to different reasons. In this context, a total of 299 elementary school fourth grade students studying in 9 classes of 2 elementary schools selected by cluster sampling method in the center of Eskişehir province formed the sample of the study. 55.84% (167) of the students in the sample were female and 44.15% (132) were male.

Criterion validity, which is one of the important conditions for the validity of the scales, was also examined in the study. Cluster sampling method was used for criterion validity. The schools that constitute the sample of the study were selected randomly. In this context, a criterion validity study was conducted with 312 fourth grade students studying in 4 public schools in Eskişehir city center in the 2021-2022 academic year. 146 (46.79%) of the students were male and 166 (53.21%) were female.

Scale development process

The procedures carried out in the validity and reliability process of the number sense scale are given in Table 1.

Table 1. Development Stages and Procedures of the Number Sense Scale

	Test Development Stages	**Procedures in the Test Development Process**
1	Determining the Number Sense Components	• Literature review • Examination of previously developed number sense scales
2	Forming the Item Pool	• Literature, number sense scales, textbook, original questions developed by the researcher
3	Establishing Content Validity	• Submitting the draft items to expert opinion in terms of language, student level and fitness to the components • Making necessary additions and corrections to the items based on the comments and suggestions

		• Conducting a pilot application of the draft articles on 32 students • Making necessary corrections in accordance with student opinions • Calculating the content validity index
4	Implementation of the Number Sense Test	• Application of the draft number sense scale to 311 students • In line with the recommendations of the class teachers, the exams of 12 students were canceled, reaching a final sample size of 299.
5	Item Analysis	• Scoring and ranking the test • Determining the upper and lower groups for item analysis • Calculation item difficulty, item discrimination, item variance, average difficulty, variance values • KR-20 coefficient calculation
6	Construct Validity Analyses	• Testing the model fit of the theoretical structure of the number sense test using CFA
7	Reliability Analysis	• Croncbach α coefficient calculation
8	Criterion Validity Analyses	• Deciding on the criterion validity of the number sense test by examining its correlation with the problem solving achievement test •Finalizing the number sense test as a result of the analysis

When Table 1 is examined, the components that will form the scale were decided at the first stage. Reys and Yang (1998) considered the comprehension of the meaning and size of numbers as a single component. Yang (1995), on the other hand, considered this factor as two different factors: comprehension of the meaning of numbers and the size of numbers. Based on the expert opinions received, it was thought that the first factor should be comprehension of the meaning and size of numbers as in Reys and Yang (1998). The second factor was decided to be the decomposition and recombination of numbers factor determined by Yang (1995), and the fourth factor was decided to be the effects of operations factor determined by Reys and Yang (1998). It was decided that the third and fifth factors of the study would be the use of reference points and flexibility in calculation as identified by both Reys and Yang (1998) and Yang (1995), and the sixth factor would be the sensibility of the numbers developed by Greeno (1991). Following the creation of the main components, an item pool was created to determine the scale questions. The item pool consists of 105 questions included in number sense scales (Reys & Yang, 1998; Yangs 1995; Dede & Şengül, 2016; Harç, 2010; Peker, 2019; Kayhan Altay & Umay, 2010; Alkaş Ulusoy & Şahiner, 2016; Çekirdekçi et al. 2016; Can, 2012), mathematics textbooks and developed by the researcher. A draft scale of 31 items was formed by selecting 6 items for the first factor and 5 items for each factor from the item pool.

The draft scale was first examined by two Turkish language experts in terms of expression disorders and spelling mistakes and reorganized according to the suggestions of the experts. The revised draft scale was reviewed by three experts who had completed their doctoral studies in mathematics education and had previously worked on number sense skills. The first expert stated that 4 items of the scale were included in more than one factor and 3 questions could cause misconceptions, the second expert stated that 5 questions were included in more than one factor and the third expert stated that 2 questions were included in more than one factor and 3 questions could cause misconceptionsous. According to the expert opinions, some questions were removed from the test and new questions were added from the item pool. Some questions were reorganized according to the deficiencies stated by the expert. After the expert opinion was completed, the

draft scale was reviewed by the classroom teachers in terms of level appropriateness. As a result of the teachers' opinions, it was seen that the scale was appropriate for the student level. Although the teachers stated that the scale was appropriate for the level of the students, it was thought that it would be useful to apply it to a pilot group. In this context, the scale was applied to a pilot group of 32 fourth grade elementary school students. As a result of the application, it was seen that there were comprehension problems in 2 items. The items were made comprehensible according to student feedback. In the last stage of the content validity process, the content validity ratio developed by Lawshe (1975) was examined.

Lawshe (1975) technique is to submit the draft scale to expert opinion for the last time before applying it to the main sample. In the Lawshe technique, the content validity rate of the draft scale whose items and factor structure are finalized is determined. In the first stage of the implementation process of the Lawshe technique, the expert group to evaluate the draft scale is determined. A total of 13 people, including 2 mathematics field experts, 8 classroom teachers and 3 mathematics teachers, constituted the evaluation group. The experts were asked to give 3 points to the item if their opinion on the item was related to the target factor, 2 points to the item if their opinion was "related to the target factor but unnecessary", and 1 point to the item if their opinion was "not related to the target factor". In the Lawshe technique, firstly, the number of experts stating that the item is necessary is divided by half of the total number of experts, then 1 is subtracted from the result to reach the content validity rate. In other words, if 11 of 13 experts say that item 1 is necessary, the average content validity of this item will be (11/6.5)-1 =.69. The content validity average of 31 items in the number sense test is between .69 and 1.00. The content validity ratios of the factors are calculated by summing the content validity ratio of each item in the factor and dividing it by the number of items in the factor. In this context, the content validity value was .85 for the meaning and size of numbers, .75 for decomposing and combining numbers, .77 for determining reference points, .89 for the effect of operations on numbers, .96 for the flexible use of numbers, and .75 for the sensibility of the result. The sufficiency of the content validity index obtained is compared with the table showing the minimum criterion values corresponding to the number of experts by Veneziano and Hooper (1997). The minimum values for content validity averages at .05 level of significance are given in Table 2.

Table 2. Minimum Values for Coverage Validity Ratios Corresponding to The Number of Experts

Number of experts	Minimum Value	Number of experts	Minimum Value
5	0.99	13	0.54
6	0.99	14	0.51
7	0.99	15	0.49
8	0.78	20	0.42
9	0.75	25	0.37
10	0.62	30	0.33
11	0.59	35	0.31
12	0.56	40+	0.29

Table 2 shows that the minimum content validity ratio corresponding to 13 experts is .54. The content validity average of all factors in the scale is higher than .54. This finding allowed us to see that the content validity of the number sense scale is sufficient. As a result of the high content validity of the scale, it was decided to apply the number sense test to the main sample group. As a result of the high content validity of the scale, it was decided to apply the number sense test to the main sample group. In the next stage, information was given about the factor structure of the scale applied to the students and the items in the scale.

