Developing a formative, digital tool to assess children's number sense when starting school

Gunnhild Saksvik-Raanes and Trygve Solstad

Norwegian University of Science and Technology, Norway; gunnhild.b.saksvik@ntnu.no, trygve.solstad@ntnu.no

We investigate how digital technologies can enrich teachers' formative assessment of number sense by describing the development process of a digital assessment tool for children starting school (five- and six-year-olds). Studying different aspects of validity, we focus on scoring and digital affordances. The quantitative analysis of the preliminary data from 101 children evidences the technical validity of the tool. We find that interactive assessment items add to the content validity of the tool. The interactive items provide qualitative data about students' number sense which cannot be captured by quantitative measures. At the same time, children may have greater difficulty interpreting more complex items. Our results support the view that a digital tool can be a useful supplement to the assessment of number sense. Further developments and approaches to investigating additional aspects of validity are discussed.

Keywords: Number sense, formative assessment, educational technology, elementary school mathematics.

Introduction

Children start school with considerable knowledge related to learning mathematics (Clarke et al., 2006). Assessing each child's number competence enables the teacher to plan effective and engaging teaching, but assessment can also be a challenging and time-consuming task. The rapid advance of digital technologies brings new opportunities for assessment in mathematics education. Developing digital formative assessment instruments has been highlighted as an area to prioritise in early mathematics education research (Ginsburg & Pappas, 2016). Specifically, to inform teachers about children's Foundational Number Sense (FoNS), Sayers et al. (2016) point out the need to "develop a diagnostic tool for teachers to assess individual grade one children's FoNS-related understanding" (p. 389).

We address this need by developing and analysing a digital assessment tool. The purpose of the tool is to inform teachers about the mathematical knowledge that children have already acquired when they start school. In a classroom setting, a digital tool can be an efficient supplement in the initial assessment process, especially since time typically does not permit one-to-one assessment interviews. In this paper, we focus on the opportunities and challenges associated with technology supplementing the teacher's formative assessment of students' number sense. We discuss how interactive tasks and records of qualitative data about the children's solutions and solution processes can strengthen the evidence for the construct validity of a digital assessment tool. Based on preliminary data from an on-going study, we address the following research question:

How do digital tasks support aspects of validity in the assessment of first-graders' number sense?

Frameworks

The FoNS model for operationalising number sense

To operationalise number sense we build on the FoNS model which describes the number-related skills that require instruction (Andrews & Sayers, 2015). The FoNS model provides a multi-layered, flexible, and relational definition of the number sense concept with eight interrelated categories: Number identification (NI), systematic counting (SC), number-quantity relationships (NQ), quantity discrimination (QD), representing numbers (RN), estimation (ES), simple arithmetic competence (AC) and awareness of number patterns (NP).

Assessment validity

In our context, assessment validity concerns the extent to which theory supports inferences made from test scores and how the evidence supports interpretations (Wolfe & Smith, 2007a). Evidence for validity can be found by looking at eight different aspects of validity: 1) content, 2) substantive, 3) structural, 4) generalisability, 5) external, 6) consequential, 7) interpretability, and 8) responsiveness (Wolfe & Smith, 2007b). In this paper, we focus on the content, substantive, and interpretability aspects of validity. The content aspect of validity addresses the relevance, representativeness, and the technical quality of the items. Documenting the purpose of the tool and the development process with expert reviews are part of the evidence that relates to the content aspect of validity. The substantive aspect of validity relies on the theoretical model we based the item development on. Here, we look at how the different items reflect children's overall number sense as described in the FoNS model. We focus on how the digital format can enhance content and substantive validity. Finally, we address how the number sense score can be interpreted by considering features that affect item difficulty.

Formative assessment of young children's knowledge

Formative assessment can be considered both an instrument and a process (Bennett, 2011). In this context, formative assessment refers to how the results of the assessment process are used to promote further learning (Black & Wiliam, 2018). Therefore, in addition to informing teachers about the children's present competence, a formative assessment tool should also support teachers in adjusting their teaching to meet the children's needs. Previous research has examined how task-based, one-to-one assessment interviews can provide crucial information to help teachers facilitate students' learning (Clarke et al., 2011). The digital medium can provide teachers with some of the features of assessment interviews, such as screen recordings of the students' solution processes on interactive items, as well as more traditional skill-based assessments characteristic of pencil-and-paper tests. Certain researchers highlight the transformative improvements that software can have on early mathematics education, both for helping children learn mathematics, and for providing guidance to teachers (Ginsburg, 2016). Hence, a digital assessment tool can guide instruction and improve students' opportunities to learn mathematics.

