Factors that influence the difficulty level of digital arithmetic assessment items for first-grade students

Gunnhild Saksvik-Raanes[1] and Trygve Solstad[2]

Norwegian University of Science and Technology, Norway;

gunnhild.b.saksvik@ntnu.no[1], trygve.solstad@ntnu.no[2]

Digital technologies enable new possibilities for the assessment of mathematical competence. When designing an assessment, it is essential to know how different design elements affect both the item difficulty and the strategies used by the children. In this paper, we investigate digital items that were designed to measure arithmetic competence as a component of the Foundational Number Sense (FoNS) framework for five- and six-year-old children. A Rasch analysis of the performance of 302 Norwegian children showed that the type of arithmetic problem and the magnitude of the answer strongly affected an item's difficulty level. Our qualitative observations indicated that certain additional design elements of the items might have influenced both the items' difficulty and the children's solution strategies. From a mixed methods perspective, we discuss the potential of different design elements to better assess children's understanding of numbers.

Keywords: Assessment, digital technology, numbers sense, arithmetic competence, primary school.

Introduction

Digital technologies bring both constraints and affordances to assessment in mathematics education (Threlfall et al., 2007). When assessing young children's mathematical competence, using a digital medium can alleviate the effect of irrelevant demands, such as reading or writing skills. At the same time, we might add elements in the design process of a digital item that could affect the assessment in unintended ways. Carefully designing digital assessment items might enable us to improve assessments and tell us more about the children's solution processes (Saksvik-Raanes & Solstad, 2021). To realise the full potential of digital assessments, we need to know more about how children perceive the different design elements of digital items and what strategies they use to solve them.

In this paper, we investigate digital arithmetic items for five- and six-year-olds from a mixed methods perspective and pose the following research question: Which design elements influence the level of difficulty of digital arithmetic assessment items, and how do the design elements influence the strategies that first-grade children use to solve these items?

Frameworks

Arithmetic competence as a part of the FoNS model

The Foundational Number Sense (FoNS) model describes the number-related skills that require instruction (Andrews & Sayers, 2015). In FoNS, the number sense concept is defined as multi-layered, flexible and relational. The FoNS model divides the number-related skills into eight interrelated categories: number identification, systematic counting, number–quantity relationships, quantity discrimination, representing numbers, estimation, simple arithmetic competence, and awareness of number patterns. In this paper, we focus on simple arithmetic competence, which is described as a child's ability to manipulate small sets through addition or subtraction.

Item design

The arithmetic items presented in this paper were designed based on four main categories: change and combine (sum items) as well as compare and equalise (difference items) (Carpenter & Moser, 1984). Previous studies have shown that children in kindergarten are highly capable of solving such items through modelling the situations in the problem (Carpenter et al., 1993). Sum items are composed in two ways. Either with one initial quantity and an action that causes a change (join), or with two initial quantities that may be considered separately or as a part of a whole. Difference items involve comparing two quantities to determine the difference between them. Equalise problems include an additional action that is to be performed to make the two sets equal.

In addition to problem type, the difficulty of the items was expected to depend on three further design elements. Some items used pictorial representations of numbers, and other items used symbolic representations of numbers. The items involved different numerical values between one and twenty. We balanced the number of items involving small (< 10) and large (≥ 10) numbers and items having ordered and unordered response buttons (see Figure 3).

Methods

Participants and procedure

Fifteen arithmetic items were solved by 302 first-grade children who were a part of a larger study that investigated 368 children's number sense using digital assessment tools. To select participants for the project, we invited about 50 elementary schools in and around Trondheim municipality in Norway to participate in the project. Eight of the interested schools were chosen to participate and all the 1st grade children in these schools carried out the assessment. The children were five and six years old.

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 in such a manner that they would not be 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 it at any time. Pre-recorded voice instructions were given for each item. The arithmetic items appeared at the end of the full assessment. There was no time limit for the items, but the time taken for each item was recorded. The children typically spent between 15 and 25 minutes on the full assessment, of which about one-fifth comprised arithmetic items.

Qualitative data from individual interviews conducted with 19 first- grade children solving the arithmetic items, were collected independently as a part of a master's degree project (Schjølberg, 2021). The goal of the interviews included in the master's project was to get an overview of the children's strategies. One of the strategies applied by one of the students who participated in the master's project is included in this paper. The interviews were carried out at about the same time as the main data collection.