Meaning and Size of Numbers: Within the meaning of numbers in this factor, students are expected to discover the relationships between different representations of numbers, and within the meaning of the size of numbers, students are expected to sort numbers and compare the distance of numbers to each other (Reys and Yang , 1998). Item 1 was taken from TIMSS (2007) and asked students to determine where the fraction 18/20 comes on the number line. Students with number sense are expected to answer that it comes to point M because it is the closest fraction to 1 or the only fraction greater than half. The 3rd question was taken from TIMSS (2011), in which students with number sense were expected to draw Ayşe's score on a graph based on Ceren and Ahmet's rankings without performing any operations. Item 4 was taken from Phipps (2008), in which students were asked to draw the fraction 21/4 based on the relationship between 1/2 and 1/4. Item 5 was adapted from Kerslake (1986) and asked the students who had acquired number sense to reach a conclusion based on the relationship between the numerator and denominator of fractions greater than 1. Question 6 was adapted from Harç (2010) and asked students to compare the fraction 3/5 with different versions of 1 / 2.

Decomposing and Combining Numbers: In this factor, individuals are expected to comprehend different representations of number in order to make practical calculations (Yang, 1995). Item 8 was taken from Çekirdekçi et al (2016). In the question, students were asked to reach the whole by combining 4 1/4 fractions. Item 9 was developed by the researchers. In this item, students were asked to associate the numbers decomposed over the same amount with the whole in a practical way. The 10th item was developed by the researcher and students who had acquired number sense were expected to realize that the combination of two unknown numbers being 1000 did not affect the new situation. In item 11, students were expected to realize that the fraction corresponds to 1/3 fraction by combining the parts of each unit of the fraction given in the figure.

Determining reference point: In this factor, the ability to use a different object with a known result to measure an object is measured (Reys and Yang 1998; Yang 1995). In this context, the 12th item was taken from Harç (2010) and the students were asked to find the area of the lake with reference to squares. The 13th item was developed by the researcher and the students were asked to find the distance between Kütahya and Erzurum by taking the distance between Kütahya and Antalya as a reference. Item 15 was developed by the researcher and students were asked to estimate the fraction given in the painted area with reference to concepts such as half and quarter. Question 16 was developed by the researcher and asked the students to estimate the total

number of watermelons in the field with reference to the number of watermelons in the width and length of the field.

Effect of operations on numbers: In this factor, the student is expected to recognize the effects of a change in a number or operation on the result (Yang ,1995). Item 17 was taken from Can (2012). In the question, students are expected to state without performing any operation that the result of the 24x9 operation is the same as the result of the 12x18 operation. Item 18 was developed by the researcher, where option a refers to rule-based solutions and option c refers to number sense-based solutions. In the question, students with high number sense were expected to state in their solutions that multiplying by 25 and multiplying by 100 and dividing by 4 express the same situation. Question 19 was taken from Can (2019). In the question, the student was expected to predict the result of the second operation based on the result of the first one. The 20th question was developed by the researcher and it was emphasized that each multiplication operation does not cause a quantitative growth in the number and each division operation does not decrease the numbers.

Flexible use of numbers: In this factor, students are expected to reach the result by using appropriate mental processing strategies (Reys and Yang 1998). Item 22 was developed by Alsawaie (2012). In this question, students were expected to add the numbers in the hundreds digit and add the numbers in the tens digit to reach the number of digits. Question 23 was taken from Can (2012). students were expected to make mental solutions by using the strategy of grouping numbers. Question 24 was developed by the researcher and students were expected to determine the winner by looking at the differences scored by the teams in each period. In question 26, students were expected to make a quick comparison by comparing the number of buttermilk in the supermarket. Here, the students were asked to solve the question "How much would the buttermilk sold in 20s be if it was sold in 10s?" or "How much would the buttermilk sold in 10s be if it was sold in 20s?".

Sensibility of the result: In this factor, students were expected to evaluate the numerical result in real life conditions (Greeno, 1991). Item 27 was developed by the researcher and students were expected to make a realistic estimation about the weight of the baby at the end of the 4th day. Item 28 was developed by the researcher and aimed to determine the age of the grandfather based on the life expectancy of people. 29th item was developed by the researcher and the students were expected to think that 2 workers together can do a job faster than 1 worker and were expected to directly mark the option with less than 20. Item 30 was developed by Verschaffel, Greer, and De Corte (2000), and the students were expected to make a realistic analysis based on the shape of the container rather than reaching a conclusion directly by establishing a ratio.

Data mining process

It was determined that there was no ethical problem in conducting the study with the decision of Dumlupınar University commission meeting 2022/02 on 16.02.2022. Necessary permissions were obtained from Eskişehir

National Education Directorate for the implementation of the study. Before the implementation process, it was explained to the students that answering the question was not enough, they had to explain how they solved the question. During the implementation process, necessary preventive measures were taken to minimize students' interactions with each other. Students were informed that the implementation was not for grading purposes and that it was not necessary for them to write their names. Since 50 minutes was sufficient as the time to answer the questions in the pilot implementation process, this time was decided to be sufficient in the main implementation process.

Data analysis

Item analysis

After the application, the answers given by the students were transferred to the computer environment for item analysis, which is a prerequisite for validity and reliability. During the item analysis, 0 points were given to empty and incorrect answers and 1 point was given to correct answers. Then, the correct answers given to each question were summed and the total score of the number sense test was obtained for each student. The obtained scores were arranged from the students with the highest scores to the students with the lowest scores, 81 (27%) students with the highest scores formed the upper group, 81 (27%) students with the lowest scores formed the lower group. 137 (46%) students in the middle group were not included in the analysis. The item difficulty index (pj), item incorrectness rate (qj), item variance (sj2) and item discrimination index (rjx) were calculated based on the data obtained from 162 students in the upper and lower groups.

Construct validity

Construct is defined as the relational pattern between items considered to be related to each other; construct validity is defined as the correlation between students' answers to the items in the test (Tekin, 2000). There is a widespread belief that factor analysis among statistical methods should be used in construct validity studies (Anastasi, 1988; Dancey & Reidy, 2004; Reio & Wisell, 2006; Urbina, 2004). Factor analysis aims to conceptually transform a large number of interrelated variables into fewer variables (Büyüköztürk, 2006). Factor analysis is divided into exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Çokluk, Şekercioğlu, Büyüköztürk, 2012). EFA is a discrimination technique that generates factors appropriate for the data set using the relationships between items (Byrne, 1994; Green, Salkind, & Akey, 1997; Jöreskog & Sörbom, 1993). CFA tests the model fit of the factor structure defined based on a strong theoretical and empirical framework (Jöreskog & Sörbom, 1993; Raykov & Marcoulides, 2008).

