Development and evaluation of a numerosity estimation test for elementary school students

Leonie Brumm, Elisabeth Rathgeb-Schnierer

Development and evaluation of a numerosity estimation test for elementary school students

Leonie Brumm[1] and Elisabeth Rathgeb-Schnierer[1]

[1]University of Kassel, Germany; brumm@uni-kassel.de

Numerosity Estimation is not only important for everyday activities but is also assumed to be important for the development of mathematical skills. To examine influencing factors on numerosity estimation, it is required to conduct numerosity estimation accuracy in different tasks. Since no standardized numerosity estimation test exists to date, the goal is to develop a standardized test that reliably and validly measures the accuracy of numerosity estimation of elementary school children. Therefore, we developed a digital numerosity estimation test including seven different types of tasks in three number ranges. 196 German third-grade students from thirteen classes participated in the numerosity estimation test. In this paper, we will describe the test development and the evaluation of testing construct validity and reliability as specific test theory-based quality criteria.

Keywords: Estimation, estimation test, elementary school mathematics, mathematics education, numerosity estimation.

Introduction

Estimation is relevant for everyday activities in the lives of children and adults as well as a core skill in everybody's life (Andrews et al., 2021; Siegler & Booth, 2005). In general, estimation is defined as mental comparison and measurement (Schipper, 2009). This project focuses on numerosity estimation as one out of four types of estimation (Sayers et al., 2020). Numerosity estimation describes the aspect of estimating discrete quantities (Andrews et al., 2021; Crites, 1992) and "is considered as a perceptual process that leads to relatively quick but less accurate numerosity judgments" (Luwel & Verschaffel, 2008, p. 320). Thus, an estimation encourages methods that lead to reasonable results, but no exact answer is required (Schipper, 2009).

Researchers emphasize that estimation is a complex problem-solving process requiring flexible thinking (Siegler & Booth, 2005; Luwel & Verschaffel, 2008). This goes in line with Crites (1992) who focuses on numerosity estimation.

> "Students who made accurate estimates tended to have good number sense and metacognitive skills, isolated important components of a problem while ignoring irrelevant information, were aware of the reasonableness of their answers, and were flexible in their thinking" (Crites, 1992, p. 614).

This quote also addresses a reference to number sense. This coincides with the agreement that (numerosity) estimation is related to number sense (e.g., Sayers & Andrews, 2015; Crites, 1992). Sayers et al. (2016) identified estimation as one of eight components of foundational number sense which Sayers and Andrews (2015) conceptualized for first-grade students. Foundational number sense describes a set of number-related essential competencies that require instruction (Sayers & Andrews, 2015). It is assumed to be substantial for estimation, but also for understanding mathematics.

Overall, it is adopted that estimation is connected to several aspects of mathematical skills, particularly arithmetic skills (Siegler & Booth, 2005). Fostering estimation skills can greatly impact the development of mathematical skills (e.g., Luwel et al., 2005; Siegler & Booth, 2005). In this vein, estimation is considered a determinant of later mathematical, especially arithmetical achievement (Andrews et al., 2021; Siegler & Booth, 2005). Results suggest that students who are gifted estimators show better math achievement (Booth & Siegler, 2006) and strategy flexibility (Siegler & Booth, 2005; Luwel & Verschaffel, 2008). Bartelet et al. (2014) account that Kindergarten students' efficiency in numerosity estimation explains a unique part of the variance in arithmetic achievement in first class. Nevertheless, no significant association between estimation and arithmetic achievement was found (Bartelet et al., 2014). Moreover, Wong et al. (2016) argue that different estimation abilities affect arithmetic achievement for six-year-old children. Furthermore, Barth et al. (2009) indicate that counting is fundamental for a successful process of numerosity estimation. However, the relationship between math achievement and numerosity estimation has been little studied, especially with respect to different types of tasks that map numerosity estimation. In order to quantitatively investigate this relationship or other factors influencing numerosity estimation accuracy, an instrument is needed to measure estimation accuracy in a large sample.