Methods

Procedure

About 50 schools in and around Trondheim municipality in Norway were invited to participate in the project, out of which eight of the interested schools were chosen. In this paper, we present preliminary data from 101 of the first-grade children (five to six years old) who participated in the study. A researcher visited the schools over a period of two months at the beginning of the school year. Groups of six to eight children carried out the assessment on separate tablet computers. The participants were seated so as not to get disturbed by each other's screens or sounds. All children were given the same instructions before they started the assessment and were free to finish at any time. Pre-recorded voice instructions were given for each item. For technical reasons, the assessment was presented in three separate units with different FoNS categories and increasing difficulty. Each child could decide whether to continue to the next unit or not. Most children completed the first two units, some completed all three units, and a few children completed only the first unit. There was no time limit for the items, but time on task was recorded for each item. The children typically spent between 15 and 25 minutes on the assessment in total.

Analytical procedures

Rasch measurement was used for quantitative analysis of the children's responses, using the Winsteps software. The Rasch model is a probabilistic measurement model that provides interval-scale measures of item difficulty and person skill on the same measurement scale in the unit of logits (Wright, 1977). On a basic level, Rasch analysis involves calculating the probability that a person with competence B answers correctly on an item with difficulty D. A person with higher competence always has a higher probability of successfully answering any item than a person with lower competence. An item that is more difficult, always has a lower probability of being successfully answered than an item that is less difficult, regardless of person ability.

All items were scored dichotomously, meaning that the children received one point for a correct answer and zero points for a wrong answer. In addition, to investigate and illustrate the potential of interactive items for enriching digital assessment, observations and screen recordings of the children's solutions underwent qualitative analysis.

Results and discussion

Development of the assessment tool: from framework to data collection

We developed items for each of the eight FoNS categories. The items were adapted from different cognitive and educational studies on number sense, standardised number sense tests, and formative assessment instruments (Ginsburg & Pappas, 2016; Davison et al., 2012).

A selection of items was presented to expert groups for them to adjust and determine the items that were the most suitable for measuring the number-sense construct. The expert reviews were performed by researchers in mathematics education familiar with the number-sense concept, target population and instrument development. Based on the first reviews, items from five number sense categories were selected for the first pilot study: Number identification, systematic counting, number and quantity, quantity discrimination, and arithmetic competence. Several pilot studies were conducted to refine the content and selection of items. The qualitative observations obtained from the first pilot study led to further changes, predominantly associated with technical issues and voice instructions. Certain items that were often misunderstood by the children were removed. The first pilot study was carried out at the end of the academic year, whereas the assessment was designed for children who were almost a year younger. To provide better estimates of the difficulty parameters, we conducted a second pilot study in a preschool right before the end of the academic year. This second pilot study led to further adjustments. The level of difficulty had to be adjusted to enable the overview of the number sense competence of all first graders. Subitising had also been included as a separate category at this point since conceptual subitising has been highlighted as a central aspect of number sense (Sayers et al., 2016). After analysis in expert groups, the decision was made to include the last three FoNS categories in the assessment tool: Estimation, number patterns, and representing number. The items from these categories were piloted in a fourth pilot study, along with the rest of the assessment, before commencing the main data collection in the fall of 2020. An overview of the final 78 items included in the tool is presented in Table 1. Two items did not contribute to the number sense measure as intended and were removed from the present analysis. We based the subsequent analysis on the remaining 76 items. The items varied from typical skill-based items, asking the child to identify a certain number symbol or quantity, to items capturing the child's solution process (later referred to as process items). The process items were designed to exploit the digital medium's potential for capturing some of the characteristic aspects of one-to-one assessment interviews. Of the 78 items, 10 were process items, and two of them will be discussed in more detail.

Scale	**Content**	**N(SB)**	**N(P)**
Number identification (NI)	Recognise numeral and meaning.	8	0
Systematic counting (SC)	Ordinality. Count to twenty and back from an arbitrary digit.	6	2
Number and quantity (NQ)	Cardinality. One-to-one correspondence between symbol and quantity.	6	4
Quantity discrimination (QD)	Compare quantities. Vocabulary: larger, smaller, more than, less than.	6	2
Representing number (RN)	Different representations of numbers and part - whole aspects.	4	1
Estimation (ES)	Estimate the size of a set and position on a number line.	6	0
Arithmetic competence (AC)	Operate on small sets by using addition or subtraction	14	1
Number patterns (NP)	Continue or complete a number sequence.	7	0
Subitising (SU)	Perceive quantity without counting. Perceptual and conceptual. Timed.	11	0