There are different opinions on the use of EFA and CFA in construct validity studies (Hurley et al., 1997; Kline, 2005; Jöreskog & Sörbom, 1993; Staplaton, 1997; Schumacker & Lomax, 1996). The basis of the different opinions lies in the fact that EFA and CFA group items with very different techniques. EFA leaves the loading value of each item free for all factors, which allows the items to load on all factors in the scale (Hovardaoğlu, 2000; Thampson, 2004; Tucker & Maccallum, 1997, Schumacker & Lomax, 1996). CFA, on the other hand,

allows the item to load only on the desired factor on the basis of the theoretical structure, and the loading value on other factors is stabilized to 0 (Brannick, 1995; Kelloway, 1995; Williams, 1995). Since factor loadings are free in scale development studies, EFA determines the factor in which the items in the scale will be included and the number of important factors in the scale independently of the researcher. In CFA, where factor loadings are controlled, the factor in which the items in the scale will be included and the number of factors in the scale are determined based on the theoretical structure created by the researcher. In this context, the disagreements on the use of EFA and CFA are based on the question "what should be the role of the scale developer in the formation of factor structures?" (Hurley et al., 1997; Kline, 2005; Jöreskog & Sörbom, 1993; Staplaton, 1997; Schumacker & Lomax, 1996). In the next stage, different perspectives based on the answers to this question were analyzed.

In scale development studies, if there is very limited information about the subcomponents that make up the theoretical framework, it is recommended to first conduct EFA to explore the structure (Büyüköztürk, 2002; Çokluk et al., 2012; Hair, Anderson, Tatham, & Black, 1998; Brown, 2006; Kelloway, 1995). There are also opinions stating that the factor structure obtained as a result of EFA should be supported by CFA (Bollen & Long, 1993; Maruyama, 1998; Hurley et al. 1997; Schumacker & Lomax, 1996). At this point, Kline (2005) and Erkuş (2003) stated that CFA is a much stricter statistical technique than EFA. They stated that the results of the analysis obtained with EFA mostly failed to pass through the CFA filter, while the scales that passed through the CFA filter increased the positive opinions about their validity. It is also recommended to use CFA in the first stage of the scale development process because it can show the relationships that do not exist in the researcher's mind, the problematic variables in the model and how well the theory in the researcher's mind and reality match (Şimşek, 2007; Gerbig & Hamilton, 1996). A different view argues that the hypothetical structure in the mind of the scale architect is much more meaningful than the structure formed by the numbers. This view recommends the use of CFA in the first stage of the scale development process (Hurley et al., 1997, Erkuş, 2003; Kline, 2005; Schumacker & Lomax, 1996; Gerbig & Hamilton, 1996). Gerbig and Hamilton (1996) stated that CFA is, in reality, partly EFA and partly CFA because the resulting model consists partly of theory and partly of analyses based on model fit. There is a view that there is no absolute truth about the choice of EFA and CFA, and that the decision should be left to the researcher, provided that the reasons are well explained (Çokluk et al., 2012; Hurley et al., 1997, Schumacker & Lomax, 1996). In this study, CFA was used since the number sense test was tested for its fitness to the theoretical structure developed by Reys and Yang (1998), Yang (1995) and Greeno (1991). In the next stage of the research, technical information about CFA is given.

In CFA analysis, each item in the scale is called the observed variable and each factor formed on the basis of the common characteristics of the items is called the latent variable (Bollen & Long, 1993; Maruyama, 1998; Hurley et al. 1997; Schumacker & Lomax, 1996). For this reason, in the next stage of the study, items will be referred to as observed variables and factors as latent variables. When analyzed in terms of CFA types, it is seen that it is

divided into two as first and second order. First-order CFA focuses on the relationship between the observed variables and the latent variables and between the latent variables themselves. Second-order CFA is required to determine the compatibility of the latent variables with the general structure (Çokluk et al., 2012, Şimşek, 2007; Yurdugül & Aşkar, 2008). In this context, both first- and second-order CFA of the number sense scale were conducted.

In order to make the data obtained from the number sense test ready for CFA, blank answers were scored as "0", rule-based incorrect solutions as "1", rule-based correct solutions as "2", strategy-based incorrect solutions as "3", and strategy-based correct solutions as "4". The scoring was carried out by three field experts who completed their doctoral studies in the field of mathematics education. Weighted Kappa coefficient was examined to provide the reliability of the scoring. In this context, the experts were asked to score the answers of 50 students to the number sense test. There was no interaction between the experts during the scoring process. Then, the scores given by the experts were transferred to the computer environment and the kappa coefficient was calculated. The data obtained from kappa coefficient are interpreted as "Poor agreement=< 0.20; Acceptable agreement=0.20-0.40; Moderate agreement=0.40-.60; Good agreement=0.60-0.80; Very good agreement=0.80-1.00" (Şencan, 2005, p. 485). Accordingly, the inter-rater agreement was found to be .83. This result showed high inter-rater agreement. Due to the high agreement, the remaining 244 scales were scored by a single expert.

Reliability analysis

The number sense test was developed for both 0-1 scoring and 0-1-2-3-4 Likert-type scoring. According to Büyüköztürk (2002), the reliability of tests scored as 0-1 is calculated by KR-20, while the reliability of Likert-type scales is calculated by looking at Croncbach α coefficient. Therefore, it was necessary to determine whether the reliability varies according to the scoring type. Therefore, reliability according to both scoring types was calculated for the whole scale and its latent variables.

Criterion validity

Criterion validity is the determination of the relationship between a developed test and another valid scale (Karaca, 2006; Tekin, 1997; Yılmaz, 1998). Number sense is defined as individuals' flexible use of numbers and strategies (Greeno, 1991; Reys et al., 1999; Kalchman, et al., 2001; Yang, 2019). Another skill that requires strategy skills in mathematics education is problem solving (Verschaffel & DeCorte, 1993; Yıldızlar, 2001; Gök & Sılay, 2008; Altun, 1995; Yazgan & Bintaş, 2005). In this context, the relationship between the factor structure obtained as a result of the construct validity studies and the problem solving scale developed by Ulu (2017) was examined. The problem solving scale developed by Ulu (2017) is one-dimensional and consists of 10 questions. The questions in the scale allow the use of different strategies such as writing math sentences, working backwards, prediction and control, pattern searching, elimination and systematization. It was determined that

the scale explained 66.32% of the variance in problem solving variance and the reliability of the scale was .84. In this context, the number sense test and the problem solving test was applied to a total of 312 fourth grade students. The data obtained were transferred to the computer environment and the relationship between the number sense scale and the problem solving scale was analyzed using Pearson correlation coefficients.

FINDINGS

Findings from item analysis studies

Within the scope of item analysis studies, 81 (27%) students with the highest score on the number sense test constituted the upper group, and 81 (27%) students with the lowest score constituted the lower group. The data obtained from 162 students were analyzed by computing item difficulty index (pj), error rate (qj), item variance (sj2) and item discrimination index (rjx); the findings are given in Table 3.

