https://doi.org/10.53656/math2023-6-3-the

Educational Technologies
Образователни технологии

THE CONSTRUCTION OF VALID AND RELIABLE TEST FOR THE DIVISIBILITY AREA

Dr. Daniela Zubović, Dr. Dina Kamber Hamzić
Faculty of Science, University of Sarajevo (Bosnia and Hercegovina)

Abstract. Understanding divisibility at the primary school level is a strong predictor of students' mathematical achievements in secondary education. To correctly measure students' understanding and achievements, a valid and reliable test is needed. This research focuses on the construction of valid and reliable test for the divisibility area studied at the primary school level. After constructing three pilot tests according to learning outcomes and standards for divisibility, and qualitative validation, tests were distributed in six primary schools, with 380 participating students (ages 12 – 13). The results were used for reliability and quantitative item analysis, and the final version of the test, which covered standards of students' achievement and had all items of appropriate difficulty and discriminative validity, was created. This test can be used by mathematics teachers in classrooms but also in large scale testing, like state or international testing.

Keywords: divisibility; primary school; reliability; test instrument; validity

1. Introduction

The concept of divisibility is the most important one in the elementary number theory. According to studies (Siegler et al. 2011) and (Ellis et al. 2018), there is a very strong correlation between the knowledge of divisibility that students gain in primary school and students' mathematical achievements in secondary school. Rules of divisibility are important for understanding factorization, fractions, and prime numbers. In (Roscoe & Feldman 2016) it was noticed the advantage of factorization into prime numbers as a conceptually rich tool for understanding divisibility. In (Young-Loveridge & Mills 2012) it is reported how students' understanding of the rules of divisibility affects their deeper understanding of multiplication and division of integer numbers. In (Lo 2020) it was emphasized that numbers and operations with numbers are the most common mathematical content in primary school and problematizes solving textual problems in this area.

The efficiency of teaching and learning has to be tested to see how the teaching improved students' knowledge and skills (Simanjuntak et al. 2019). A test can be conducted in a simple classroom context or as a part of a large international research (Broadfoot & Black 2004). Testing serves as a means of communication between the world of education and the wider social community, and for the consequences

of testing to be acceptable to society, the results of testing must be trustworthy (Broadfoot & Black 2004). In order to achieve that, tests have to be valid, reliable, sensitive and with appropriate difficulty levels (Simanjuntak et al. 2019).

The goal of this research is to develop a valid and reliable test for the divisibility area studied in primary school. This paper describes the development of the test according to learning outcomes and standards of students' achievements, its pilot testing and test and item analysis. The created test allows one to measure achievements in the area of divisibility at the micro-level – in a classroom, but also at the state level or even within the frame of international testing. At the moment of writing this paper, the authors have not found a similar test instrument for the divisibility area.

1.1. Divisibility Area in Primary School

In order to ensure progress and monitoring of trends in education in the world, the Agency for Pre-Primary, Primary and Secondary Education in Bosnia and Herzegovina (APOSO) defined eight educational areas, and one of them is the mathematics area[2]. In 2015, APOSO created the Common core curriculum for the mathematics area defined on the learning outcomes. The Common core curriculum for the mathematics area consists of four fields: Numbers, sets, and operations; Algebra; Geometry and measurement; Data and probability[3]. Every field has two or three components which consist of learning outcomes and indicators. Indicators are defined according to the child's age: for the end of pre-primary education, end of the third grade of primary school, end of the sixth grade of primary school, end of the ninth grade of primary school and end of secondary education.

In Bosnia and Herzegovina, students learn about divisibility in the sixth grade, within the field of Numbers, sets, and operations, so the rest of this paper will focus on this field and this grade. The field Number, sets, and operations has two components (Sets, numbers and numeral systems, and Arithmetic operations), and four basic learning outcomes[3]:

1. The student analyzes the properties and relationships of sets in different forms of representation and applies them when solving problem tasks.
2. The student analyzes the properties and relationships of numbers and numeral systems and uses symbols and different representations.
3. The student selects and combines strategies, methods, and operations to solve problems and provides solutions in the context of the problem.
4. The student evaluates the justification and precision of the chosen strategies, methods, operations, and obtained solutions, and discusses the final solution in the context of the problem.

