CSJE1705901 VOL 0, ISS 0

Validation of an Early Numeracy Screener for First Graders

Anita Lopez-Pedersen, Riikka Mononen, Johan Korhonen, Pirjo Aunio, and Monica Melby-Lervåg

QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left are hyperlinked to the location of the query in your paper.

The title and author names are listed on this sheet as they will be published, both on your paper and on the Table of Contents. Please review and ensure the information is correct and advise us if any changes need to be made. In addition, please review your paper as a whole for typographical and essential corrections.

Your PDF proof has been enabled so that you can comment on the proof directly using Adobe Acrobat. For further information on marking corrections using Acrobat, please visit http://journalauthors.tandf.co.uk/production/acrobat.asp; https://authorservices.taylorandfrancis.com/how-to-correct-proofs-with-adobe/

The CrossRef database (www.crossref.org/) has been used to validate the references. Changes resulting from mismatches are tracked in red font.

AUTHOR QUERIES

QUERY NO.	QUERY DETAILS
Q1	Please note that the ORCID has/have been created from information provided through CATS. Please correct if this is inaccurate.
Q2	Please provide the missing Finland address for d affiliation.
Q3	Please provide the missing department for b and c affiliations.
Q4	Please provide the missing physical address for corresponding author.
Q5	The sentence “Teacher instructions in italics” has been changed to “Teacher instructions are displayed”. Please correct if this is inaccurate.
Q6	“Hannola-Sormunen” has been changed to “Hannula-Sormunen” in the citation. Please correct if this is inaccurate.
Q7	Please check that the heading levels have been correctly formatted throughout.
Q8	The reference "NCME, 2018" is cited in the text but is not listed in the references list. Please either delete the in- text citation or provide full reference details following journal style.
Q9	The spelling of “Høstmælingen” in “Evers, Hagemeister, & Høstmælingen, 2013” has been changed to match the entry in the references list, but is inconsistent with the following reference. Please provide revisions if this is incorrect.
Q10	In point (4) in the list “MLD” is defined as “mathematical learning disabilities”. Earlier in the paper it was defined as “mathematical learning difficulties”. Please indicate which is correct.

QUERY NO.	QUERY DETAILS
Q11	The year for "Hu & Bentler, 1999" has been changed to year to match the entry in the references list. Please provide revisions if this is incorrect.
Q12	In the sentence beginning “In addition, in some of the tasks …” “(i.e., one task that was omitted one point for a correct answer required the children to solve, for example, two items within the task correctly…)” has been changed to “(i.e., one task that omitted one point for a correct answer required the children to solve, for example, two items within the task correctly…)”. Please check the change conveys the intended meaning or amend.
Q13	The sentence beginning “Regarding how the screener predicts skills later on, in a general level, not solemnly children at risk of learning disabilities…” has been changed to “Regarding how the screener predicts skills later on, in a general level, not solely children at risk of learning disabilities…”. Please check the change conveys the intended meaning or amend.
Q14	The disclosure statement has been inserted. Please correct if this is inaccurate.
Q15	The CrossRef database (www.crossref.org/) has been used to validate the references. Mismatches between the original manuscript and CrossRef are tracked in red font. Please provide a revision if the change is incorrect. Do not comment on correct changes
Q16	The reference "Chan 2014" is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
Q17	Please provide missing city for the reference "Dockrell et al. 2017" references list entry.
Q18	The reference "National Council of Teachers of Mathematics 2006" is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
Q19	The reference "National Council on Measurement in Education 1999" is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
Q20	The reference "National Mathematics Advisory Panel 2008" is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
Q21	The reference "Terwee et al. 2017" is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
Q22	Please provide missing city for the reference "Zumbo and Chan 2009" references list entry.
Q23	Please supply a title for Appendix A.
Q24	Please supply a caption for Table A1.
Q25	The sentence “Teacher instructions in italics” has been changed to “Teacher instructions are displayed”. Please correct if this is inaccurate.
Q26	Please supply a title for Appendix C.

SCANDINAVIAN JOURNAL OF EDUCATIONAL RESEARCH
https://doi.org/10.1080/00313831.2019.1705901

Validation of an Early Numeracy Screener for First Graders

Q1 Anita Lopez-Pedersen[a], Riikka Mononen[a], Johan Korhonen [b], Pirjo Aunio[c,d] and Monica Melby-Lervåg[a]

Q3 Q2 [a]Department of Special Needs Education, University of Oslo, Oslo, Norway; [b]Åbo Akademi University, Turku, Finland; [c]Helsinki University, Helsinki, Finland; [d]Finland and Department of Special Needs Education, University of Oslo, Oslo, Norway

ABSTRACT

This study investigated the psychometric properties of the Early Numeracy Screener. The Early Numeracy Screener is a teacher-administered, paper-and-pencil test measuring counting skills, numerical relational skills, and basic arithmetic skills. Three hundred and sixty-six first graders took the Early Numeracy Screener at the beginning of the school year. Confirmatory factor analysis was conducted in order to examine whether the screening tool was identified as a one-factor model or a three-factor model. The confirmatory factor analysis found evidence for the three-factor model, establishing construct validity. Evidence for criterion-related validity was found in crosstabulation and correlation with the national test measuring overall mathematics performance taken towards the end of the school year. The Early Numeracy Screener may serve as an indicator of young children's performance in early numeracy. The brevity and ease of use of the Early Numeracy Screener make it suitable for classroom instructional settings.

ARTICLE HISTORY
Received 5 June 2019
Accepted 12 October 2019

KEYWORDS
Early numeracy; reliability; screener; test; validity

Coll:
QA:

Q5 Introduction

Developing well-functioning early numeracy skills is a foundation for further mathematical skills and for qualification for employment in society (Duncan et al., 2007; Geary, Bailey, & Hoard, 2009; Geary, Hoard, Nugent, Bailey, & Krueger, 2013; Ritchie & Bates, 2013; Trilling & Fadel, 2009). Mathematical skills develop in a cumulative fashion, with early skills forming a foundation for the acquisition of later skills (Aunio & Niemivirta, 2010; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Purpura, Baroody, & Lonigan, 2013). The early childhood years serve perhaps the most important developmental years in one's life (McGuire, Kinzie, & Berch, 2012). Longitudinal studies show that early numeracy skills are important for the kind of learning trajectory a child has in primary school mathematics (Aunio & Niemivirta, 2010; Desoete, Stock, Schepens, Baeyens, & Rieyers, 2009; Dougherty, 2003; Gersten et al., 2015; Jordan, Glutting, & Ramineni, 2010; Jordan et al., 2009; Krajewski & Schneider, 2009; Zhang et al., 2014). Studies have related children's mathematical achievement to specific aspects (e.g., counting skills) of their early numerical competencies (Hannula-Sormunen, Lehtinen, & Räsänen, 2015). Differences in early numeracy are displayed before the onset of formal schooling (Berch, 2005). Children who perform poorest in early numeracy skills may have serious deficits in all early number skills (Salminen, Koponen, & Tolvanen, 2018). For instance, verbal counting plays an important role as a predictor of arithmetic (Zhang et al., 2014); accordingly, an important skill for identifying children at risk of developing mathematical learning disabilities might be counting. Therefore, to assist children in establishing these skills, it is important