Table 3. Findings Related to the Item Analysis of the Elementary School Number Sense Scale

Sorular	pj	qj	Sj^2	rjx
s1	,75	,25	,19	,33
s2	,69	,31	,21	,27
s3	,39	,61	,24	,53
s4	,46	,54	,25	,58
s5	,73	,27	,20	,40
s6	,41	,59	,24	,47
s7	,90	,10	,09	,20
s8	,31	,69	,21	,40
s9	,55	,45	,25	,56
s10	,63	,37	,23	,44
s11	,35	,65	,23	,43
s12	,51	,49	,25	,35
s13	,75	,25	,19	,38
s14	,12	,88	,10	,14
s15	,33	,67	,22	,38
s16	,39	,61	,24	,58
s17	,60	,40	,24	,60
s18	,77	,23	,18	,38
s19	,57	,43	,24	,68
s20	,41	,59	,24	,72
s21	,83	,17	,14	,27
s22	,55	,45	,25	,70
s23	,64	,36	,23	,51
s24	,57	,43	,25	,59
s25	,72	,28	,20	,28
s26	,54	,46	,25	,70
s27	,35	,65	,23	,52
s28	,26	,74	,19	,32
s29	,55	,45	,25	,60
s30	,23	,77	,18	,31
s31	,09	,91	,08	-,04

In item analysis studies, the items to be used in the test or to be removed from the test are determined. in this context, item discrimination indices (rjx) were first analyzed from the data in Table 3. The item discrimination index (rjx) is a value ranging between -1.00 and +1.00. According to the item discrimination index, items below 0.30 are removed from the test, while items above 0.30 remain in the test (Karaca, 2006; Tekin, 1997; Yılmaz, 1998). When Table 3 is examined, it is seen that there are 6 questions (p2, p7, p14, p21, p25, p31) with item discrimination index (rjx) below 0.30. It was decided to remove these questions from the scale. In the next stage, item analysis was continued by examining the item difficulty indices (pj) of the remaining 25 questions in the test. Questions with item difficulty indices (pj) between 0-0.29 are considered as difficult, 0.30-0.69 as medium, 0.70-1.00 as easy (Karaca, 2006; Tekin, 1997; Yılmaz, 1998). When the values in Table 3 are analyzed, it is seen that 3 (p1, p5, p18), 20 (p3, p4, p6, p8, p9, p10, p11, p12, p15, p16, p17, p19, p20, p22, p23, p24, p26, p27, p28, p29) and 2 (p28, p30,) of the remaining 25 questions in the number sense scale are easy, medium and difficult respectively. In achievement tests, questions that serve the purpose are generally expected to be at the medium difficulty level (0.30-0.69). In this context, it can be considered that a total of 5 questions, 3 of which were easy and 2 of which were difficult, did not serve the purpose, but it was decided to keep these questions in the test because their discrimination values were sufficient and students using different number sense strategies could be revealed. The arithmetic mean of the 25-question test is 12.81. The average difficulty index value of the test obtained by dividing the arithmetic mean by the number of questions is 0.51. This value allows us to see that the test as a whole is at the medium difficulty level.

After examining the difficulty and discrimination of the items in the test, the reliability of the test was also examined. Yılmaz (1998) emphasized that if a test is item analyzed and the items are scored as 0-1, the reliability of that test should be determined by KR-20 coefficient. The KR-20 coefficient is an internal consistency coefficient that allows us to determine the consistency of the items with each other and with the overall test. The KR-20 formula is given in Figure 2

$$r_{kr\cdot 20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum pq}{\sigma^2}\right)$$

Figure 2: The KR-20 formula

KR-20 Formula;

k = number of questions

p = item difficulty index

q = percentage of item incorrect answers

σ2 = variance of the test

In Table 3, item variance (sj2) values were found by multiplying the difficulty index (pj) for each item by the percentage of incorrect answers (qj). The sum of the item variances of 25 questions is 5.67 and the total

variance of the test is 37.84. Based on these values, the KR-20 reliability coefficient calculated for the whole scale was 0.89. The KR-20 coefficient was calculated for the sub-factors. It was found 0.77 for the factor of meaning and size of numbers, 0.74 for the factor of decomposing and combining numbers, 0.81 for the factor of determining reference points, 0.76 for the factor of effect of operations on numbers, 0.84 for the factor of flexible use of numbers, and 0.72 for the factor of sensibility of operations. According to Büyüköztürk (2002), a KR-20 value of 0.70 and above indicates that the internal consistency and therefore the reliability of the test is high. The fact that the KR-20 coefficient was above 0.70 for the whole number sense scale and its sub-factors shows that our test is reliable. It was decided to examine the construct validity of the test after it was seen that the number sense scale with 6 factors and 25 items was reliable.

Construct validity analysis

As a result of the item analysis, a first level CFA was conducted to determine whether the theoretical structure consisting of 6 latent variables and 25 indicator variables was matched with the empirical structure. The model obtained is given in Figure 1.

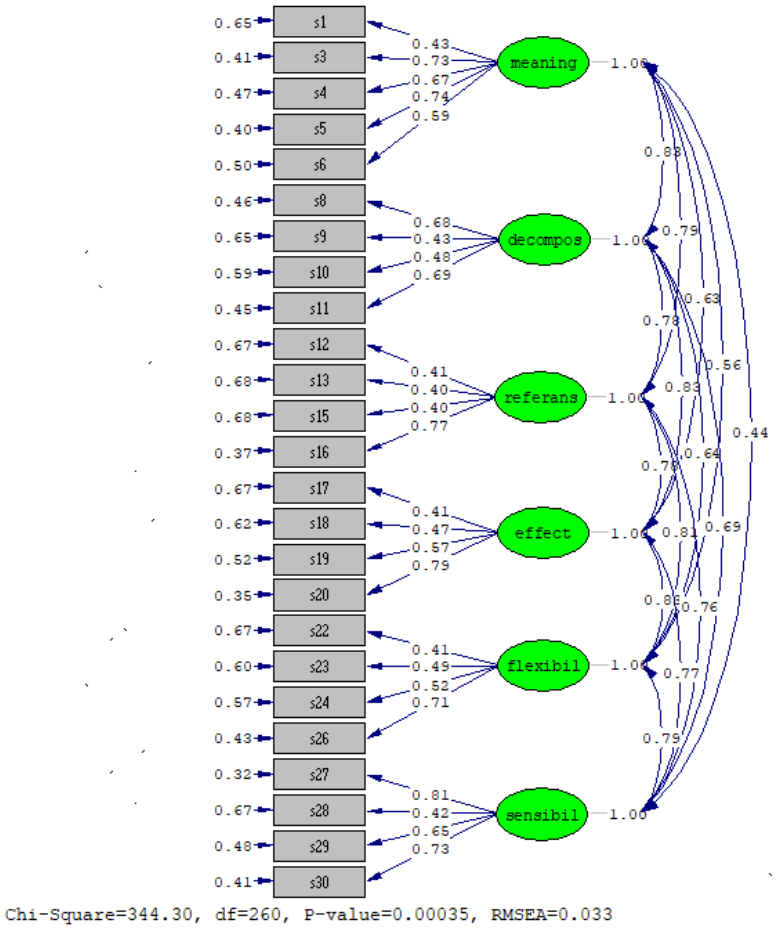

Figure 2. Unadjusted First Level Confirmatory Factor Analysis Results of the Number Sense Scale

Figure 1 shows that the factor loadings of the observed variables on the latent variable. 41 to. 81 in the latent variable. If the factor loading value produced by the observed variable in the latent variable is below .30, it indicates that the observed variable has a low fit with the latent variable. In this case, observed variables with low fit with the latent variable can be assigned to other latent variables with high fit. If they cannot be assigned, they are removed from the scale (Çokluk vd., 2012; Thompson, 2008; Kline, 2005; Schumacker ve Lomax, 1996). The factor loadings of the observed variables in the number sense scale are above .30. This shows that the indicator variables have high fit with the latent variable to which they are related. Another factor that causes the observed variables to be retained or removed from the scale is the variance of the observed variable that cannot be explained by the latent variable (error variance). If this value is too high and the t values that test the significance of the path between the observed variable and the latent variable are not significant, the observed variable is excluded from the test. Error variances of the observed variables in the scale. 32 to. 68 and the t-values expressing the paths between all indicator variables and latent variables were found to be significant ($p < .05$). These results show that the amount of error produced by the observed variables in the scale in the latent variables is at an acceptable level (Jöreskog & Sörbom, 1993; Raykov & Marcoulides, 2008; Şimşek, 2007; Thompson, 2008). As a result of the significant relationship between the observed variables and latent variables, the model data fit of the scale was also examined. In this context, the criterion values of the fit indices (Byrne, 2010; Schermelleh Engel, Moosbrugger, & Müller, 2003; Şimşek, 2007) and the model fit indices calculated as a result of the first level CFA are given in Table 4.