Each of these learning outcomes has indicators according to the child's age. For

example, one indicator for outcome 3 at the end of sixth grade is "The student applies divisibility tests for positive integers with 2, 3, 4, 5, 6 and 10 (LCM, GCD)"[3].

In 2012, APOSO published standards for native language, mathematics and science for the end of third and sixth grade[1]. Standards of students' achievements are classified into low, medium and high level standards. There are 16 standards that are related to divisibility and the application of divisibility for the end of the sixth grade: two low level standards (in the rest of this paper, they will be coded as 1LL and 2LL), six medium level standards (codes 1ML, 2ML, …, 6ML), and eight high level standards (codes 1HL, 2HL, …, 8HL). These standards can be seen at the link in NOTES[4].

1.2. Validity and Reliability of Test

It is impossible to imagine learning and teaching without adequate testing and measurement. In order for the measurement to be of high quality, the test must be reliable. The test is reliable if one can trust that it will give the same or very similar results each time it is used with the same subject and one can rely on its results when making inferences (Husremović 2016). Test reliability refers to the consistency of results one gets from testing.

The most used measure of test reliability is Cronbach's alpha, which measures the internal consistency of a test and is expressed as a number between 0 and 1. The internal consistency describes the extent to which all questions on the test measure the same construct and is related to the interconnectedness of the items in the test (Tavakol & Dennick 2011). It is commonly used that the acceptable value of Cronbach's alpha is 0.70. The lower value of Cronbach's alpha indicates there are not enough items on the test, the items are poorly connected or the measured construct is heterogeneous. The value of Cronbach's alpha that is too high (> 0.90) indicates that the test contains redundant items and can be shortened (Tavakol & Dennick 2011).

Reliable results however do not guarantee that the test measures what it is supposed to measure (Darr 2005a). The most important characteristic of test instruments is their validity, which, to put it simply, is the degree to which a test measures what it is intended to measure (Husremović 2016). There is content validity, construct validity, criterion validity and consequential validity (Darr 2005b). If one wants to determine whether students achieve desired learning outcomes, content validity should be considered. Content validity indicates the compatibility of the test content and the content of learning. It is usually verified qualitatively, in a way that a panel of experts gives its opinion on the relevancy of the test and whether the items are clear, understandable and solvable

(Husremović 2016). To quantitatively analyze the content validity, one uses quantitative item analysis and determines item sensitivity parameters like difficulty and discriminative validity (Husremović 2016).

Item difficulty index (P) is determined after the test has been administered, tests are marked and participants are ranked according to their total test score. One-third of the participants make the "higher" group – those are the participants with the best scores. One-third of the participants, with the lowest scores, make the "lower" group. Item difficulty index is then calculated according to the formula

$$P = \frac{H + L}{N} \cdot 100$$

where $H$is the number of correct answers on that item in the "higher" group, L is the number of correct answers in the "lower" group and N is the total number of all answers (correct and incorrect ones) on that item in both groups (Patel 2017). Items with index difficulty $P < 30\%$ are considered too difficult, items with an index between 30% and 70% are acceptable (difficulty index between 50% and 60% is considered an ideal one), and items with index $P > 70\%$ are considered too easy (Patel 2017).

Item sensitivity or discriminative validity indicates how much that item differentiates the students according to what the test measures. The extreme groups' method is one way to determine discriminative validity (Husremović 2016). Participants are ranked according to their total test scores and the results of the "higher" and the "lower" thirds are observed. Item discrimination index (d) is calculated using the formula:

$$d = 2 \cdot \frac{H - L}{N}$$

where H is the number of correct answers on that item in the "higher" group, L is the number of correct answers in the "lower" group and N is the total number of all answers on that item in both groups(Patel 2017). Items with a discrimination index $d \leq 0.2$ have poor discriminative validity, items with a discrimination index between $0.21$and$0.24$are acceptable, items with an index between 0.25 and 0.35 are goodand items with a value of index $d \geq 0.36$ are excellent (Patel 2017).