Table 1. Overview of the number sense items within each FoNS category. N(SB): Number of skill-based items. N(P): number of process items

Development of a specific item involving children's solution strategies

NQ10 was developed within the category of number and quantity (NQ) to assess the ability to count a number of objects and identify the number symbol that represents that amount. The item was adapted from similar items used in interview settings (Malofeeva et al., 2004). We present the first version of the item in Figure 1A. Expert groups discussed the specific number of objects that would enable insights into the child's structuring of quantities, how to give the most succinct and accurate voice instructions, and how the objects could be placed to encourage the child to manipulate them.

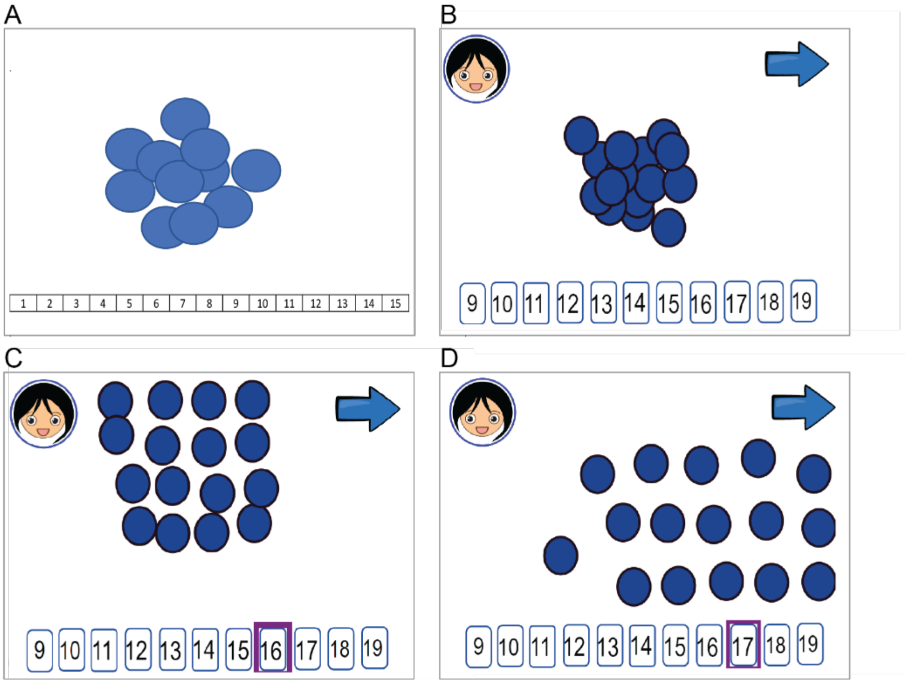

Figure 1. Development and responses to Item NQ10
A. First version; B. Final version; C. Final response of Child 1; D. Final response of Child 2

In the final version of item NQ10 (Figure 1B) the children are asked to arrange the objects in a manner that is easy to count. The software provides both a recording of the motion and the final positions, which allowed us to study how the child rearranges the objects to simplify the counting process.

In the given examples, we see how two first-graders structured the objects. Child 1 structured the 16 objects into four groups of four and selected the correct answer. Child 2 structured the objects into three groups of five, placing one object separately, and then selected 17 as the answer. If not just a random mistake, the error might be related to the number symbols (rather than the quantity involved), the aspects of cardinality, or the structure of odd and even numbers. Such access to the qualitative aspects of the responses can give the teacher valuable insight into each child's solution process and more information about the child's understanding than simply knowing whether the response was right or wrong. Therefore, these process items can provide qualitative data to support the assessment process and contribute to the evidence for both the content and substantive aspects of validity. Process items cannot replace the insights gained through human interaction in one-to-one assessment interviews. However, the qualitative recordings can support the validation of the assessment and be valuable for screening, individual follow-up, and teaching purposes.