Crites has designed a numerosity estimation test for third-, fifth-, and seventh-grade students to examine the strategies of "skilled and less skilled estimators" (1992, p. 601). The items from this test require more advanced knowledge which influences the estimation accuracy. For example, one item asks to estimate how many students are at the school. To be able to solve this item, knowledge about the school structure is presumed. In fact, some of the quantities to be estimated are in a very high number range, up to five million. In another item, the estimation of lengths is asked at the same time (Crites, 1992). Irrespective of these constructional features, the evaluation by Crites (1992) showed that the test is not reliable for third-grade students. Last, this test is based on individual interviews which is not the most efficient way to survey estimation accuracy within a large sample.

Although the importance of estimation is repeatedly emphasized in the literature, there is not a broad body of studies on numerosity estimation. Most studies focusing on numerosity estimation only consider one type of numerosity estimation task to capture accuracy. There are hardly any current studies investigating accuracy in numerosity estimation by considering different types of estimation tasks (e.g., two- and three-dimensional quantities in a structured or unstructured way). Furthermore, there is no test that reliably measures estimation accuracy in elementary school with respect to different types of perception tasks. To fill this gap, one aim of our study was to develop and evaluate a standardized online estimation test that does not require any advanced knowledge, that includes numbers in a number range that is realistic to adequately estimate for third graders, that can be administered to a whole class within a single lesson, and that addresses different types of numerosity estimation tasks. The test development and evaluation are embedded in a broader project, in which we also investigate the influence of various non-cognitive constructs on numerosity estimation accuracy and students' strategies as well as the relationship between math achievement and numerosity estimation accuracy (Brumm & Rathgeb-Schnierer, 2023). However, in this paper we emphasize one part of the project and focus on research questions concerning the test development and evaluation:

RQ1: Which subconstructs regarding content can be identified in the numerosity estimation test?

RQ2: How reliable is the numerosity estimation test for third-grade students?

RQ3: To what extent are the subconstructs related to each other?

Method

This project includes a sample of 196 German third-grade students who participated in the numerosity estimation test. Their ages ranged from 8.7 years to 11.1 years (M = 9.5, SD = 0.4). 98 students were girls (approximately 51%), and 96 were boys (approximately 49%). The total sample of students comes from five public schools and 13 classes. The data was collected during school hours from May 2022 to July 2022. The digital numerosity estimation test including a digital questionnaire was administered in one lesson (45 minutes) per class. This survey of the test took place within a broader project (Brumm & Rathgeb-Schnierer, 2023). However, the most important design features are explained here in order to be able to subsequently answer the questions of this part study.

Before the test started, we asked the students openly what they understood regarding the term ‘estimation’ and ended this phase with a standardised definition of estimation. Then we explained the test structure and gave instructions on how to answer the test (e.g., response format and time schedule). Afterward, the students carried out two test items and had the possibility to ask questions if they had problems with the test items. The test presents perception tasks. Thus, the number of elements shown in a picture is to be estimated. A crucial aim in the test development was to exclude tasks requiring knowledge external to estimation. In addition, various task characteristics were considered within the item construction of this test after researching different characteristics in the literature, summarising them, and finally structuring them. Accordingly, the quantity to be estimated can be either two-dimensional or three-dimensional, including structured or unstructured elements, and the elements, in turn, can be the same or different. Thus, a broad range of characteristics is covered to assure content validity. Since the numbers are represented by a picture, the difficulty with three-dimensional quantities is to first perceive them as three-dimensional in order to estimate the number of elements subsequently. Figure 1 displays an example of a picture that shows a three-dimensional quantity with structured arranged and equal elements as well as a two-dimensional quantity where the elements are structured and equal.

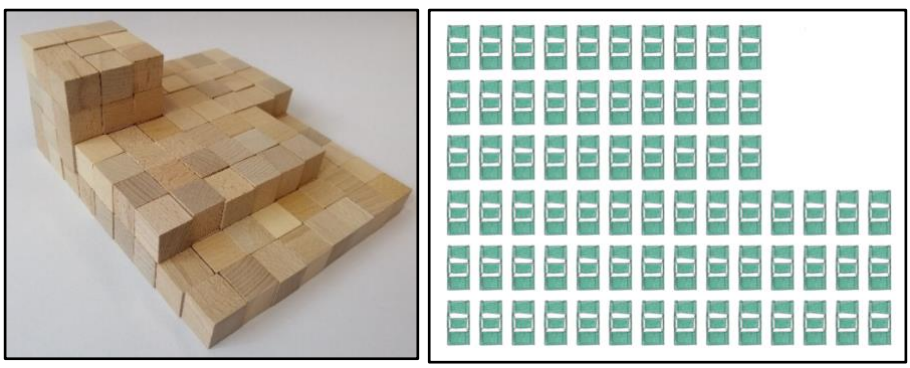

Figure 1: Example: three-dimensional, structured, equal elements and two-dimensional, structured, equal elements