2. Methodology

2.1. Participants

380 students from six primary schools in Sarajevo (ages 12 – 13) participated in this research. Since at the time of the research they were minors, consent for conducting the research was requested and obtained from the Ministry of

Education, as well as consent from the management of the schools where the research was conducted. Participating students had 45 minutes for the test and they did it in the presence of their mathematics teacher and the first author of this paper.

2.2. Tests

Three tests (Test 1, Test 2, and Test 3), with 12 items each, were created for the needs of this research. Each of these tests had two variants (Test 1-A and Test 1-B, Test 2-A and Test 2-B, Test 3-A and Test 3-B). Tests 1-A and 1-B had the same items but in a different order. The same holds for Tests 2-A and 2-B, and Tests 3-A and 3-B. This is a methodology APOSO uses inits testings, and the idea is that all items should be in the first half of some variant of the test (so that it does not happen that some item is poorly done just because it is at the end of the test). Within one class, participating students were not given the same test, i.e., within one class all three tests, Test 1, Test 2 and Test 3, were distributed.

The selection of items in the tests according to which standards they belong was as follows: there was one item for each standard of low (two standards) and medium level (six standards), and there were four items for different standards of high level. One item for each standard of high level was not chosen, because then the test would have 16 items, and half of them would be high-level items, i.e., more difficult items. Therefore, the test would not be suitable for use in a 45-minute school lesson. When creating the tests, approved literature for the curriculum according to which the students attend classes was used.

2.3. Qualitative Validation

The tests were reviewed by three experts: two university professors, of which one was an expert in number theory and the other in geometry and mathematical education, while the third expert was a mathematics teacher who worked several years in a state education agency. Experts suggested changing the wording in some items, in order to make them clearer. After changes in wording, tests were administered in schools.

2.4. Results and Discussion

The goal of this research was to create a reliable and valid test that would correspond to standards in divisibility. Out of 36 items appearing in three tests, 12 items that give the best validity and reliability had to be chosen. In addition to the results obtained from the quantitative item analysis, the criterion of representation of the selected standards also had to be met. For the quantitative analysis, the statistical software SPSS Statistics 20.0.0 was used.

Participants' tests were reviewed and each item was scored with 1 or 0 points, depending on whether they were done correctly or not. In total, 128 participants were given some variant of Test 1, 127 participants were given some variant of Test

2 and 125 were given some variant of Test 3. Since both variants of Test 1 had the same items, only in a different order, during the analysis the order of items in the two variants was adjusted and the results were observed as if it were a single test. The same was done for Test 2 and Test 3.

2.5. Reliability and Item Analysis

For Test 1 Cronbach's alpha was 0.692, for Test 2 it was 0.744 and for Test 3 it was 0.738. The values of Cronbach's alpha for Test 2 and Test 3 were acceptable, while the value of Cronbach's alpha for Test 1 showed this test needed some minor modifications and changes in items (Patel 2017).

For each item from Test 1, Test 2 and Test 3, the associated level and standard of the item, its difficulty index (P) and discrimination index d are presented in Table 1.

As can be seen from Table 1, only two items had poor discriminative validity ($d \leq 0.2$): item 11 from Test 1 ($d = 0.19$) and item 7 from Test 3 ($d = 0.14$). Both items had difficulty index $P < 30\%$. Since these items were too difficult and of poor discriminative validity, they were not considered as options for the final version of the test.