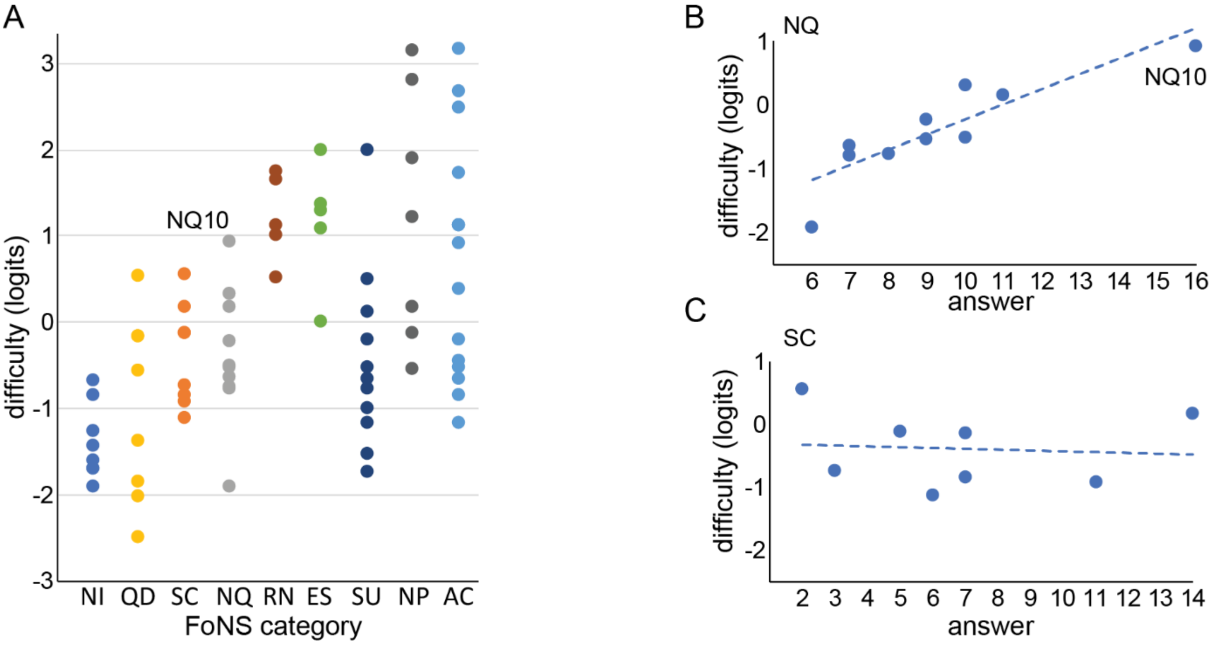

Figure 2. Item difficulty. A. Plotted as a function of the FoNS categories arranged by maximum item difficulty. B, C. Plotted as a function of the correct answer to the item in the categories number and quantity (NQ) and systematic counting (SC). The dashed line is the linear regression line. Item NQ10 is indicated in the plots (difficulty = 0.93; answer = 16). The abbreviations are defined in Table 1

Test quality: technical aspects of construct validity

The person reliability value of the 76 items analysed in this paper was 0.88 (corresponding to a Cronbach's alpha of 0.91), which is typically considered productive for measurement.

As a whole, the items were well targeted to the group of school-starters and were slightly easy, having the value of 0 ± 1.31 logits (mean ± sd), while the children scored 0.51 ± 0.93 logits (mean ± sd). The larger standard deviation of the items indicates that only a few children scored at the top and bottom of the measurement scale. This gives the assessment tool the necessary range to track the children's number sense as it develops during their first year of school, for example, by comparing the assessments from the first and second half of the school year.

If a single aggregate score is used to measure number sense, individual items need to measure the same number sense construct. Evidence for this can be found in the infit and outfit measures of the items. An item with fit values close to 1 is considered to measure the same construct as the rest of the items. For this assessment tool, most items fit well with the Rasch model, with a mean item infit mnsq of 0.98 ± 0.15 (mean ± sd) and mean item outfit mnsq of 1.02 ± 0.42.

The group of process items also measured the same construct as the other items. One potential exception was item RN4, with infit mnsq 1.21 and outfit mnsq of 2.2, which is high. Excluding RN4, the process items had a mean infit mnsq of 1.02 ± 0.11 and mean outfit mnsq of 0.99 ± 0.17.

Item RN4, in which the child uses dice to compose the quantity ”four” in three different ways, was adapted from a composing number task used in interview settings (Clements et al., 2008). Formulating clear instructions for such complex tasks was a general challenge of the item development and might be related to the misfit of this item. Representing numbers seems to be different from other number sense categories, as three out of five items from this category had infit values greater than 1.2 (one process and two skill-based items). One reason for items from this category to stand out as more misfitting than items form other categories

might be the complexity of this domain. Representing numbers in different ways is a rich domain which includes several other categories of number sense, such as arithmetic competence and connections between number and quantity (Andrews and Sayers, 2015).