Q4 CONTACT Monica Melby-Lervåg monica.melby-lervag@isp.uio.no

to identify children who struggle with numeracy skills at an early stage. Despite this need, there are surprisingly few well-validated screening tools for this purpose for non-English-speaking European countries. To ensure the validity of assessment tools to be used in different countries, even in mathematics, they preferably need to be validated for each language and country. To fill this gap, we present a study of the development and validation of an early numeracy screening tool in Norwegian to detect first graders with challenges in early numeracy skills.

Q7 Why Assessment of Early Numeracy Skills?

Mathematical skills is a continuous variable that is normally distributed in the population, and cut-offs to establish normal versus disordered development will be arbitrary. However, it is common to assume that around 15–20% of children and adults experience difficulties in developing mathematical skills to such an extent that it hampers their school or work performance (Geary, 2011). Out of these, 5–7% have problems so severe that they are often diagnosed as having specific mathematical learning disabilities or dyscalculia (American Psychiatric Association, 2013; Butterworth, Varma, & Laurillard, 2011). Identifying and remediating the early numeracy that predicts poor school-entry mathematical knowledge has the potential to reduce these risks substantially, and accordingly considerable resources have been devoted to these efforts in recent years (Clarke et al., 2016; Fuchs et al., 2013; Gersten et al., 2015; Jordan, Glutting, Dyson, Hassinger-Das, & Irwin, 2012).

Core Numerical Skills Model: What do We Need to Assess?

When it comes to developing appropriate targeted assessment tools for early numeracy skills, it has been suggested that the skills that need to be considered generally fall into three different domains: understanding numerical relations, counting skills, and basic arithmetic skills (Aunio & Räsänen, 2016; Jordan et al., 2009; National Research Council [NRC], 2009; Purpura & Lonigan, 2013). Aunio and Räsänen (2016) theorized a model of these crucial numerical factors for the development of mathematical skills among children aged 5–8 years old. Their model was based on the results of longitudinal studies and a series of analyses of standardized tests intended to measure the development of mathematical skills. Support for the content of these domains can also be found in previous research describing early numeracy (Desoete, Ceulemans, De Weerdt, & Pieters, 2012; Gersten et al., 2012; Moeller, Pixner, Zuber, Kaufmann, & Nuerk, 2011).

Understanding Numerical Relations

If we examine more closely the three putative domains that constitute early numeracy skills, the first foundation is to understand numerical relations. Numerical relations refer to the understanding of the quantitative and non-quantitative relationships between the elements in the task (Aunio & Räsänen, 2016). Numerical relational skills serve as a prerequisite to basic arithmetic skills. Knowledge of basic arithmetic principles is often referred to as the understanding of part-whole relations in addition or subtraction tasks (Canobi, Reeve, & Pattison, 2002; Wilkins, Baroody, & Tiilikainen, 2001). Numerical relational skills include a set of subskills such as the ability to compare the magnitudes of numbers, to understand cardinal value, one-to-one correspondence and early mathematical-logical principles, and to understand the meaning of the 10-base system (Aunio & Räsänen, 2016; Geary & vanMarle, 2016). Longitudinal studies clearly establish that numerical relational skills are a crucial part of early numeracy development (Aunio & Niemivirta, 2010; Desoete et al., 2009; Stock, Desoete, & Roeyers, 2009). Research has pointed out that children's understanding of numerical magnitudes predicts individual differences in mathematics achievement (e.g., De Smedt, Noël, Gilmore, & Ansari, 2013; De Smedt, Verschaffel, & Ghesquière, 2009; Halberda, Mazzocco, & Feigenson, 2008; Holloway & Ansari, 2009; Schneider et al., 2016; Vanbinst, Ghesquière, & De Smedt, 2015).

Counting Skills

The second component in early numeracy skills is counting skills. Counting skills refers to the child's knowledge of number symbols, skills in moving within the sequence of the number words, and enumeration (Aunio & Räsänen, 2016). Counting strategies are, perhaps not surprisingly, an imperative aspect of children's early numerical knowledge (Aunio & Räsänen, 2016; Wright, Martland, & Stafford, 2006). Counting reinforces the child's understanding of the relationships between numbers (Baroody, 2003, 2006; Baroody, Eiland, & Thompson, 2009). Counting also helps expand children's quantitative knowledge to larger numbers (Baroody, 2003, 2006; Baroody et al., 2009). Counting knowledge allows children to count on or up from addends to solve novel number combinations—a key arithmetic strategy in early elementary school (Geary, Hoard, Byrd-Craven, & DeSoto, 2004).

Basic Arithmetic Skills

The understanding of numerical relations and counting is a prerequisite for the core component in early mathematical skills, namely basic arithmetic. Basic arithmetic skills in 5–8-year-olds pertain to the degree to which a child masters mainly the addition and subtraction tasks with number symbols (Aunio & Räsänen, 2016). Basic arithmetic skills also depend on adequate counting skills. Frequent and successful use of counting strategies usually leads to improvements in memory representations of arithmetical facts and leads to the strategy of retrieving arithmetical facts from long-term memory (Canobi et al., 2002; Wilkins et al., 2001). Correct and fluent number word sequences, part of counting skills, are also relevant for solving basic arithmetic addition and subtraction tasks since children use counting-based strategies at the beginning when learning arithmetic, for example, number word sequences advancing forward to solve addition problems and backward when solving subtraction problems.

Quality of Educational Assessment Instruments—Validation of an Assessment Tool

A range of different systems and criteria for judging the validity of a measurement exist (AERA et al.,1999; APA et al., 1974; Cosmin, NCME, 2018). Validity is a crucial psychometric notion since it concerns the degree to which the test scores provide information that is related to the conclusions drawn from them. Validity is an evaluation of the degree to which empirical evidence and theoretical rationales support the adequacy and suitability of interpretations resulting from test scores of other models of assessment (Zumbo & Chan, 2009).