Table 1. Quantitative item analysis

Test and item	Item's standard and level	Difficulty index (P)	Discrimination index (d)
Test 1– item 1	3ML	29.1%	0.49
Test 1 – item 2	2HL	43.0%	0.63
Test 1 – item 3	1LL	54.7%	0.44
Test 1 – item 4	1ML	47.7%	0.35
Test 1 – item 5	5HL	45.3%	0.44
Test 1 – item 6	6ML	10.5%	0.21
Test 1 – item 7	2ML	65.1%	0.60
Test 1 – item 8	2LL	75.6%	0.40
Test 1 – item 9	6HL	33.7%	0.53
Test 1 – item 10	4ML	39.5%	0.65
Test 1 – item 11	8HL	9.3%	0.19
Test 1 – item 12	5ML	39.5%	0.56
Test 2 – item 1	2ML	64.3%	0.57
Test 2 – item 2	4HL	17.9%	0.36
Test 2 – item 3	2LL	77.4%	0.36
Test 2 – item 4	5ML	41.7%	0.40
Test 2 – item 5	2HL	41.7%	0.55
Test 2 – item 6	1ML	42.9%	0.67

Test 2 – item 7	6ML	28.6%	0.57
Test 2 – item 8	1LL	86.9%	0.26
Test 2 – item 9	8HL	13.1%	0.26
Test 2 – item 10	3ML	56.0%	0.50
Test 2 – item 11	1HL	54.8%	0.76
Test 2 – item 12	4ML	36.9%	0.40
Test 3 – item 1	6ML	31.3%	0.58
Test 3 – item 2	4HL	32.5%	0.41
Test 3 – item 3	1LL	63.9%	0.65
Test 3 – item 4	1ML	50.6%	0.63
Test 3 – item 5	3HL	36.1%	0.67
Test 3 – item 6	5ML	30.1%	0.51
Test 3 – item 7	2ML	9.6%	0.14
Test 3 – item 8	2LL	67.5%	0.39
Test 3 – item 9	5HL	25.3%	0.41
Test 3 – item 10	3ML	77.1%	0.48
Test 3 – item 11	7HL	30.1%	0.60
Test 3 – item 12	4ML	60.2%	0.39

Items 1 and 6 from Test 1, items 2, 7 and 9 from Test 2, and item 9 from Test 3 were also too difficult. Item 8 from Test 1, items 3 and 8 from Test 2, and item 10 from Test 3 were too easy (P >70%). The other 24 items had at least an acceptable difficulty index and an acceptable discrimination index.

The next step was to choose 12 test items in a way that two of them represent different low-level standards, six items represent different medium-level standards, and four items represent different high-level standards. If for some standard there was more than one option, the one with a better difficulty index and/or better discrimination index would be chosen.

2.6. Final Selection of Items

Starting from items representing low-level standards, one can notice there were two options for standard 1LL: item 3 from Test 1 ($P = 54.7\%, d = 0.44$) and item 3 from Test 3 ($P = 63.9\%, d = 0.65$). Both items had excellent discriminative validity, but item 3 from Test 1 had an ideal difficulty index, so it was chosen for the final version of the test. On the other hand, there was only one option for standard 2LL, item 8 from Test 3. The other two items representing this standard were discarded since they were too easy according to their difficulty index.

Using this idea, the following items representing medium-level standards were chosen: for standard 1ML – item 4 from Test 1, for 2ML – item 1 from Test 2, for 3ML – item 10 from Test 2, for 4ML – item 12 from Test 3, for 5ML – item 4 from

Test 2, and for 6ML – item 1 from Test 3.

The following step was to choose four items representing different high-level standards. First, there were 12 high-level items (four high-level items in each test), but after eliminating items that were too easy or too difficult, or with poor discriminative validity, there were eight high-level items left. Tests where those items were initially placed, the standards they represent, and their difficulty and discrimination indices are presented in Table 2.

For the final test, the following items were chosen: items 2 (2HL) and 5 (5HL) from Test 1, item 11 from Test 2 (1HL), and item 11 from Test 3 (7HL). This completed the test with 12 items that covered all low level and medium level standards, and four different high level standards.