Altogether, a total of seven different types of tasks are implemented in the test, with each type of task presented in three different number ranges (up to 50, 50 to 100, and up to 150). Thus, the test includes 21 items. In some cases, items are based on the items Luwel and Verschaffel (2008) used in their study.

In Germany, the number range is extended from 100 to 1000 in the third grade. Consequently, the chosen number range is assumed to be appropriate for third-grade students. To avoid counting as a sole strategy, on the one hand, each task was presented for 20 seconds (Luwel & Verschaffel, 2008). On the other hand, the quantity to be estimated in the range between 31-144 hinders counting all elements (Albarracín & Gorgorió, 2019). The picture disappears after 20 seconds, but the students have 40 additional seconds to adjust their results. Accordingly, the students have a total of one minute to estimate a quantity. For each item, the students can set a number between 0 and 500 as the result with a slider. The test introduction and the test execution take about half an hour.

In our project (e.g., Brumm & Rathgeb-Schnierer, 2023), we calculated the value for an item of the numerosity estimation test as follows:

$$\frac{Actual\ value - Estimated\ value}{Standard\ deviation\ (in\ relation\ to\ acutal\ value\ instead\ of\ mean\ value)}$$

The actual value corresponds to the number of elements to be estimated and the estimated value mirrors the number that the students estimated. Overall, this calculation reflects a standardized absolute deviation because we assume that the deviation increases with a larger number of elements. Accordingly, we assume that the size has an influence on the deviation and relative errors occur more frequently with small quantities than large ones. For this reason, we assume that size also affects the mean, which is why the standard deviation was calculated concerning the actual value.

During test development, we verified the validity and reliability of the test as specific test theory-based quality criteria (Moosbrugger & Kelava, 2020). Due to that, we measured the construct validity by exploratory factor analysis. We chose the principal-axis factor analysis using oblimin rotation to allow the factors to correlate. The principal-axis analysis is one of the most frequently used extraction methods (Brandt, 2020). Among others, the Kaiser-Meyer-Olkin test was calculated. The Kaiser-Meyer-Olkin value is greater than .9 signifying that the sample is adequate to conduct an exploratory factor analysis. The number of factors was extracted based on the Sree-Test and eigenvalue (Brandt, 2020). The reliability was assessed with Cronbach's Alpha. To answer our third research question, a correlation analysis (Pearson's *r*) was conducted.

Results

Four factors were extracted in the exploratory factor analysis. Six items load solely on one factor (*I13_3uu_35* with .74; *I19_2se_45* with .65; *I15_2ud_43* with .60; *I2_3se_31* with .58; *I17_2sd_37* with .52; and *I7_2ue_45* with .50, see the note of Table 1 for the explanation of the abbreviations). Another five items load exclusively on the second factor (*I8_2ud_132* with .61; *I5_2se_126* with .53; *I10_2sd_108* with .52; *I4_3ue_90* with .48; and *I3_2sd_82* with .48).

The remaining results of the exploratory factor analysis of the additional ten items are shown in Table 1.

Table 1: Results of the exploratory factor analysis

	Factor			
	1	2	3	4
I6_3ud_68			.71	
I14_2ue_123			.69	-.40
I12_2se_72	.33		.54	
I11_3ue_42	.39		.50	
I1_2ud_64			.46	.33
I16_3sd_140				-.68
I20_3ud_112				-.56
I21_2ue_84	.37			-.50
I18_3ue_144				-.49
I9_3se_72				-.38
Note. I = Item, 3 = three-dimensional, 2 = two-dimensional, s = structured, u = unstructured, e = equal elements, d = different (unequal) elements. The number at the end of the item description represents the quantity to be estimated.				