The European Federation of Psychologists' Associations (EFPA) (Evers, Hagemeister, & Hostmealingen, 2013; Evers, Muñiz, Høstmælingen, Sjöberg, & Bartram, 2013) has provided a description and a thorough assessment of the psychological assessment tests, namely "EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests". EFPA provides a manual for examining the quality of the test materials per se as well as psychometric properties of the identified documented instruments. In this study, validity and reliability are evaluated through EFPA's quality criteria.

EFPA emphasizes two types of validity—construct validity and criterion validity. Construct validity refers to whether the items represent the theoretical constructs that they are designed for. Construct validity, according to EFPA, is whether the test actually measures the intended construct or something else. Criterion-related validity is required for all kinds of tests and demonstrates the extent to which test instruments correlate with relevant valid instruments used for the same purpose to predict whether future or current performance is related to another measure of the same construct. In addition, criterion-related validity includes predictive validity, that is, whether the test is applicable in the sense that it serves its purpose, in this case to identify children at risk of developing mathematical learning difficulties (MLD). The types of validity mentioned here cannot be evaluated

in isolation, they are complementary. Overall validity refers to the appropriateness, meaningfulness, and usefulness of inferences based on scores (Cronbach & Meehl, 1955; Messick, 1995).

In addition, it is also important to assess the extent to which a measure is prone to random measurement error. Reliability refers to the extent to which an instrument produces random measurement errors and is expressed either as a standard error of measurement or as a reliability coefficient. Reliability is crucial if an assessment is to be useful, as a test that is not reliable can never be a valid instrument (Thorndike & Thorndike-Christ, 2014).

Existing Assessments and Screening Instruments of Early Numeracy

Most European countries have mandatory math assessments or screening such as national tests. National assessments are designed for various purposes, not necessarily targeted to identify those who are at risk of developing MLD, but at the very least to identify children in need of extra support. Still, these tests are often introduced later in children's mathematical development and are not suitable for detecting those who are struggling at an early stage, more or less before formal math teaching starts. As for instruments other than the mandatory assessments, a review of early numeracy assessments shows that there is a scarcity of well-validated tests and screening tools (Dockrell et al., 2017). Dockrell et al. (2017) identified 23 omnibus tests and 16 tests assessing number concepts and skills, published in English. Dockrell et al. (2017) reviewed the extent to which the different instruments covered critical domains in numeracy, namely counting, transcoding (i.e., knowledge of number sequence, reading numerals, writing numerals, matching numerals to numbers), comparing numerical magnitudes, and simple arithmetic.

In summary, only 4 of the 23 tests of numeracy featured items in all 4 areas that Dockrell et al. (2017) emphasized as critical domains; none of them included all types of item and none was well-validated. The tests included in Dockrell et al.'s (2017) review were both screeners and numeracy subtests as part of a larger battery of mathematical tests (e.g., Keymath-3 [Connely, 2007]). The review concludes that choosing a test that does not cover all areas of numeracy skills can lead to misidentification of children with difficulties as well as challenges in planning the content of support.

To ensure validity of assessments tools in mathematics to be used in different countries, they preferably need to be validated for each language and country. A challenge when validating tests in continental Europe is that many different languages are used, demanding considerable validation resources. The measures included in the review by Dockrell et al. (2017) were in English, but there are also validated screening instruments in other European languages. One widespread tool is the Dutch Early Numeracy Test (Van Luit, Van de Rijt, & Pennings, 1994; Van de Rijt, Van Luit, & Pennings, 1999) and the translation and norming of this into Finnish (Aunio, Hautamäki, Heiskari, & van Luit, 2006), German (Van Luit, Van de Rijt, & Hasemann, 2001). Brankaer, Ghesquière, and De Smedt (2017) validated the SYMP test, a magnitude processing screening, and had a satisfactory test–retest reliability as well as construct- and criterion-related validity.

In Norway, few studies exist on validation of screening instruments. The most commonly used tool for early numeracy is the summative national assessment test. Another commonly used instrument is Alle Teller! (McIntosh, 2012), but this test along with others is neither validated nor normed, hence leaving the teacher few options in assessing numeracy in early schooling. This does not necessarily imply that these assessment tools are of poor quality, but if they are not evaluated, for instance, relative to EFPA's criteria, they provide little or no information when it comes to the quality of the assessment tool. Thus, Norwegian schools and teachers are in an unfortunate situation where they are unable to consider the quality of the assessment tools.

The Current Study: Aims and Research Questions

The Early Numeracy Screener has a research-based theoretical foundation with three core components (Aunio & Räsänen, 2016), and can thus be used to improve the efficacy of assessment

as well as help teachers structure children's need of support more comprehensively in relation to the three core factors. The aim of the present study was to evaluate the psychometric properties of the Early Numeracy Screener for Norwegian first graders. More specifically we investigated:

(1) item-level difficulty and the internal consistency of the measure and its subscales (reliability);
(2) whether the hypothesized three-factor structure of the measure fits the data best (construct validity);
(3) whether the measure showed measurement invariance across gender and age (known group validity);
(4) how the measure is related to the national tests in mathematics and whether the measure identifies children who are at risk of developing mathematical learning disabilities (MLD) (compared with the national tests) (criterion-related validity).

Q1

Altogether, this provides insights into the Early Numeracy Screener as an evaluator of construct validity, criterion-related validity, and reliability. Additionally, information about invariance will ensure whether the test has known groups validity, that is, the extent to which a measurement is sensitive to differences and similarities in various groups such as gender, age groups, and so on. Thus, it is important to test invariance to ensure that the test works equally for boys and girls, and for children born at different times of the year (younger and older children within Grade 1). If successfully validated, the Early Numeracy Screener would meet the need for a valid instrument that targets early numeracy skills for first graders that teachers would be able to use easily and efficiently. Our study will then contribute to educational practice in several ways. This study enables Norwegian schools to use flexibly a validated screening tool and a screening tool that targets early numeracy skills separately and not unidimensionally. This enables targeted support for children at risk of developing learning difficulties in mathematics.