All the described studies have been approved by the Norwegian Centre for Research Data, and the necessary guidelines related to depersonalisation and parental consent have been followed.

Items

The 15 arithmetic items were designed to investigate the different aspects of the children's arithmetic competence that could influence item difficulty. Four items involved the difference between two numbers. Two of these 'difference items' were compare problems, and two were equalise problems (Carpenter & Moser, 1984). The compare items involved small numbers (< 10), while the equalise problems involved large numbers (≥ 10).

Eleven items asked for the sum of two numbers. Eight of these 'sum items' included the systematic variation of three design elements: (i) small (< 10) or large (≥ 10) answer, (ii) symbolic or pictorial representation of the problem and (iii) ordered or unordered response buttons (see Figure 3).

A priori, we expected the difference items to be more difficult than the sum items, the items with large numbers to be more difficult than those with small numbers, the items with symbolic representations of numbers to be more difficult than those with pictorial representations and the items with unordered response buttons to be more difficult than those with ordered response buttons.

Analysis

All items were scored dichotomously, meaning that the children received one point for a correct answer and zero points for a wrong answer. Rasch measurement was used for the quantitative analysis of the children's responses using the Winsteps software (Linacre, 2017). 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 units of logits (Wright, 1977). The probability that person v scores 1 point on item i depends on the difference between the skill of person v, β_v and the difficulty of item i, δ_i according to

$$\boldsymbol{P}\{\boldsymbol{X_{vi}} = \mathbf{1}|\boldsymbol{\beta_v}, \boldsymbol{\delta_i}\} = \frac{\boldsymbol{e}^{(\beta v - \delta i)}}{\mathbf{1} + \boldsymbol{e}^{(\beta v - \delta i)}}$$

Winsteps implements the joint maximum likelihood estimation (JMLE) algorithm to estimate the parameters of this model.

The excerpt from the individual interviews demonstrates how some children used the available resources on the screen to find the right answer to the problems. The qualitative data was analysed using a thematic analysis (Bryman, 2016).

Results and discussion

Task type

From Figure 1, we see that the type of task strongly influenced the difficulty of the items. As expected, the four difference items were also the four most difficult arithmetic items (Figure 1, orange markers). An independent samples t-test between the four difference items and four comparable summation items (symbolic representations involving large and small numbers and ordered and unordered response buttons) showed that this difference was significant ($p = 0.026$; $df = 6$).

Surprisingly, within the difference category, both compare items had higher difficulty than the two equalise items. The compare items were more difficult despite involving small numbers, while the equalise items involved large numbers and came with more complex voice instructions.