How can the number sense score be interpreted? Here, we give two examples supporting the interpretability aspect of validity of the assessment. First, different FoNS categories had different ranges of difficulty (Figure 2A). This means that a child's aggregated measure is indicative of whether that child can solve the tasks from each category. Second, for some FoNS categories, such as number and quantity, the difficulty correlated with the numerical value of the answer (Figure 2B). The relation between numerical value and item difficulty suggests that the aggregate measure is predictive of the range of numbers the child can process confidently. For other categories, such as systematic counting, the difficulty was not clearly related to the numerical value of the answer (Figure 2C), indicating that task difficulty was largely determined by the content of the tasks in some categories.

Conclusions

Presenting parts of an on-going study, we have described the development of a tool to assess first grader's number sense. The digital assessment tool is based on the FoNS model (Andrews & Sayers, 2015) and has been subjected to expert reviews. The use of interactivity provides opportunities for simulating aspects of interview tasks, which adds to the content aspect of the validity of the assessment. The Rasch analysis of the first graders' responses indicates that the technical quality of the assessment tool is high. Taken together, these results indicate that a digital assessment tool has the potential to provide teachers with information about their students' number sense.

We argue that the digital format can supplement the teachers' informal assessments and offer a valid, reliable, and more complete alternative to paper-and-pencil assessment by incorporating aspects of one-to-one assessment interviews. The use of interactive process items gives us the opportunity to adapt tasks that were previously reserved for one-to-one assessments to a setting in which teachers can gain information about their students' number sense in a less time-consuming manner. It remains to investigate how the digital assessment may affect test conditions and whether the level of engagement in the digital assessment is different from that of comparable written assessments.

An important validity aspect of a formative assessment tool is how the tool is used to improve children's learning. Describing formative assessment as a process and validating a tool to be used in this process, entails presenting the scores in a manner that is useful for the teacher. In a classroom setting, the teacher's informal assessments are often based on several different interpretations, which leads to a measurement issue in the formative assessment (Bennett, 2011). The integration of fundamental measurement principles with digital technology may help ameliorate this issue. To be able to ”develop curriculum support tools for teachers to plan an explicit incorporation of FoNS categories in their teaching” (Sayers et al., 2016, p. 389), the other aspects of the validity of the tool must be considered. As a next step, we need to assess the consequential validity of the claim that the tool can usefully supplement the teacher's formative assessment of number sense.

Acknowledgements

We would like to thank everyone involved in this project for their contribution to aspects such as technical implementation, design, expert reviews, data collection and analysis.

References

Andrews, P. & Sayers, J. (2015). Identifying Opportunities for Grade One Children to Acquire Foundational Number Sense: Developing a Framework for Cross Cultural Classroom Analyses. *Early Childhood Education Journal*, *43*(4), 257–267. https://doi.org/10.1007/s10643-014-0653-6

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Black, P. & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Clarke, B., Cheeseman, J. & Clarke, D. (2006). The mathematical knowledge and understanding young children bring to school. *Mathematics Education Research Journal*, *18*(1), 78–102. https://doi.org/10.1007/BF03217430

Clarke, D., Clarke, B. & Roche, A. (2011). Building teachers' expertise in understanding, assessing and developing children's mathematical thinking: The power of task-based, one-to-one assessment interviews. *ZDM*, *43*(6), 901–913. https://doi.org/10.1007/s11858-011-0345-2

Ginsburg, H. (2016). Helping early childhood educators to understand and assess young children's mathematical minds. *ZDM*, *48*(7), 941–946. https://doi.org/10.1007/s11858-016-0807-7

Ginsburg, H. & Pappas, S. (2016). Invitation to the birthday party: Rationale and description. *ZDM*, *48*(7), 947–960. https://doi.org/10.1007/s11858-016-0818-4

Malofeeva, E., Day, J., Saco, X., Young, L. & Ciancio, D. (2004). Construction and Evaluation of a Number Sense Test With Head Start Children. *Journal of Educational Psychology*, *96*(4), 648–659. https://doi.org/10.1037/0022-0663.96.4.648

Sayers, J., Andrews, P. & Boistrup, L. B. (2016). The role of conceptual subitising in the development of foundational number sense. In W. A. Meaney T., Helenius O., Johansson M., Lange T. (Eds.), *Mathematics education in the early years* (pp. 371–394). Springer. https://doi.org/https://doi.org/10.1007/978-3-319-23935-4_21

Wolfe, E. W. & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I--instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

Wolfe, E. W. & Smith, J. E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II--validation activities. *Journal of Applied Measurement*, *8*(2), 204–234.

Wright, B. D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*, *14*(2), 97–116. http://www.jstor.org/stable/1434010