After the factor analysis, we decided to eliminate three of the 21 items (*I1*, *I4*, *I21*) due to similar double loading, content fit, or negatively influencing the reliability. It becomes clear that one of the items (*I11_3u_42*), which loads positively on two factors, also belongs to the comparatively small number range and would therefore also fit the first factor in terms of content. Due to the clearly higher loading on factor 3, we decided to include this item in the *Mix* scale resulting from factor 3. The latter also applies to item *I12_2se_72*. All items of one factor were tested again with a principal component analysis to ensure content unidimensionality within that factor. Consequently, the reliability of one scale of items of one factor was measured by Cronbach's Alpha. Table 2 shows the four resulting scales incorporating the factor scores as measures for accuracy in numerosity estimation.

Table 2: Reliability of estimation scales

Scale	Items	Cronbach's Alpha α
SmallN	6	.78
2DlargeN	4	.69
Mix	4	.75
3DlargeN	4	.71

The first scale *SmallN* represents six items in the number range from 31 to 45 and has an internal consistency of α = .78. Scale *2DlargeN* shows a reliability of α = .69 and includes four items in the number range from 82 to 132 which are all two-dimensional. Scale *2DlargeN* is the only one that has questionable reliability. The third scale, *Mix*, also comprises four items with an internal consistency of α = .75. Two items in *Mix* are in the number range of around 70, one contains 123 elements to estimate, and another one shows 42 unstructured elements. In this scale, two items are arranged in an unstructured way and two are arranged in a structured way. Finally, the scale *3DlargeN* (α = .71) comprises four three-dimensional items in the number range from 72 to 144. Table 3 displays the correlations between these four estimation scales.

Table 3: Pearson correlation between the four estimation scales

	1.	2.	3.	4.
1. SmallN	--			
2. 2DlargeN	.08	--		
3. Mix	.49**	.10	--	
4. 3DlargeN	.33**	.41**	.27**	--
*. $p < .05$. **. $p < .01$.				

There is a medium positive correlation between Scale *SmallN* and *3DlargeN* ($r = .33$, $p < .01$), Scale *SmallN* and Scale *Mix* ($r = .49$, $p < .01$), as well as Scale *2DlargeN* and Scale *3DlargeN* ($r = .41$, $p < .01$), which is highly significant. Furthermore, a highly significant, weak positive correlation exists between scale *Mix* and scale *3DlargeN* ($r = .27$, $p < .01$). The only two pairs of scales with no correlation are *SmallN* and *2DlargeN* as well as *2DlargeN* and *Mix*.

Discussion

To answer which subconstructs regarding content can be identified in the numerosity estimation test and to examine construct validity, we performed an exploratory factor analysis (RQ1). Four factors became apparent through the factor analysis, resulting in four scales as a measurement tool for estimation accuracy. In the comparison, three of the four scales can be clearly distinguished from each other in terms of content. One scale contains items whereby quantities must be estimated in a comparatively low number range (*SmallN*). Another scale contains items with two-dimensional quantities in a comparatively higher number range (*2DlargeN*), and a matching scale contains only three-dimensional items in a likewise higher number range (*3DlargeN*). The last scale includes items in different number ranges, which are also not uniformly two- or three-dimensional (*Mix*). It is interesting to note that the structure of the items to be estimated does not seem to affect the response behavior of the students.

Regarding the second research question, it can be answered that three of the four scales have satisfactory reliability. Scale *2DlargeN* has questionable reliability. However, the value is very close to a satisfactory value and plausible in terms of content, which is why this scale will be included in further analyses (Brumm & Rathgeb-Schnierer, 2023).

We performed a correlation analysis to answer to what extent are the subconstructs related to each other (RQ3). It was interesting to see that the *3DlargeN* scale correlates with all three other scales. It was to be expected that *3DlargeN* would correlate with *2DlargeN* since both scales cover a similar number range. Surprisingly, *3DlargeN* correlates with *SmallN*, but *SmallN* does not correlate with *2DlargeN*. We expected scales *SmallN* and *2DlargeN*, in particular, to be related since we thought it was possible that estimation accuracy for small quantities could most likely influence estimation accuracy for two-dimensional quantities because students presumably had more experience with two-dimensional quantities than three-dimensional ones. Summing up, the subconstructs are partially related to each other, but only significantly weak or moderately. Note that the factor loadings for 3DlargeN are all negative. This means that a high score on the item is associated with a low score on the factor (Brandt, 2020).