Table 4. Fit Index Criteria Values and Fit Index Values for the First Level CFA Results of the Number Sense Scale

Fit indices	**Excellent fit**	**Good fit**	**Values obtained from Model 1**
χ^2/sd	$0 \le \chi^2/sd \le 2$	$0 \le \chi^2/sd \le 5$	1.32
RMSEA	$0 \le RMSEA \le .05$	$0 \le RMSEA \le .08$	.33
NFI	$.95 \le NFI \le 1.00$	$.90 \le NFI \le .94$	.86
NNFI	$.95 \le NNFI \le 1.00$	$90 \le NNFI \le .94$	.96
CFI	$.95 \le CFI \le 1.00$	$.90 \le CFI \le .94$	.96
IFI	$.95 \le IFI \le 1.00$	$.90 \le IFI \le .94$	.96
GFI	$.95 \le GFI \le 1.00$	$.90 \le GFI \le .94$	.92
AGFI	$.95 \le AGFI \le 1.00$	$.90 \le AGFI \le .94$	.91

When Table 4 is examined, the first level CFA results of the number sense scale show that it produced excellent fit on four indices ($\chi^2/sd=1.32$, RMSEA =.33, NNFI =.96; CFI=.96; IFI=.96), good fit on two indices (GFI= .92; AGFI= .91) and poor model-data fit on one index (NFI=.86). In the model, it was investigated why the NFI value produced low data fit and it was determined that this value produced low fit in small samples. NFI is a value that produces a fit index based on the difference between the independent model in which there is no relationship between the variables and the model created by the researcher. It is stated that using the NNFI value, which makes calculations by taking into account the degrees of freedom in small samples, is more appropriate than the NFI value (Çokluk et al., 2012; Sümer, 2000; Tabachnick & Fidell, 2001). In this context,

the excellent NNFI value of the scale tolerates the low fit obtained for the NFI value. As a result of the adequacy of the indices showing model-data fit, standardized regression coefficients showing the regression coefficient between the latent variables in the scale were examined. Standardized regression coefficients are given in Table 5.

Table 5. Standardized Regression Coefficients between Latent Variables

	Meaning	Decomposition	Reference	Effect	Flexilibility	Sensiblelity
Meaning	1.00					
Decomposition	.83	1.00				
Reference	.79	.78	1.00			
Effect	.63	.83	.78	1.00		
Flexilibility	.56	.64	.81	.83	1.00	
Sensiblelity	.44	.69	.76	.77	.79	1.00

When Table 5 is analyzed, it is seen that the correlation between the latent variables varies between .44 and. 83. According to Çokluk et al. (2012), a standardized regression coefficient of .85 and above between two latent variables indicates that the variables measure similar constructs. The similarity of the latent variables makes it difficult to differentiate the dimensions, which reduces validity. When such situations are encountered in the CFA process, it is recommended to remove one of the similar latent variables. In the number sense scale, there is no relationship of .85 and above between the latent variables. This finding shows that the scale does not have a multi-connection problem. First-order CFA focuses on the relationship between the observed variables and the latent variables and between the latent variables themselves. Second-order CFA is required to determine the fit of the latent variables with the overall structure (Çokluk et al., 2012, Şimşek, 2007; Yurdugül & Aşkar, 2008). In this context, a second level CFA was conducted to determine the relationship between the latent variables of meaning and size of numbers, decomposing and combining numbers, setting reference points, effect of operations on numbers, flexible use of numbers, and sensibility of operations with the general structure of number sense scale. The model obtained is given in Figure 2.