Besides covering desired standards, all chosen items had at least acceptable difficulty (three had ideal difficulty), and 11 out of these 12 items had excellent discriminative validity. The remaining item had good discriminative validity. The final version of the test is given at the link in NOTES[5].

Table 2. Difficulty and discrimination indices of eight highlevel items

Standard	Test 1	Test 2	Test 3
1HL		Item 11 $P = 54.8\%, d = 0.76$	
2HL	Item 2 $P = 43\%, d = 0.63$	Item 5 $P = 41.7\%, d = 0.55$	
3HL			Item 5 $P = 36.1\%, d = 0.67$
4HL			Item 2 $P = 32.5\%, d = 0.41$
5HL	Item 5 $P = 45.3\%, d = 0.44$		
6HL	Item 9 $P = 33.7\%, d = 0.53$		
7HL			Item 11 $P = 30.1\%, d = 0.60$

3. Conclusion

This research focused on constructing a reliable and valid test for the divisibility area covered in primary school. The goal was to develop a test for efficient assessment of students' understanding of divisibility, as well as for identification of eventual difficulties or deficiencies in their knowledge.

The first step in the test construction was determining learning outcomes and standards covering the divisibility area learned in the sixth grade of primary school.

Then, three versions of the test were developed, which were administered to participants, after qualitative validation by three experts.

The reliability of tests, i.e., consistency of testing results, was checked using Cronbach's alpha. The validity of tests, i.e., how precisely and correctly tests measured what they were intended to measure, was checked qualitatively (experts' opinions) and quantitatively (item analysis). Using difficulty and discrimination indices, and the idea that the test had to have 12 items covering all low-level and medium-level standards, as well as four different high-level standards, the final version of the test was constructed.

The results of this research show that the constructed test for the divisibility area is a reliable and valid test instrument. It has the potential to be used in classroom settings or at the state level, as an effective tool for the assessment of students' understanding of divisibility, which can provide useful information to teachers and school administration.

This test can be improved or changed according to the testing needs. One option is to put open-ended questions. In further research, the plan is to use this test to determine students' achievements in divisibility depending on the curriculum according to which they attend classes in Bosnia and Herzegovina.

NOTES

1. AGENCY FOR PRE-PRIMARY, PRIMARY AND SECONDARY EDUCATION – APOSO, 2012. Stručno izvješće: Definiranje standarda učeničkih postignuća za treći i šesti razred devetogodišnjeg obrazovanja iz bosanskog/hrvatskog/srpskog jezika, matematike i prirodnih znanosti [The expert report: Defining student achievement standards for the third and sixth grade of primary school in the Bosnian/Croatian/Serbian language, mathematics and science]. Online. Sarajevo: Agency for Pre-Primary, Primary and Secondary Education. Available from: https://aposo.gov.ba/sadrzaj/uploads/definiranje-standarda-ucenickih-postignuca-za-3.-i-6.-razred-devetogodisnjeg-obrazovanja-HRVATSKI.pdf

2. AGENCY FOR PRE-PRIMARY, PRIMARY AND SECONDARY EDUCATION – APOSO, 2015a. Smjernice za provedbu zajedničke jezgre nastavnih planova i programa definirane na ishodima učenja [Guidelines for the implementation of Common core curriculum defined on learning outcomes]. Online. Sarajevo: Agency for Pre-Primary, Primary and Secondary Education. Available from: https://aposo.gov.ba/sadrzaj/uploads/Smjernice-za-provedbu-ZJNPP-1.pdf

3. AGENCY FOR PRE-PRIMARY, PRIMARY AND SECONDARY

EDUCATION – APOSO, 2015b. Zajednička jezgra nastavnih planova I programa za matematičko područje definisana na ishodima učenja [Common core curriculum for mathematics area defined on the learning outcomes]. Online. Sarajevo: Agency for Pre-Primary, Primary and Secondary Education. Available from: https://aposo.gov.ba/sadrzaj/uploads/ZJNPP-matematičko-područje-BOSANSKI.pdf