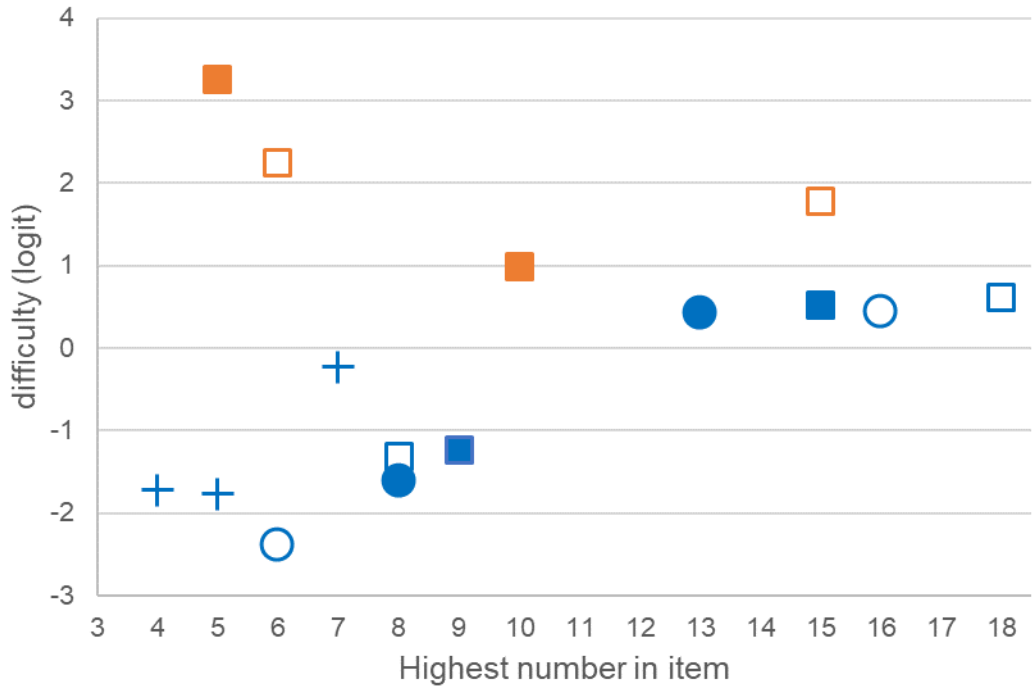

Figure 1. Item difficulty ordered by the highest number involved in the item

The items are categorised as (i) sum item (blue markers) and difference item (orange markers), (ii) pictorial representation (circular markers) and symbolic representation (square markers), and (iii) ordered response buttons (open markers), unordered response buttons (filled markers) and no response buttons (plus markers)

The compare and equalise items were visually identical (Figure 2), and three design elements differentiated them: the magnitude of the answer, the order of the response buttons and the voice instruction given. The answer was less than 10 for both compare items, while the answer was greater than 10 for both equalise items. The following voice instruction was given for the compare items: “How many more marbles are there in the blue box?”. For the equalise problems, the following voice instruction was given: “There should be an equal number of marbles in each box. How many more should the red box have?”.

In the design process, the difference items were challenging to create in a way that would enable all children to understand the given voice instructions. We wanted to keep the instructions as simple as possible to adapt to the attention span of the target group. At the same time, the compare and equalise problems represent two semantically different problems. The equalise problems involve one more step than the compare problems, as an action is performed on one of the two groups when comparing

the quantities. We therefore expected that the equalise items would be more difficult. However, Figure 1 shows that the compare items were the most difficult.

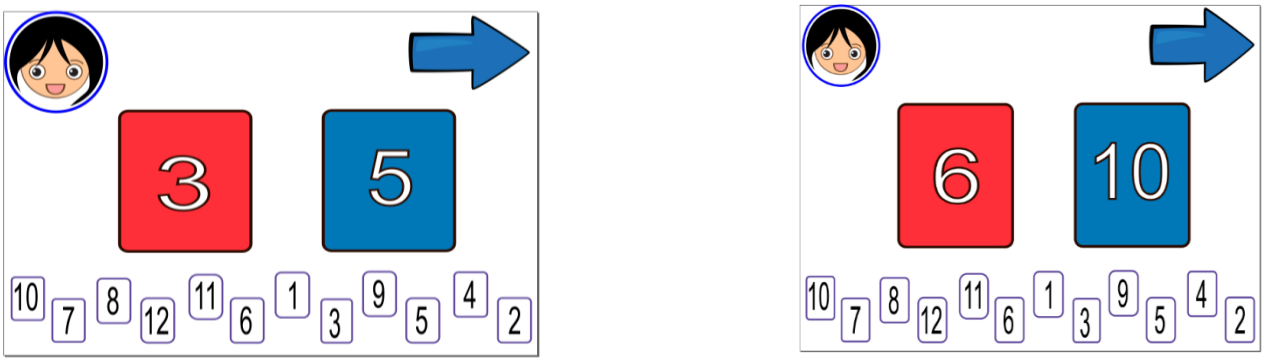

Figure 2. Two difference items

Left: item 13 (compare). Right: item 15 (equalise)

One reason why the compare items appear to be more difficult could be that the added action in the instructions for the equalise items make them more concrete, and this might have aided the children's comprehension of the items. Carpenter et al. (1993) found the kindergarteners in their study to be highly competent in solving compare problems through modelling. It is also possible that some children ignored the "more" word in the instructions for the compare tasks and interpreted it to mean "how many marbles are there in the blue box?". These results indicate that simplified instructions in word problems may lead to more misunderstandings and reduce the child's possibilities for modelling the situation.

The level of abstraction in the illustrations of these four items might also have contributed to their relatively high difficulty compared to the sum items. The children's previous experiences could also have played a role in determining the level of difficulty, as it seems that they were more familiar with the language-related problems that involved addition than subtraction.

Taken together, these results underline the importance of carefully investigating the various design elements when developing digital assessment items.