Method

Participants

All children born in 2010 and attending first grade in two municipalities in Norway were invited to participate in the study. This is not a random sample, as we first contacted heads of the two municipal affairs, and from those municipalities each school's principal agreed to participate. Teachers in each class distributed and collected information about the study and a letter of consent to the parents. This resulted in a sample size of 366 participants (mean age 6.36 years, 55.7% boys). None of the participants had been diagnosed with learning disabilities as this is often diagnosed later in school. Ethical approval was obtained from the Norwegian Social Science Data Services, and informed parental consent was obtained for each child in this study. A survey about the educational level of the parents and home language was distributed. The students were predominantly of Norwegian nationality and13.9% of the students had Norwegian as a second language (Table 1). The children were recruited from a district that was close to the national average on variables related to parental education (Norway: 25.8% secondary school, 37.2% high school, 27% bachelor degree, and 10% master's degree, recruitment area 26.9% secondary school, 37.9% high school, bachelor degree 26.1%, and master's degree 10.1% [Statistics Norway, 2019]; our sample secondary school 4.2%, high school 31.5%, bachelor degree 38.9%, master's degree 20.1 5, PhD 0.8%). Notably, 18.8% did not reply to our survey, and previous studies indicate that those with low educational levels are less likely to answer these kinds of surveys (Goyder, Warriner, & Miller, 2002). However, notably, the educational level is close to the national average with high school. Thus, since we had a cohort of 81.2% of children it could be likely that it is approaching the national average, even if our survey does not indicate so.

Table 1. Background Information: Home Language, Parental Educational Background.

Home language	%	Educational background	%
Norwegian	86.10		
English	2.70		
Lithuanian	1.30	Secondary school	4.20
Urdu	1.30	High school	31.50
Somali	1.00	Bachelor	39.20
Tamil	1.00	Master	20.10
Bosnian	0.06	PhD	4.20
Dutch	0.06	*Missing*	0.08
Kurdish	0.06		
Polish	0.06		
Sign language	0.06		
Arabic	0.03		
Burmese	0.03		
Finnish	0.03		
Icelandic	0.03		
Portuguese	0.03		
Romanian	0.03		
Russian	0.03		
Swedish	0.03		
Missing	0.03		

Notably, in Norway children enter elementary school at the age of six and most of them have been to kindergarten from the age of one until they begin first grade. In a Norwegian context, this does not include a formalized kindergarten education program, but basically day care. There is no detailed common curriculum in mathematics, so the ways of introducing children to early numeracy may differ from kindergarten to kindergarten. Moreover, the main focus in Norwegian kindergartens is on play and play-based learning rather than formal instruction.

Measures

Early Numeracy Screener

The screening measure was developed based on Aunio and Räsänen's (2016) core numerical skills model. The tasks are representative of the range of skills assessed by other early numeracy measures (Clements, Sarama, & Liu, 2008; Jordan et al., 2009). The screener consists of 52 items measuring core early numeracy skills—numerical relational skills, counting skills, and basic skills in arithmetic (addition and subtraction). One point was given per correct answer and zero per wrong answer. The three components in the screener were as follows:

Numerical relational skills were assessed using 14 tasks. The children were asked to do tasks such as comparing numbers, for example, "tick the box with the smallest number" and comparing concepts such as "one more than", "as many as", "one less than" (see Appendix B for item examples). One point was given per correct answer and zero per wrong answer.

Counting skills were measured with 28 tasks. The children were asked to do various counting tasks with ordinal numbers in number sequences, for example, "tick the third triangle", "tick the seventh star", and tasks measuring the number–quantity correspondence (see Appendix B). One point was given per correct answer and zero per wrong answer.

Arithmetic skills were assessed with 10 tasks, 6 of them in addition and 4 in subtraction. The sums and differences in the addition and subtraction tasks ranged from 0 to 15, that is, the answers were in this number range (see Appendix B). One point was given per correct answer and zero per wrong answer.

National Test in Mathematics

The national test is a curriculum-based standardized summative assessment in mathematics for first graders and is used in Norwegian schools in April of each year (Udir, 2016). It consists of 50 different

early numeracy-related items (Cronbach's α = .903, from our sample); measuring number line, counting, number word knowledge, and addition and subtraction fluency (Udir, 2016). The tasks in the national test also measure fluency. All the items are timed; they have, for instance, 1.5 min to solve a set of counting tasks. Item examples for counting skills and basic arithmetic are shown in Appendix C (Udir, 2016). Scoring of the national test varies somewhat throughout the test in the sense that some tasks are scored by giving one point for correct answers and zero for a wrong answer. However, some of the tasks require a correct sum score on two or three different tasks to obtain a score of one. Consequently, in those 50 tasks there are tasks with several items all of which the children have to get correct to score a correct answer, which is given one point.

Procedure

Data for this validation study were collected in October 2016. The teachers administering the screener were given a 3-hr-long training session in their respective schools. All the teachers administering the screener had prior experience with group-based assessment before this training because the screener is administered in the same way as the national test in mathematics. The screener is a group-based, teacher-instructed, paper-and-pencil test that requires about 30–45 min to complete. The children were given instructions before each task and were told to write down the answers themselves. Teachers collected the answer sheets and sent them to the research group for correction and coding. The research group was thus involved with correcting both the Early Numeracy Screener and the national tests in an attempt to ensure that both tests were correctly scored.

Analysis

All analyses were performed in *Mplus* 7.3 (Muthén & Muthén, 1998–2017). Owing to categorical data, weighted least squares means and variance estimation (WLSMV) was used as estimator in all the analyses. Confirmatory factor analysis (CFA) was used to examine the factor structure of the test instrument. To evaluate the fit of the structural equation models that contained latent variables, we considered the common guidelines for model fit. These guidelines suggest an acceptable fit to the data if the Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI) exceed .95, the Root Mean Square Error of Approximation is less than .08, and the Standardized Root Mean Square Residual (SRMR) is less than .10 (Hu & Bentler, 1991; Marsh, Hau, & Grayson, 2005).

Q1

Multigroup CFA with categorical data was conducted to test for measurement invariance across gender and age according to the guidelines of Muthén (Muthén & Muthén, 1998–2012, p. 485). We used a model which assumes the same factor structure but allows factor loadings and item thresholds to vary freely across groups as a baseline model (configural). The configural model was then compared with a model where factor loadings and item thresholds are constrained to equality across groups (scalar invariance). When comparing nested models, a change of more than .01 in CFI and .015 in RMSEA indicates significant differences between the models (Chen, 2007).

To investigate whether the Early Numeracy Screener identifies the same children at risk of developing MLD as does the national test, configural frequency analysis was conducted. In this analysis, the children are grouped according to their performance in the Early Numeracy Screener into at-risk and typical performing groups, and the same categorization is done with the national test scores. By means of configural frequency analysis we can then compare the extent to which the same children are identified with both tests. More specifically, the observed frequencies are compared with expected frequencies in crosstabulation and analyzed to ascertain whether cell frequencies are larger or smaller than could be expected with a base model. The base model selected for frequency comparison was the first-order CFA, which assumes that all variables under study may show the main effects and are independent of each other (Von Eye, 1990; Von Eye, Spiel, & Wood, 1996). Thus, we focus on whether children identified as at risk (or typical) with the Early Numeracy Screener are also

identified as at risk (or typical) with the national test (i.e., individual stability) as well as whether there are changes across the groups that cannot be explained by chance fluctuations (i.e., individual change).