4. https://drive.google.com/file/d/1xgOqsWgt-nLjc4NCJBjKUmQMpwiUUAJp/view
5. https://drive.google.com/file/d/1xjlYY1k35pY_q4W_eHVEcBjM75qR6Arp/view?pli=1

REFERENCES

BROADFOOT, P. &BLACK, P., 2004. Redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy & Practice*, vol. 11, no. 1, pp. 7 – 26. Available from: https://doi.org/10.1080/0969594042000208976

DARR, C., 2005a. A hitchhiker's guide to reliability. *SET: Research Information for Teachers*, vol. 3, no. 5, pp. 59 – 60. Available from: https://doi.org/10.18296/set.0623

DARR, C., 2005b. A hitchhiker's guide to validity. *SET: Research Information for Teachers*, vol. 2, pp. 55 – 56. Available from: https://doi.org/10.18296/set.0639

ELLIS, A., et al., 2018. Sharing as a model for understanding division. *Neuroreport*, vol. 29, no. 11, pp. 889 – 893. Available from: https://doi.org/10.1097/WNR.0000000000001050

HUSREMOVIĆ, Dž., 2016. *Basics of psychometrics for psychology students* (in Bosnian, *Osnove psihometrije za studente psihologije*). Sarajevo: Faculty of Philosophy.

LO, W.Y., 2020. Unpacking Mathematics Pedagogical Content Knowledge for Elementary Number Theory: The Case of Arithmetic Word Problems. *Mathematics*, vol. 8, no. 10, 1750, pp. 1 – 13. Available from: https://doi.org/10.3390/math8101750

PATEL, R.M., 2017. Use of Item analysis to improve quality of Multiple Choice Questions in II MBBS. *Journal of Education Technology in Health Sciences*, vol. 4, no. 1, pp. 22 – 29. Available from doi: 10.18231/2393-8005.2017.0007

ROSCOE, M. & FELDMAN, Z., 2016. Strengthening prospective elementary

teachers' knowledge of divisibility: An interventional study. In: *NCTM Research Conference*, 11-13 April 2016, San Francisco. Available from: https://nctm.confex.com/nctm/2016RP/webprogram/Manuscript/Session42256/RoscoeFeldman2016.pdf

SIEGLER, R.S., THOMPSON, C.A. & SCHNEIDER, M., 2011. An integrated theory of whole number and fractions development. *Cognitive Psychology*, vol. 62, no. 4, pp. 273 – 296. Available from: doi.org/10.1016/j.cogpsych.2011.03.001

SIMANJUNTAK, E., HUTABARAT, H.D.M. & HIA, Y., 2019. The effectiveness of test instrument to improve mathematical reasoning ability of mathematics student. *Journal of Physics: Conference Series*, vol. 1188, no. 1, p. 012048. Available from: doi.org/10.1088/1742-6596/1188/1/012048

TAVAKOL, M. & DENNICK, R., 2011. Making sense of Cronbach's alpha. *International journal of medical education*, vol. 2, pp. 53 – 55. Available from doi.org/10.5116/ijme.4dfb.8dfd

YOUNG-LOVERIDGE, J. & MILLS, J., 2012. Deepening students' understanding of multiplication and division by exploring divisibility by nine. *The Australian Mathematics Teacher*, vol. 68, no. 3, pp. 15 – 20. Available from: https://core.ac.uk/download/pdf/29201407.pdf

✉ **Dr. Daniela Zubović, Senior expert associate**
ORCID iD: 0000-0002-4076-3075
Faculty of Science, University of Sarajevo
Zmaja od Bosne 33 – 35
71 000 Sarajevo, Bosnia and Hercegovina
E-mail: dzubovic@pmf.unsa.ba

✉ **Dr. Dina Kamber Hamzić, Assist. Prof.**
ORCID: 0000-0003-4846-2733
Faculty of Science, University of Sarajevo
Zmaja od Bosne 33 – 35
71 000 Sarajevo, Bosnia and Hercegovina
E-mail: dinakamber@pmf.unsa.ba