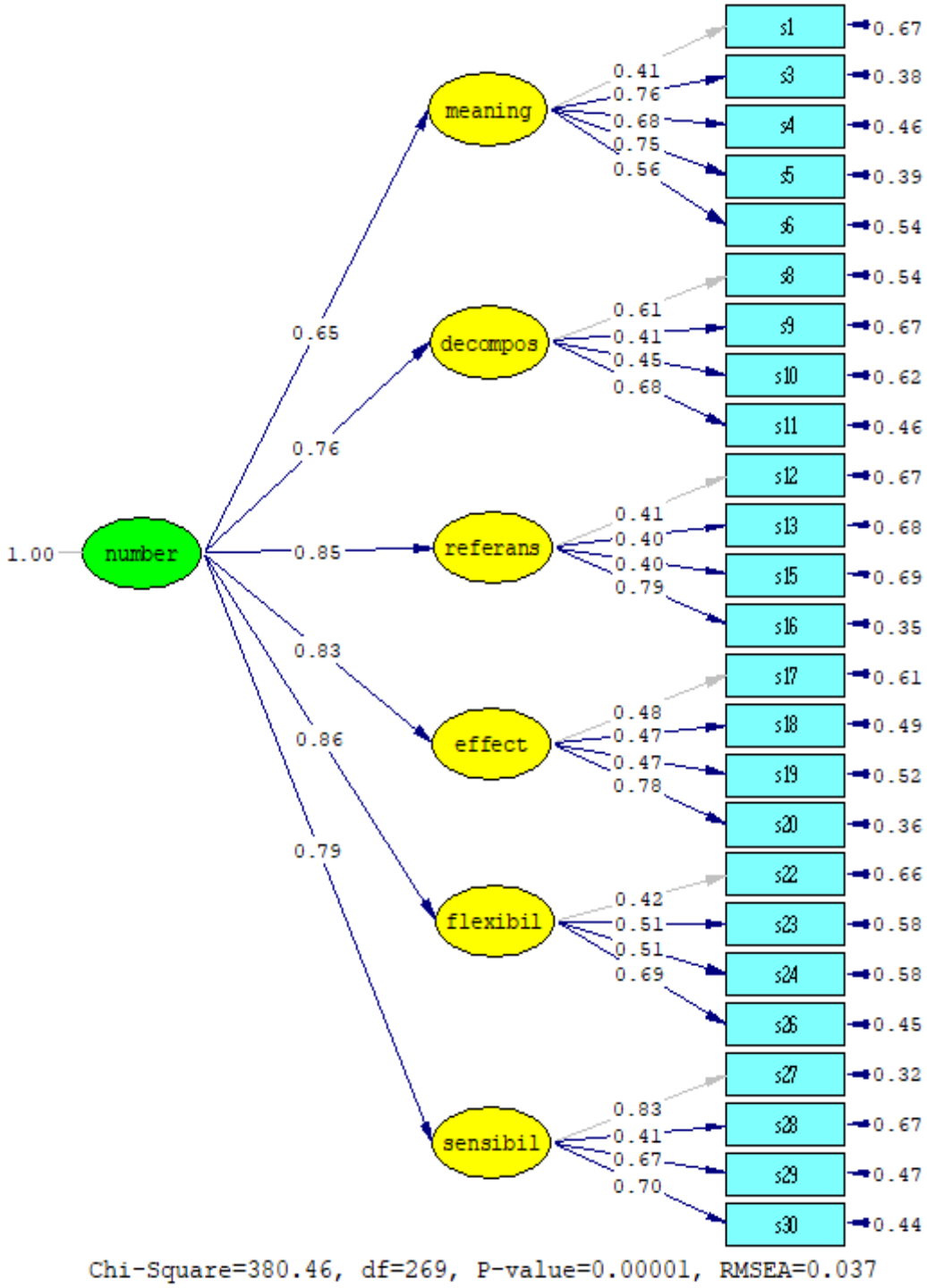

Figure 2. Adjusted Second Level Confirmatory Factor Analysis Results of the Number Sense Scale

When Figure 3 is examined, it is seen that the paths from items to latent variables and from latent variables to the general structure are significant as a result of the second level CFA conducted to determine the relationship between the latent variables and the general structure. In the next step, the model fit indices of the second level CFA were examined. The findings are given in Table 6.

Table 6. Fit Index Criteria Values and Fit Index Values for the Second Level CFA Result of the Number Sense Scale

Fit indices	Excellent fit	Good fit	Values obtained from Model 1
χ^2/sd	$0 \le \chi^2/sd \le 2$	$0 \le \chi^2/sd \le 5$	1.41
RMSEA	0 ≤ RMSEA ≤ .05	0 ≤ RMSEA ≤ .08	.37
NFI	.95≤ NFI≤ 1.00	.90≤ NFI≤ .94	.86
NNFI	.95≤ NNFI≤ 1.00	90≤ NNFI≤ .94	.95
CFI	.95 ≤ CFI≤ 1.00	.90 ≤ CFI≤ .94	.95
IFI	.95 ≤ IFI ≤ 1.00	.90 ≤ IFI ≤ .94	.95
GFI	.95 ≤ GFI ≤ 1.00	.90 ≤ GFI ≤ .94	.91
AGFI	.95 ≤ AGFI ≤ 1.00	.90 ≤ AGFI ≤ .94	.90

When Table 6 is examined, the second level CFA results of the number sense scale show that it produced excellent fit in five indices (χ^2/sd=1.41, RMSEA =.33, NNFI =.95; CFI=.95; IFI=.95), good fit in two indices (GFI= .91; AGFI= .90) and poor fit in one index (NFI=.86). As stated before, NNFI can be an alternative to NFI in small sample groups (Çokluk et al., 2012; Sümer, 2000; Tabachnick & Fidell, 2001), in this context, the fact that the NNFI value produced an excellent fit showed that there was a difference between the independent model and the defined model. As the second level CFA of the scale showed that the model-data fit was adequate, the standardized regression coefficients showing the relationship between the latent variables in the scale and the general structure of the scale were examined. Standardized regression coefficients are given in Table 7.

Table 7. Standardized Regression Coefficients between Latent Variables and General Structure

	Meaning	Decomposition	Reference	Effect	Flexibility	Sensible	General
Meaning	1.00						
Decomposition	.53	1.00					
Reference	.64	.79	1.00				
Effect	.64	.80	.76	1.00			
Flexibility	.64	.79	.74	.76	1.00		
Sensiblity	.55	.69	.72	.83	.72	1.00	
General	.65	.76	.85	.83	.86	.79	1.00

When Table 7 is examined, it is seen that the latent variable that explains the number sense skill at the highest level is the flexible use of numbers (ξ=.86), followed by determining the reference point (ξ=85), effect of operations on numbers (ξ=83), sensibility of numbers (ξ=.79), decomposition and combination (ξ=.76) and the meaning and size of numbers (ξ=.65). Findings regarding the final model of the number sense scale is given in Table 8.

Table 8. Findings Related to the Final Model of the Number Sense Scale

Latent variable	Item	Faktor loadings λ	Error variance ψ	Determination coefficient R^2	Regression coefficient ξ	Factor variance $\sigma 2$	Croncbach α
Meaning and Size of Numbers	S1	.41	.67	.33	.65	.43	.79
	S3	.76	.38	.62			
	S4	.46	.68	.32			
	S5	.75	.39	.61			
	S6	.56	.54	.46			
Decomposing and	S8	.61	.54	.46	.76	.58	.75
	S9	.41	.67	.33			

Combination	S10	.45	.62	.38			
	S11	.68	.46	.54			
Determining the Reference Point	S12	.41	.67	.33	.85	.72	.86
	S13	.40	.68	.32			
	S15	.40	.68	.32			
	S16	.79	.35	.65			
Effect of Operations on Numbers	S17	.48	.61	.39	.83	.69	.74
	S18	.47	.49	.51			
	S19	.57	.52	.48			
	S20	.78	.36	.64			
Flexible Use of Numbers	S22	.42	.66	.34	.86	.74	.81
	S23	.51	.58	.42			
	S24	.51	.58	.42			
	S26	.69	.45	.55			
Sensibility of Numbers	S27	.83	.32	.68	.79	.62	.73
	S28	.41	.67	.33			
	S29	.67	.47	.53			
	S30	.70	.44	.56			

When Table 8 was examined, it was seen that the factor loadings of the observed variables in the final model of the number sense scale ranged between .40 and .83, the coefficient of determination explaining the variance of the observed variables in the latent variable ranged between .32 and .68, and all of the paths from the observed variables to the latent variables were significant ($p<.05$). It is seen that the latent variable that explains the most variance in the number sense scale is the flexible use of numbers ($\sigma 2=.74$) followed by the variable of setting a reference point ($\sigma 2=72$), effect of operations on numbers ($\sigma 2=.69$), sensibility of numbers ($\sigma 2=.62$), decomposition and combination ($\sigma 2=.69$), and meaning and size of numbers ($\sigma=.43$).

The reliability of the scale was previously determined with the KR-20 coefficient for the case where the test was used by scoring 0-1, but the number sense test can also be used with the Likert-type scoring method. In order to determine whether the test is reliable for Likert-type, Croncbach α coefficient was calculated both for the whole scale and for each latent variable. Croncbach α found .88 for the whole number sense scale, .79 for the meaning and size of numbers, 0.75 for the factor of decomposing and combination, .86 for the factor of determining reference points, 0.74 for the the effect of operations on numbers, 0.81 for flexible use of numbers, and 0.73 for the sensibility of numbers. According to Büyüköztürk (2002), Croncbach α coefficient .70 and above indicates that the internal consistency and therefore the reliability of the test is high. The fact that Croncbach α değerlerinin was above .70 both for the whole scale and for each factor allowed us to see that the number sense scale also produced reliable results when it was scored in Likert form.