Results

Table 2 shows the range, means, standard deviations, skewness, kurtosis, and reliabilities for all measures, and Table 3 shows correlations between them. Internal consistency was satisfactory. The same was the case for the distribution of the variables. Appendix A shows the skills measured on item level.

Item-Level Difficulty and Reliability

First, we examined the extent to which the different items discriminated between the children. We used the 95% pass or fail as a criterion in order to secure variance, since we wanted to capture individual differences with the screener, hence the items cannot be so easy that more than 95% solve them either correctly or incorrectly. Four of the items were removed because they were passed or failed by more than 95% of the children and thus were not useful in discriminating between them (Table A1). Notably, all the description and results are thus based on the screener consisting of 52, not the original 56, items. In addition, as shown in Table 2, the internal consistencies of all variables were satisfactory.

Construct Validity and Measurement Invariance Across Gender and Age for the One-Factor Model

Second, we examined the measurement model for the early numeracy construct as a one-factor model with all the items as indicators for a single overall early numeracy factor. The fit of the measurement model was χ^2 (1,274, $N = 366$) = 2,956.486, $p < .001$, RMSEA = .060 (90% CI = .057–.063), CFI = .911, TLI = .908; thus, this model fitted the data well. The multigroup CFAs indicated that the one-factor model showed measurement invariance across gender and age (see Table 4) and, more importantly, the model fit did not worsen when constraining the factor loadings and item threshold to equality across gender, ΔCFI = .002, ΔRMSEA = .000.

Table 2. Minima, Maxima, Means, SD, Skewness, Kurtosis, and Reliability for One-Factor Model, Three-Factor Model and National Test.

	Min	Max	Mean	SD	Skewness	Kurtosis	Cronbach's α
Age	5.87	6.88	6.38	0.30	0.087	−1.239	
One-factor model, full scale	2	52	30.85	11.58	−0.108	−0.764	.943
Three-factor model							
Counting skills	2	28	18.31	6.60	−0.406	−0.704	.904
Numerical relational skills	0	14	7.95	3.21	0.385	−0.692	.792
Arithmetic skills	0	10	4.58	3.19	0.803	−0.935	.901
National test	6	50	43.55	7.29	−1.467	4.459	.903

Table 3. Correlations of the Three-Factor Model and the National Test.

	Counting skills	Numerical relations	Arithmetic	National test
Counting skills	1			
Numerical relations	.744**	1		
Arithmetic	.785**	.712**	1	
National test	.415**	.395**	.300**	1

Note: **Correlation is significant at the .01 level.

Table 4. Model Fit Statistics for the Test of Invariance Across Gender and Age—One-Factor Model.

Model	$\chi 2$ (*df*)	CFI	RMSEA (90% CI)	TLI	ΔCFI	ΔRMSEA
Gender						
Configural invariance	2,548	.934	.051 (.047–.054)	.931		
Scalar invariance	2,598	.932	.051 (.047–.056)	.931	.000	.000
Age						
Configural invariance	2,548	.913	.052 (.049–.056)	.909		
Metric invariance						
Scalar invariance	2,598	.914	.052 (.048–.055)	.912	−.001	.000

Construct Validity and Measurement Invariance Across Gender and Age for the Three-Factor Model

Next, we compared a three-factor model consisting of counting skills, numeric relational skills, and arithmetic skills factors with the one-factor model. The three-factor model fitted the data better than the one-factor model, ΔCFI = .027, ΔRMSEA = .01, χ^2 (1,271, N = 366) = 2,439.967, p = .00, RMSEA = .050 (90% CI = .047–.053), CFI = .938, TLI = .934. The correlation between the factors ranged from .70 to .77. These results indicate that early numeracy can be treated as a multidimensional construct and that the Early Numeracy Screener differentiates between the three subskills. Table 5 shows factor loadings, thresholds, and variance in the three-factor model. As for the measurement invariance, the three-factor model was found to be invariant across gender and age (Table 6).

Criterion Validity

First, we examined the correlations between the three early numeracy factors and the national test in mathematics taken 6 months after the early numeracy test. The fit of this model was good, χ^2 (1,320, N = 366) = 2,474.714, p = .00, RMSEA = .049 (90% CI = .046–.052), CFI = .939, TLI = .936. The correlations between the national test scores and the early numeracy factors were low, r = .25, p < .001 (counting skills); r = .20, p < .001 (relational skills); r = .19, p < .001 (basic arithmetic skills).

Next, we grouped the children into at risk of MLD and typical-performing groups based on the Early Numeracy and the national test scores using the 20th percentile established as a critical limit in the national test as a cut-off. This resulted in four different configurations (typical/typical; typical/at-risk; at-risk/typical; at-risk/at-risk), for which the observed frequencies were compared with the expected frequencies (Table 7). To account for the increased risk of Type-I error, Bonferroni correction was applied to the significance testing (0.05/4 = .013) when comparing the observed and expected frequencies.

The configural frequency analysis identified one stable configuration and two antitypes of change. Of the 67 children who were identified as at risk with the Early Numeracy Screener, 34 (51%) were also identified as at risk with the national test. This configuration was occurring more than expected by chance, indicating stability in classification of at risk for MLD. This result was further supported by the fact that moving from at-risk status to the typical group occurred less frequently than expected by chance (antitype of change) (N = 33, 9.2%). However, the results for the typically performing group were mixed as movement from the typical group (Early Numeracy) to the at-risk group (national test) was occurring less frequently than expected by chance (N = 39, 10.9%); while at the same time the stability of belonging to the typical group in both tests was not significant (N = 253, 70.4%). Although, it is important to note that the p-value (p = .024) for the stable configuration was very close to the Bonferonni-adjusted critical value of .013, indicating that it is rather likely to be identified as typically performing in both tests. Overall, these results indicate that despite the rather

Table 5. Factor Loadings, Thresholds and Variance in the Three-Factor Model.