Magnitude of the answer

For the sum items, we found that difficulty was strongly correlated with the magnitude of the answer of an item. The Pearson correlation between difficulty and answer magnitude was $r = 0.88$ ($p < 0.001$) for all 11 sum items and $r = 0.96$ ($p < 0.001$) for the eight sum items that had a shared problem structure (Figure 1). In particular, the four sum items with a large answer were 2.1 logits more difficult than the corresponding four sum items with a small answer on average. An independent samples t-test showed that this difference was significant ($p < 0.001$; $df = 6$).

Number representations

A pictorial representation of a number is often thought to be easier to understand than its more abstract, symbolic representation. However, in the group of the eight sum items that shared a problem

structure, we found no significant difference in the difficulty between the four items with symbolic representations (blue squares in Figure 1) and the four corresponding items with pictorial representations (blue circles in Figure 1) ($p = 0.65$; $df = 6$; independent samples t-test). One reason for this might be that the response buttons were written in the symbolic representation. Thus, knowing the correct answer only verbally would not be sufficient to provide a correct response. Indeed, from the qualitative data, we observed that some children knew how to verbally count to 20 without recognising the corresponding written numerals (see the next section).

Order of the response buttons

Based on pilot studies, we had the a priori expectation that unordered response buttons would increase the difficulty of the items because they do not easily allow children to rely on verbal counting strategies. However, at least at first glance, the structure of the response buttons did not seem to strongly influence the difficulty of the items (Figure 1; open vs filled markers). An independent samples t-test between the four sum items with ordered response buttons and the four sum items with unordered response buttons was not significant ($p = 0.88$; $df = 6$).

On closer inspection, the four sum items with large answers were found to be of similar difficulty (Figure 1) even though the two items with ordered response buttons had larger answers than the two items with unordered response buttons. It is therefore possible that the ordered response buttons made the two tasks with the largest answers easier to solve. The latter interpretation is substantiated by qualitative analyses of the children's solution strategies. One example is item 10, which involved numbers that some children were not very familiar with. After the voice instruction "What is sixteen and two altogether?", the child was to choose the correct answer from the response buttons at the bottom of the screen. The qualitative observations gathered during the data collection led us to carry out a small qualitative interview study on the children's solution strategies.

Qualitative observations of Agnes's strategies

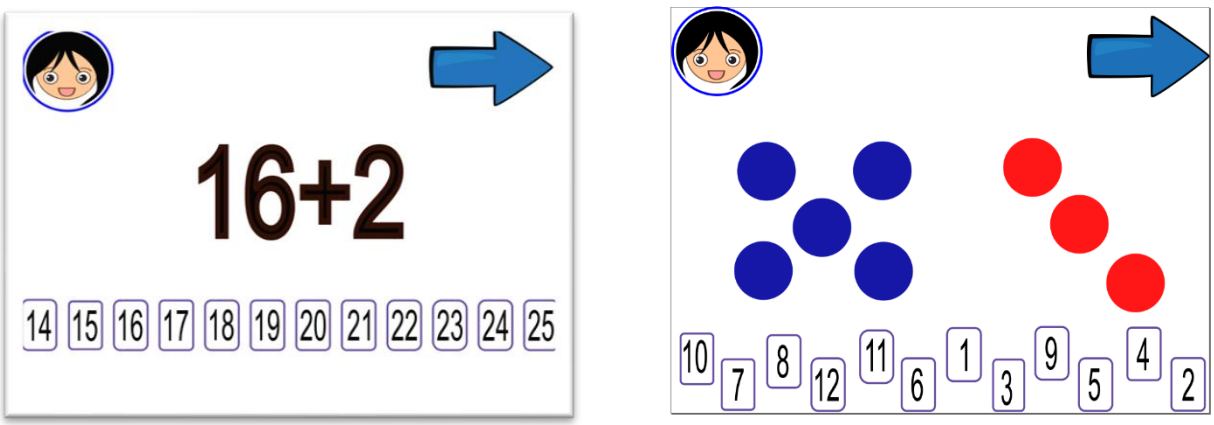

Figure 3. Sum items with systematic variation in the design elements

Left: item 10 with large answer, symbolic representations and ordered response buttons. Right: item 5 with small answer, pictorial representations, and unordered response buttons