Criterion validity of the Number Sense Scale

In order to determine the criterion validity of the number sense scale, its relationship with the problem solving scale was determined. In this context, Pearson correlation coefficients between the sub-dimensions of the number sense scale and the problem solving test were examined and the findings are given in Table 9.

Table 9. Pearson Correlation Coefficients between the Sub-Dimensions of the Number Sense Scale and the Problem Solving Scale

	Meaning	Decomposition	Reference	Effect	Flexibility	Sensible	Problem
Meaning	1	.424**	.385**	.402**	.466**	.413**	.559**
Decomposition	.424**	1	.380**	.427**	.418**	.437**	.504**
Reference	.385**	.380**	1	.403**	.469**	.416**	.474**
Effect	.402**	.427**	.403**	1	.537**	.518**	.497**
Flexibility	.466**	.418**	.469**	.537**	1	.638**	.619**
Sensible	.413**	.437**	.416**	.518**	.638**	1	.559**
Problem	.559**	.504**	.474**	.497**	.619**	.559**	1

When Table 9 is examined, it is seen that there is a significant (p<.01) relationship between the problem solving scale requiring strategic thinking and the meaning and size of numbers (r=.559), decomposition and combination (r=.504), setting a reference point (r=.474), the effect of operations on numbers (r=.497), flexible use of numbers (r=.619) and sensibility of the result (r=.559). The relationship between the factors of the scale developed for criterion validity and the criterion scores should not be lower than .30 (Karaca, 2006; Tekin, 1997; Yılmaz, 1998). It is seen that there is no relationship lower than .30 between the factors of the number sense scale and the problem solving achievement test. In this context, it can be said that the number sense scale, which requires strategic thinking skills, has criterion validity with the problem solving skill that measures strategic skills.

CONCLUSION AND DISCUSSION

In the study, the number sense model consisting of 6 factors and 31 items, which was created by synthesizing the theoretical structures of Reys and Yang (1998), Yang (1995) and Greeno (1991), was tested. The appropriateness of the questions in the scale to the factor structure was examined by mathematics field experts. The content validity coefficients of the factors in the scale were determined using Lawshe (1975) technique and it was seen that the content validity of the factors was high. As a result of the item analysis, 6 items were found to be non-discriminative and it was decided to remove them from the test. CFA was conducted to test the remaining 25-item structure and it was seen that the model fit of the theoretical structure was high. The calculated reliability coefficients showed that the scale was reliable both as a whole and on a factor basis. Finally, the criterion validity of the scale was determined and found to be sufficient by examining the relationship between problem solving skills, which measure strategic skills just like the number sense scale. In the light of all these findings, it was concluded that the elementary school number sense scale with its 26-item, 6-factor structure was valid in terms of content, structure and criterion, and reliable in terms of internal consistency.

The scale developed by Çekirdekçi et al. (2016), which measures elementary school number sense skills, was designed in accordance with the 6-factor theoretical structure developed by Reys and Yang (1998), and the scale developed by Can (2012) was designed in accordance with the 5-factor theoretical structure developed by Yang (1995). However, both of them reached a 3-factor structure different from the theoretical structure as a result of EFA. The elementary school number sense scale consists of a structure with 6 factors. The reason why the factor structure of the scale conflicts with the structures obtained by Çekirdekçi et al. (2016) and Can

(2012) is related to the different factor analysis techniques used. Reaching the factor structure directly with the CFA technique in the study may create doubts about the construct validity of the scale. In scale development studies, it is stated that if there is very limited information about the theoretical basis and the subcomponents that make up the theoretical basis, EFA should be conducted first to discover the structure (Büyüköztürk, 2002; Çokluk et al., 2012; Hair, Anderson, Tatham, & Black, 1998; Brown, 2006; Kelloway, 1995). However, there are many studies on number sense and its subcomponents (Reys & Yang, 1998; Yangs 1995; Dede & Şengül, 2016; Harç, 2010; Peker, 2019; Kayhan Altay & Umay, 2010; Alkaş Ulusoy & Şahiner, 2016; Çekirdekçi et al. 2016; Can, 2012). In areas where there is sufficient research, it is recommended to use CFA to see how well the theory and reality match (Şimşek, 2007; Gerbig & Hamilton, 1996). A different view, which argues that the hypothetical structure in the mind of the scale architect is much more meaningful than the structure formed by the numbers, suggests using CFA in the first stage of the process (Hurley et al., 1997, Erkuş, 2003; Kline, 2005; Schumacker & Lomax, 1996; Gerbig & Hamilton, 1996). In this context, the theoretical structure of the scale was tested using CFA. The high criterion validity of the scale eliminates the doubts about validity. Criterion validity can continue to be tested by using the scale with different variables.

ETHICAL TEXT

In this study, all the rules set by the journal regarding ethical situations were complied with. Any ethical violation that may arise belongs to the authors of the article. It was determined that there was no ethical problem in conducting the study with the decision of Dumlupınar University commission meeting 2022/02 on 16.02.2022.

Author(s) Contribution Rate: In this study, the contribution rate of the first author is 50% and the contribution rate of the second author is 50%.

REFERENCES

Alsawaie, O. N. (2012). Number sense-based strategies used by high-achieving sixth grade students who experienced reform textbooks. *International Journal of Science and Mathematics Education, 10*, 1071-1097.

Altay, M. K., & Umay, A. (2013). İlköğretim ikinci kademe öğrencilerine yönelik sayı duyusu ölçeği'nin geliştirilmesi. *Eğitim ve Bilim, 38*(167),241-255.

Altun, M. (1995). İlköğretim matematik programının değerlendirilmesi. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi,10*(1), 143-154

Altun, M., Bintaş, J., Yazgan, Y., & Arslan, Ç. (2004). *İlköğretim Çağındaki Çocuklarda Problem Çözme Gelişiminin İncelenmesi.* Bursa: Uludağ Üniversitesi, Bilimsel Araştırma Projeleri Birimi.

Baki, A., & Kartal, T. (2004). Kavramsal ve işlemsel bilgi bağlamında lise öğrencilerinin cebir bilgilerinin karakterizasyonu. *Türk Eğitim Bilimleri Dergisi, 2*(1), 27-46.

Baştürk, S., & Taştepe, M. (2013). *Evren ve örneklem. Bilimsel Araştırma Yöntemleri*, Ankara: Vize Yayıncılık, 129, 159.

Birgin, O., & Peker, E. S. (2022). Development of Number Sense Test for Eighth-Grade Students: A Validity and Reliability Study. *Cukurova University Faculty of Education Journal, 51*(1), 187-219.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models* (C. 154). Sage 1993.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16*(3), 201-213.

Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö., Karadeniz, Ş., & Demirel, F. (2008). *Bilimsel araştırma yöntemleri.* Ankara: Pegem Akademi.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*(3), 289-311.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Third Edition (3rd ed.). New York:Routledge

Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental psychology, 27*(5), 777-786.