Item	Factor 1 Relational	Factor 2 Counting	Factor 3 Arithmetic	Item threshold	Residual variance	Explained variance (R^2)
1						
2	.703			−0.705	.506	.494
3	.596			−1.429	.645	.355
4	.701			−1.468	.509	.491
5		.552		−0.914	.695	.305
6		.596		−1.174	.645	.355
7		.528		−0.679	.722	.278
8		.441		−1.001	.805	.195
9		.539		0.048	.709	.291
10		.473		0.110	.776	.224
11	.670			−1.392	.551	.449
12	.761			−0.041	.422	.578
13	.793			0.200	.372	.628
14	.728			0.263	.470	.530
15	.543			−0.461	.705	.295
16	.551			0.796	.696	.304
17						
18	.773			−0.014	.403	.597
19	.668			0.089	.553	.447
20	.604			0.158	.635	.365
21	.627			0.228	.607	.393
22	.355			0.313	.874	.126
23		.486		−0.768	.764	.236
24		.666		−0.645	.556	.444
25		.674		−0.103	.546	.454
26		.741		−0.873	.451	.549
27		.781		−0.571	.390	.610
28		.751		0.014	.436	.564
29		.667		−0.492	.556	.444
30		.625		−0.172	.609	.391
31		.671		0.179	.550	.450
32		.751		−0.207	.435	.565
33		.862		0.357	.256	.744
34		.845		0.371	.285	.715
35		.733		−1.001	.463	.537
36		.721		−0.645	.480	.520
37		.678		−0.371	.540	.460
38		.765		−1.059	.415	.585
39		.785		−0.555	.384	.616
40		.726		0.027	.473	.527
41		.703		−1.160	.506	.494
42		.584		−0.978	.659	.341
43		.818		−0.285	.332	.668
44		.815		0.221	.336	.664
45			.895	−0.603	.198	.802
46			.891	−0.466	.205	.795
47			.845	−0.342	.285	.715
48			.974	−0.563	.050	.950
49			.870	−0.130	.244	.756
50			.835	0.379	.303	.697
51			.924	0.560	.146	.854
52			.957	0.698	.084	.916
53			.966	0.813	.067	.933
54						
55			.993	0.861	.015	.985
56						

Note: Item numbers 1, 17, 54, and 56 were taken out of the analysis.

low correlations between the Early Numeracy Screener and the national tests, children that are identified as at risk for MLD with the Early Numeracy Screener are very likely to perform below the 20th percentile in the national test (Table 7).

Table 6. Model Fit Statistics for the Test of Invariance Across Gender and Age—Three-Factor Model.

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	TLI	ΔCFI	ΔRMSEA
Gender						
Configural invariance	2,542	.956	.041 (.037–.056)	.954		
Scalar invariance	2,588	.956	.041 (.037–.045)	.954	.000	.000
Age						
Configural invariance	2,542	.938	.044 (.040–.048)	.935		
Metric invariance						
Scalar invariance	2,588	.938	.044 (.040–.047)	.937	.000	.000

Table 7. Change and Stability in the Mathematics Achievement Groups Across the Early Numeracy Test and the National Test.

Configuration (EN–NT)	o	e	z	*p*(z)	
TA – TA	253	232.62	2.25	0.0243	
TA – MLD	39	59.38	−2.89	0.0038	Antitype
MLD – TA	33	53.38	−3.02	0.0025	Antitype
MLD – MLD	34	13.62	5.63	0.0001	Type

Note: EN = Early Numeracy test; NT = national test; o = observed frequencies; e = expected frequencies; TA = typical-achieving group; MLD = at risk for mathematical learning difficulties group.

Discussion

The main purpose of this study was to investigate the psychometric properties and to validate the Early Numeracy Screener. In line with EFPA (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013) the study evaluated reliability, construct validity, and criterion-related validity. First, the items showed acceptable discriminant abilities. Regarding construct validity, the analysis suggested a three-factor model over a one-factor model. When it comes to criterion validity, the correlations between the three factors are below .85, which is often used as a cut-off (Brown, 2006). Additionally, high correlations between the core numerical skills are to be expected since the skills are related. The results indicate that the Early Numeracy Screener reliably measures three distinct subskills of early numeracy: counting skills, numerical relational skills, and basic arithmetic skills. Moreover, these subskills were related to the national test scores, and children identified as at risk of developing mathematical learning disabilities with the Early Numeracy test were likely to perform below the 20th percentile in the national test in mathematics 6 months later. The scores on the Early Numeracy Screener were on average lower than the national test. This is probably due to the different times of the school year in which the two tests are administered. The Early Numeracy Screener was used at the very beginning in first grade, after only 5 weeks of schooling, and the children had little experience with mathematics instruction since this is not obligatory in Norwegian kindergartens. The national test is also designed with a ceiling effect, in which the sole aim is to identify those performing under the 20th percentile (Udir, 2016).

Construct Validity and Measurement Invariance

Concerning construct validity, first, we examined whether the Early Numeracy Screener worked best as a one-factor model or a three-factor model. The CFAs showed better model fit for the three-factor model based on counting skills, numerical relational skills, and basic arithmetic skills. This supports Aunio and Räsänen's (2016) theoretical model, which suggests that early numeracy skills are based on these three components. Support for assessing these skills is not found in Aunio and Räsänen's (2016) study alone; support is also found in previous research (Desoete et al., 2012; Gersten et al.,

2012; Moeller et al., 2011). Here we see that although the three-factor model had a slightly better fit than the one-factor model, these are indeed highly related skills. Still, the three-factor model here had a better fit to the data and therefore this was chosen over the more parsimonious one-factor model.

Criterion-Related Validity

The analyses of strong concurrent and predictive relationships between the Early Numeracy Screener and the national test showed respectable criterion validity for the Early Numeracy Screener. However, the correlation between the Early Numeracy Screener and the national test proved to be quite low. There could be at least two reasons for this. First, item types in the Early Numeracy Screener and national test do not overlap very much and this could imply that these two measures do not quite measure the same thing at item level; as all the items in the national test are timed, the children are given a certain amount of time, hence the national test also measures fluency, as opposed to the Early Numeracy Screener that measures accuracy and is without time limits. Many of the tasks in the national test required children to use more than one operation for solving (e.g., using both counting skills and relational skills within one task), which is not the case with the Early Numeracy Screener. In addition, in some of the tasks in the national test children were given one point if they had solved tasks consisting of multiple items (i.e., one task that omitted one point for a correct answer required the children to solve, for example, two items within the task correctly; if they manQ12 aged only one out of two, the task was corrected with zero points). Second, the national test has a large ceiling effect, and this attenuates correlations. Also, the national screening was taken 6 months after the Early Numeracy Screener.

However, when we look at the relationship between the Early Numeracy Screener and those beyond the critical limit on the national test, there is a high correspondence between the two measures. Given that the Early Numeracy Screener measures core skills, it is not surprising that the children identified as at risk of developing mathematical learning disabilities with the Early Numeracy Screener also have problems in national tests that are curriculum-based (the significant stability of the at-risk/at-risk configuration). At the same time, it is reasonable that some children that do not have problems in core skills might still struggle with the curriculum-based national test. However, they are probably not at risk of developing MLD; their difficulties might be related to other factors. Regarding how the screener predicts skills later on, in a general level, not solely children at risk of learning disabilities, the ceiling effect of the national test makes this impossible to say Q13 something about. Internal consistency of the Early Numeracy Screener and its subscales was also demonstrated to be adequate, and thus met the criteria evaluated through EFPA (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013).