In item 10, one of the children, Agnes, used the number alternatives on the screen to find the right answer, but she did not know what numeral she ended up with:

Researcher: What is 16 and 2 altogether?
(..)
Agnes: It is... 16 and 2…
(..)
Agnes: Wait… and then we go 1-2.
(Agnes points to 16 and makes two jumps with her finger on the numerals to the right)
(..)
Researcher: What are you thinking?
Agnes: That one.
(Points to 18)
Researcher: Do you know what number that is?
Agnes: 1-2-3-4-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18. Eighteen!
(Counting to 14 and then pointing to the numerals on the screen)

Agnes used the buttons to find the numeral that displayed the answer when she was unable to recollect the numerals after 14. In the design process, we did not expect the children to use the number alternatives in this way. These observations also emphasise the importance of investigating the available resources and how these could affect children's solution strategies.

To obtain a more fine-grained analysis of the role of ordered and unordered response numerals, we need to investigate an instrument in which items with ordered and unordered response buttons are designed with identical arithmetic problems.

Conclusions

To investigate the factors that influence the level of difficulty in digital assessment items for arithmetic, we have looked at the role of problem type, representations, numerical values and differently ordered response buttons. We have also considered how children may use the ordered response buttons to find the correct answer for an item.

The strongest determinant of item difficulty was the type of problem: difference items were more difficult than sum items, and compare items were more difficult than equalise items. The second strongest determinant of item difficulty was the numerical value of the item's answer. Whether the problem was presented in a symbolic or pictorial form did not affect item difficulty. Finally, although we could not conclusively determine the influence of ordered or unordered response buttons, our data indicates that ordered response buttons allow children who have not yet acquired mastery over large numerals to use these buttons as a number sequence that helps them solve the problem. Including both kinds of response buttons might help distinguish between the children's knowledge of large numerals and their reasoning regarding the number sequence or with a number line.

While digital technologies continue to influence the assessment of students' mathematical competence with its new possibilities, it is also important to consider the technical and methodological challenges involved in this development (Nortvedt & Buchholtz, 2018). There are many aspects to consider when investigating the various elements that affect children's solution processes when interacting with digital technologies. To ensure the validity of such assessments, it is

important that future research looks more into how the various possibilities that digital technologies enable can both improve and hinder students' performance. One way forward could be to compare items that have different design elements but similar answers. Looking more directly at a greater variety of both word problems and symbolic items can allow us to determine how the different contents affects the difficulty of the items. With the use of digital technology, we could also look more closely into young children's competence for solving digital word problems. For instance, one could record the children's solution process while they are introduced to a variety of word problems with more elaborate instructions and pictorial representations. The use of different digital aids, with a larger degree of interactivity, could enable us to study in more detail how the children model and use different strategies to solve the problems.

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

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Carpenter, T. P., Ansell, E., Franke, M. L., Fennema, E., & Weisbeck, L. (1993). Models of problem solving: A study of kindergarten children's problem-solving processes. *Journal for Research in Mathematics Education*, *24*(5), 428–441. https://doi.org/10.5951/jresematheduc.24.5.0428

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15*(3), 179–202. https://doi.org/10.2307/748348

Dewey, J., Alexander, T. M., & Hickman, L. A. (1998). *The essential Dewey: 1: Pragmatism, education, democracy* (Vol. 1). Indiana University Press.

Linacre, J. M. (2017). *Winsteps® Rasch measurement computer program. User's Guide.* Beaverton, OR: https://www.winsteps.com/

Nortvedt, G. A., & Buchholtz, N. (2018). Assessment in mathematics education: Responding to issues regarding methodology, policy, and equity. *ZDM – Mathematics Education*, *50*(4), 555–570. https://doi.org/10.1007/s11858-018-0963-z

Saksvik-Raanes, G., & Solstad, T. (2022). Developing a formative, teacher-oriented, digital tool to assess number sense in school starters. *NORMA20 Proceedings*, Oslo Norway.

Schjølberg, A. (2021). *"Jeg må prøve å telle" En kvalitativ studie om elevers strategier i arbeid med tallforståelsesoppgaver på første trinn* (p. 87). [Master thesis] Norwegian University of Science and Technology.

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, *66*(3), 335. https://doi.org/10.1007/s10649-006-9078-5

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