Can, D. (2017). İlkokul dördüncü sınıf öğrencilerinin sayı duyularının bağlam temelli ve bağlam temelli olmayan problem durumlarında incelenmesi. *İlköğretim Online. 18*(4), 1751-1765.

Çekirdekçi, S., Şengül, S., & Doğan, M. C. (2016). 4. Sınıf Öğrencilerinin Sayı Hissi İle Matematik Başarıları Arasındaki İlişkinin. *Qualitative Studies, 11*(4), 48-66.

Çekirdekçi, S., Şengül, S., & Doğan, M. C. (2017). 4. Sınıf sayı hissi testi'nin geliştirilmesi. *Kalem Eğitim ve İnsan Bilimleri Dergisi, 7*(2), 441-473.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları* (C. 2). Pegem Akademi.

De Corte, E., Verschaffel, L., & Greer, B. (2000). *Connecting mathematics problem solving to the real world.* Proceedings of the International Conference on Mathematics Education into the 21st Century: Mathematics for Living.

Dede, H. G., & Şengül, S. (2016). İlköğretim ve ortaöğretim matematik öğretmen adaylarının sayı hissinin incelenmesi *1. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 7*(2), 285-303.

Erdem, Z. Ç., & Duran, H. (2015). Yetişkinlerin zihinden hesaplama becerilerinin özellikleri üzerine karşılaştırmalı bir çalışma. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 6*(3), 463-482.

Erkuş, A. (2003). Psikometri Üzerine Yazılar: Ölçme ve psikometrinin tarihsel kökenleri, güvenirlik, geçerlik, madde analizi, tutumlar. *Türk Psikologlar Derneği.*

Gök, T., & Silay, I. (2008). Effecs of problem-solving strategies teaching on the problem-solving attıtudes of cooperative learning groups in physıcs education. *Journal of Theory & Practice in Education (JTPE), 4*(2), 253-266.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research In Mathematics Education, 22*(3), 170-218.

Griffin, S. A., Case, R., & Siegler, R. S. (1994). *Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure*. The MIT Press.

Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education, 3*(1), 1-49.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multivariate data analysis. Englewood cliff. New jersey, USA, 5(3), 207-2019.

Harç, S. (2010). *6. sınıf öğrencilerinin sayı duygusu kavramı açısından mevcut durumlarının analizi* (Yayınlanmış Yüksek Lisans Tezi) Marmara Üniversitesi.

Hovardaoğlu, S. (2000). *Davranış bilimleri için araştırma teknikleri*. VE-GA Yayınları.

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and individual differences, 20*(2), 82-88.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Lawrence Erlbaum Associates, Inc.

Kabaca, T., Çontay, E. G., & İymen, E. (2011). Dinamik matematik yazılımı ile geometrik temsilden cebirsel temsile: Parabol kavramı. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 30*(30), 101-110.

Kalchman, M., Moss, J., & Case, R. (2001). *Psychological models for the development of mathematical understanding: Rational numbers and functions. Cognition and instruction*: Twenty-five years of progress, 1-38.

Karaca, E. (2006). Öğretimde planlama ve değerlendirme dersine yönelik bir tutum ölçeği geliştirme. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 25*, 119-128.

Kelloway, E. K. (1995). Structural equation modelling in perspective. *Journal of Organizational behavior, 16*(3), 215-224.

Kerslake, D. (1986). *Fractions: Children's Strategies and Errors. A Report of the Strategies and Errors in Secondary Mathematics Project.* NFER-NELSON Publishing Company, Ltd.

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage publications.

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

Mohamed, M., & Johnny, J. (2010). Investigating number sense among students. *Procedia-Social and Behavioral Sciences, 8,* 317-324.

National Council Of Teachers Of Mathematics (2000). *Principles And Standards For School Mathematics.* Reston, VA: NCTM.

Olkun, S., Fidan, E., & Özer, A. B. (2013). 5-7 yaş aralığındaki çocuklarda sayı kavramının gelişimi ve saymanın problem çözmede kullanımı. *Eğitim ve Bilim, 38*(169).

Olkun, S., Mutlu, Y., & Sarı, M. H. (2017). The relationships between number sense and mathematics achievement. *International Conference on Education and New Developments,* 24-26.

Phipps, M. C. (2008). *A phenomenological investigation on eighth graders' number sense of fractions* (Doctoral dissertation), University of Northern Colorado.

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis.* Routledge.

Reio Jr, T. G., Petrosko, J. M., Wiswell, A. K., & Thongsukmag, J. (2006). The measurement and conceptualization of curiosity. *The Journal of Genetic Psychology, 167*(2), 117-135.

Reys, R. E., Reys, B. J., Nohda, N., & Emori, H. (1995). Mental computation performance and strategy use of Japanese students in grades 2, 4, 6, and 8. *Journal for research in mathematics education, 26*(4), 304-326.

Reys, R. E., & Yang, D.-C. (1998). Relationship between computational performance and number sense among sixth-and eighth-grade students in Taiwan. *Journal for Research in Mathematics Education, 29*(2), 225-237.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online, 8*(2), 23-74.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Lawrence Erlbaum Associates, Mahwah.

Sümer, N. (2000). Yapisal Eşitlik Modelleri: Temel Kavramlar ve Örnek Uygulamalar. *Türk Psikoloji Yazilari, 3*,49-73.

Tekin, H. (1997). *Eğitimde ölçme ve değerlendirme*. Yargı Yayınları.

Tucker, L. R., & MacCallum, R. C. (1997). *Exploratory factor analysis.* Unpublished manuscript, Ohio State University, Columbus.

Ulu, M. (2017). The Effect of Reading Comprehension and Problem Solving Strategies on Classifying Elementary 4th Grade Students with High and Low Problem Solving Success. *Journal of Education and Training Studies, 5*(6), 44-63.

Ulusoy, Ç. A., & Şahiner, Y. (2017). Sayı Duyusuna Yönelik Özyeterlilik Ölçeğinin Geliştirlmesi. *Kastamonu Eğitim Dergisi, 25*(1), 17-32.

Veneziano, L. (1997). A method for quantifying content validity of health-related questionnaires. *American Journal of Health Behavior, 21*(1), 67-70.

Verschaffel, L., & De Corte, E. (1993). A decade of research on word problem solving in Leuven: Theoretical, methodological, and practical outcomes. *Educational psychology review, 5,* 239-256.

Whitacre, I., & Nickerson, S. D. (2016). Prospective elementary teachers making sense of multidigit multiplication: Leveraging resources. *Journal for Research in Mathematics Education, 47*(3), 270-307.

Yang, D.-C. (1995). *Number sense performance and strategies possessed by sixth-and eighth-grade students in Taiwan.* University of Missouri-Columbia.

Yang, D.-C. (2019). Development of a three-tier number sense test for fifth-grade students. *Educational Studies in Mathematics, 101*(3), 405-424.

Yazgan, Y., & Bintaş, J. (2005). İlköğretim dördüncü ve beşinci sınıf öğrencilerinin problem çözme stratejilerini kullanabilme düzeyleri: Bir öğretim deneyi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 28*(28), 210-218.

Yurdugül, H., & Aşkar, P. (2008). An investigation of the factorial structures of pupils' attitude towards