Implications and Future Studies

In this study, the Early Numeracy Screener demonstrated a three-factor structure enabling a brief screener with broader content, albeit not every aspect of mathematics, but a screener that arguably focuses on the core competencies that define early numeracy in 6-year-olds (Gersten, Jordan, & Flojo, 2005). To ensure the validity of assessment tools to be used in different countries, they preferably need to be validated for each language and country. A challenge when validating tests in continental Europe is that there is a large number of languages used, and therefore validation is resource demanding. However, there are validated screening instruments in other European countries (Aunio et al., 2006; Brankaer et al., 2017; Van Luit et al., 1994; Van de Rijt et al., 1999;Van Luit et al., 2001). For future studies it would be interesting to increase the validity even further with a test–retest design of the screening tool and perhaps open for doing ROC-curve analysis adding standardized mathematical measures in that respect. In this sense, we did not have any other criterion-related measure with which to relate the Early Numeracy Screener (e.g., SYMP test [Brankaer et al., 2017]). However, the Early Numeracy Screener gives the teacher an opportunity to identify in which specific aspects of

early numeracy a child needs remediation. The brevity and ease of use of the Early Numeracy Screener make it well suited for classroom instructional settings. The present study contributes to the practice field in that it offers schools a screening tool for measuring early numeracy skills that the schools can use whenever they want, that is, unrestrained guidelines as to when pupils can take the national screener in mathematics that is normally administered at the end of the school year.

Disclosure Statement

Q14 potential conflict of interest was reported by the authors.

ORCID

Johan Korhonen http://orcid.org/0000-0001-9666-9400

Q15 References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Aunio, P., Hautamäki, J., Heiskari, P., & van Luit, J. (2006). The early numeracy test in Finnish: Children's norms. *Scandinavian Journal of Psychology*, *47*(5), 369–378.

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, *20*(5), 427–435.

Aunio, P., & Räsänen, P. (2016). Core numerical skills for learning mathematics in children aged five to eight years – a working model for educators. *European Early Childhood Education Research Journal*, *24*(5), 684–704.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody, & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 1–34). Mahwah, NJ: Erlbaum.

Baroody, A. J., Eiland, M., & Thompson, B. (2009). Fostering at-risk preschoolers' number sense. *Early Education and Development*, *20*, 80–128. doi:10.1080/10409280802206619

Baroody, J. (2006). *Why children have difficulties mastering the basic number combinations and how to help them*. National Council of Teachers of Mathematics. Retrieved from https://www.jstor.org/stable/41198838?origin=JSTOR- pdf&seq=1#metadata_info_tab_contents

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities*, *38*(4), 333–339. doi:10.1177%2F00222194050380040901

Brankaer, C., Ghesquière, P., & De Smedt, B. (2017). Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP test). *Behavior Research Methods*, *49*(4), 1361–1373.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Butterworth, B., Varma, S., & Laurillard, D. (2011). Dyscalculia: From brain to education. *Science*, *332*(6033), 1049–1053. doi:10.1126/science.1201536

Canobi, K. H., Reeve, R. A., & Pattison, P. E. (2002). Young children's understanding of addition concepts. *Educational Psychology*, *22*(5), 513–532. doi:10.1080/0144341022000023608

Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In B. Zumbo, & E. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences. Vol 54. Social indicators research series* (pp. 9–24). Switzerland: Springer.

Q16 Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. doi:10.1080/10705510701301834

Clarke, B., Doabler, C., Smolkowski, K., Nelson, E. K., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the immediate and long-term efficacy of a tier 2 kindergarten mathematics intervention. *Journal of Research on Educational Effectiveness*, *9*, 607–663. doi:10.1080/19345747.2015.11160 34

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based early maths. *Assessment, Educational Psychology*, *28*(4), 457–482.

Connely, A. J. (2007). *KeyMath-3 diagnostic assessment: Manual forms A and B*. Minneapolis, MN: Pearson.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. Retrieved from https://psycnet.apa.org/doi/10.1037/h0040957

De Smedt, B., Noël, M.-P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non- symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*(2), 48–55. doi:10.1016/j.tine.2013.06.001

De Smedt, B., Verschaffel, L., & Ghesquière, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *103*(4), 469–479. doi:10.1016/j.jecp.2009.01.010

Desoete, A., Ceulemans, A., De Weerdt, F., & Pieters, S. (2012). Can we predict mathematical learning disabilities from symbolic and non-symbolic comparison tasks in kindergarten? Findings from a longitudinal study. *British Journal of Educational Psychology*, *82*, 64–81. doi:10.1348/2044-8279.002002

Desoete, A., Stock, P., Schepens, A., Baeyens, D., & Rieyers, H. (2009). Classification, seriation, and counting in grades 1, 2, 3 as two-year longitudinal predictors for low achieving in numerical facility and arithmetical achievement? *Journal of Psychoeducational Assessment*, *27*(3), 252–264. doi:10.1177%2F0734282908330588

Dockrell, J., Llaurado, A., Hurry, J., Cowan, R., Flouri, E., & Dawson, A. (2017). *Review of assessment measures in the early years. Language and literacy, numeracy, and social-emotional development and mental health, education endowment foundation (EEF), Institute of education*. UCL.

Dougherty, C. (2003). Numeracy, literacy, and earnings: Evidence from the national longitudinal survey of youth. *Economics of EducationReview*, *22*, 511–521.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. doi:10.1037/0012-1649.43.6.1428

Evers, A., Hagemeister, C., & Hostmealingen, A. (2013). *EFPA a review model for the description and evaluation of psychological and educational tests*. Tech. Rep. Version 4.2.(6). Brussels: European Federation of Psychology Associations.

Evers, A., Muñiz, C., Høstmælingen, A., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, *25*(3), 283–291.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., … Changas, P. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology*, *105*, 58–77. doi:10.1037/a0030127

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology*, *47*(6), 1539–1552. doi:10.1037/a0025510

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool. *Journal of Psychoeducational Assessment*, *27*(3), 265–279. doi:10.1177%2F0734282908330592

Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology*, *88*(2), 121–151. doi:10.1016/j.jecp.2004.03.002

Geary, D. C., Hoard, M. K., Nugent, L., Bailey, D. H., & Krueger, F. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS ONE*, *8*(1), 1–8.