It is important to further develop and evaluate the test. Due to the sample size, splitting the sample and performing exploratory and confirmatory factor analysis with the data set at one point was impossible. Therefore, it would be necessary to survey the test within a larger sample in order to conduct a confirmatory factor analysis to verify the factors found in the exploratory factor analysis.

The overall purpose of developing the test is to be able to examine estimation accuracy in elementary school, its development, and factors influencing estimation accuracy or relationships to other constructs in more detail. It is also easy to change the size of the numbers to be estimated to make the test accessible to even younger or older students. Regardless of test development, we intend this paper to emphasize the importance of numerosity estimation and the need for further studies in this field.

References

Albarracín, L. & Gorgorió, N. (2019). Using large number estimation problems in primary education classrooms to introduce mathematical modelling. *International Journal of Innovation in Science and Mathematics Education*, *27*(2), 45–57. https://doi.org/10.30722/IJISME.27.02.004

Andrews, P., Constantinos X., & Judy S. (2021). Estimation in the primary mathematics curricula of the United Kingdom: Ambivalent expectations of essential competence. *International Journal of Mathematical Education in Science and Technology*, *53*(8), 2199–2225. https://doi.org/10.1080/0020739X.2020.1868591

Bartelet, D., Vaessen, A., Blomert, L., & Ansari, D. (2014). What basic number processing measures in kindergarten explain unique variability in first-grade arithmetic proficiency? *Journal of Experimental Child Psychology*, *117*, 12–28. https://doi.org/10.1016/j.jecp.2013.08.010

Barth, H., Starr, A., & Sullivan, J. (2009). Children's mappings of large number words to numerosities. *Cognitive Development*, *24*(3), 248–264. https://doi.org/10.1016/j.cogdev.2009.04.001

Booth, J. L. & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Brandt, H. (2020). Exploratorische Faktorenanalyse. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]* (pp. 575–614). Springer.

Brumm, L. & Rathgeb-Schnierer, E. (2023). The relationship between accuracy in numerosity estimation, math achievement, and math interest in primary school students. *Frontiers Psychology*. *14*(1146458). https://doi.org/10.3389/fpsyg.2023.1146458

Crites, T. (1992). Skilled and less skilled estimators` strategies for estimating discrete quantities. *The Elementary School Journal*, *92*(5), 601–619. https://doi.org/10.1086/461709

Luwel, K., Lemaire, P., & Verschaffel, L. (2005). Children's strategies in numerosity judgment. *Cognitive Development*, *20*(3), 448–471. https://doi.org/10.1016/j.cogdev.2005.05.007

Luwel, K. & Verschaffel, L. (2008). Estimation of 'real' numerosities in elementary school children. *European Journal of Psychology of Education*, *23*(3), 319–338. https://doi.org/10.1007/BF03173002

Moosbrugger, H. & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen ("Gütekriterien"). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]* (pp. 13–38). Springer.

Sayers, J., Petersson, R., Rosenqvist, E., & Andrews, P. (2020). Estimation: An inadequately operationalised national curriculum competence. In R. Marks (Ed.), *Proceedings of the British Society for Research into Learning Mathematics*, *40*(1). BSRLM.

Sayers, J., Andrews, P., & Björklund Boistrup, L. (2016). The role of conceptual subitising in the development of foundational number sense. In T. Meany, O. Helenius, M. L. Johansson, T. Lange, A. Wernberg (Eds.). *Mathematics Education in the Early Years: Results from the POEM2 Conference*, 2014 (pp. 371–394). Springer.

Sayers, J. & Andrews, P. (2015). Foundational number sense: Summarising the development of an analytical framework. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of the European Society for Research in Mathematics* (pp. 361–367). Charles University in Prague.

Schipper, W. (2009). *Handbuch für den Mathematikunterricht an Grundschulen [Handbook for teaching mathematics at primary schools]*. Schroedel.

Siegler, R. S. & Booth, J. L. (2005). Development of numerical estimation: A review. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 197–212). CRC Press.

Wong, T. T.-Y., Ho, C. S.-H., & Tang, J. (2016). Consistency of response patterns in different estimation tasks. *Journal of Cognition and Development*, *17*(3), 526–547. https://doi.org/10.1080/15248372.2015.1072091