Geary, D. C., & vanMarle, K. (2016). Young children's core symbolic and nonsymbolic quantitative knowledge in the prediction of later mathematics achievement. *Developmental Psychology*, *52*(12), 2130–2144. doi:10.1037/dev0000214

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children*, *78*, 423–445. doi:10.1177/001440291207800403

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with math difficulties. *Journal of Learning Difficulties*, *38*, 293–304. doi:10.1177%2F00222194050380040301

Gersten, R., Rolfhus, E., Clarke, B., Decker, L., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal*, *52*, 516–546. doi:10.3102/0002831214565787

Goyder, J., Warriner, K., & Miller, S. (2002). Evaluating socio-economic status (SES) bias in survey nonresponse. *Journal of Official Statistics*, *18*(1), 1–11.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *International Journal of Science*, *455*, 665–668.

Hannula-Sormunen, M. M., Lehtinen, E., & Räsänen, P. (2015). Preschool children's spontaneous focusing on numerosity, subitizing, and counting skills as predictors of their mathematical performance seven years later at school. *Mathematical Thinking and Learning*, *17*(2-3), 155–177. doi:10.1080/10986065.2015.1016814

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology*, *103*(1), 17–29. doi:10.1016/j.jecp.2008.04.001

Hu, L.-t., & Bentler, P. M. (1991). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*, 647–660. doi:10.1037/a0029018

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

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcome. *Developmental Psychology*, *45*(3), 850–867.

Krajewski, K., & Schneider, W. (2009). Exploring the impact of phonological awareness, visual-spatial working memory, and preschool quantity-number competencies on mathematics achievement in elementary school: Findings from a 3-year longitudinal study. *Journal of Experimental Child Psychology*, *103*, 516–531. doi:10.1016/j.jecp.2009.03.009

Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares, & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

McGuire, P., Kinzie, M. B., & Berch, D. B. (2012). Developing number sense in pre-k with five-frames. *Early Childhood Education*, *40*(4), 213–222.

McIntosh, A. (2012). *Alle teller!* Trondheim: Matematikksenteret.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Moeller, K., Pixner, J., Zuber, J., Kaufmann, L., & Nuerk, H.-C. (2011). Early place-value understanding as a precursor for later arithmetic performance – A longitudinal study on numerical development. *Research in Developmental Disabilities*, *38*, 1837–1851. doi:10.1016/j.ridd.2011.03.012

Muthén, L. K., & Muthén, O. (1998) 2004. *Mplus user's guide*. (3rd ed.). Los Angeles, CA: Muthèn & Muthèn.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Virginia: Reston.

National Council on Measurement in Education. (1999).

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: US Department of Education.

National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Washington, DC: The National Academies Press.

Purpura, D. J., Baroody, A. J., & Lonigan, C. J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology*, *105*(2), 453–464. doi:10.1037/a0031753

Purpura, D. J., & Lonigan, C. J. (2013). Informal numeracy skills: The structure and relations among numbering, relations and arithmetic operations in preschool. *American Educational Research Journal*, *50*(19), 178–209. doi:10.3102%2F0002831212465332

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301–1308. doi:10.1177%2F0956797612466268

Salminen, J. B., Koponen, T. K., & Tolvanen, A. L. (2018). Individuality in the early number skill components underlying basic arithmetic skills. *Frontiers in Psychology*, *9*, 1–11. doi:10.3389/fpsyg.2018.01056

Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S., Stricker, J., & De Smedt, B. (2016). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*. doi:10.1111/desc.12372

Statistics Norway. (2019). Retrieved from https://www.ssb.no/utniv

Stock, P., Desoete, A., & Roeyers, H. (2009). Detecting children with arithmetic disabilitiesfrom kindergarten: Evidence from a 3-year longitudinal study on the role ofpreparatory arithmetic abilities. *Journal of Learning Disabilities*, *43*, 250–268. doi:10.1177/0022219409345011

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., De Vet, H. C. W., Westerman, M. J., Patrick, D. L., … Mokkink, L. B. (2017). COSMIN standards and criteria for evaluating the content validity of health-related Patient-reported Outcome measures: A Delphi study. *Quality of Life Research*, in press.

Thorndike, R. M., & Thorndike-Christ, T. (2014). *Measurement and evaluation in psychology and education* (8th ed.). Harlow: Pearson.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco: Jossey-Bass.

Udir. (2016). The Norwegian Directorate of Education and Training.

Vanbinst, K., Ghesquière, P., & De Smedt, B. (2015). Does numerical processing uniquely predict first graders' future development of single-digit arithmetic? *Learning and Individual Differences*, *37*, 153–160. doi:10.1016/j.lindif.2014.12.004

Von Eye, A. (1990). *Introduction to configural frequency analysis. The search for types and anti-types in cross classification*. Cambridge: University Press.

Von Eye, A., Spiel, C., & Wood, P. H. (1996). Configural frequency analysis. *Applied Psychological Research*, *45*(4), 310–327.

Van de Rijt, B. A. M., Van Luit, J. E. H., & Pennings, A. H. (1999). The construction of the Utrecht early mathematical competence scale. *Educational and Psychological Measurement*, *59*, 289–309. doi:10.1177%2F00131644995920006

Van Luit, J. E. H., Van de Rijt, B. A. M., & Hasemann, K. (2001). *OTZ Osnabrücker test für zahlbegriffsentwicklung* [German version of the Uterecht test of number sense]. Göttingen: Hogrefe-Verlag.

Van Luit, J. E. H., Van de Rijt, B. A. m., & Pennings, A. H. (1994). *Utrechtse Getalbegrip Toets* [Utrecht test of number sense]. Doetinchem: Graviant.

Wilkins, J. L. M., Baroody, A. J., & Tiilikainen, S. (2001). Kindergartner's understanding of additive commutativity within the context of word problems. *Journal of Experimental Child Psychology*, *79*(1), 23–26.

Wright, R. J., Martland, J., & Stafford, A. K. (2006). *Early numeracy. Assessment for teaching and intervention* (2nd ed.). London: Sage.

Zhang, X., Koponen, T., Räsänen, P., Aunola, K., Lerkkanen, M.-K., & Nurmi, J.-E. (2014). Linguistic and spatial skills predict early arithmetic development via counting sequence knowledge. *Child Development*, *85*(3), 1091–1107. doi:10.1111/cdev.12173

Zumbo, B. D., & Chan, E. K. H. (2009). Setting the stage for validity and validation in social, behavioral, and health sciences: Trends in validation practices. In B. Zumbo, & E. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences*. Social indicators research series, vol 54. (pp. 3–8). Springer.

Q22

Appendices

Q23 *Appendix A*

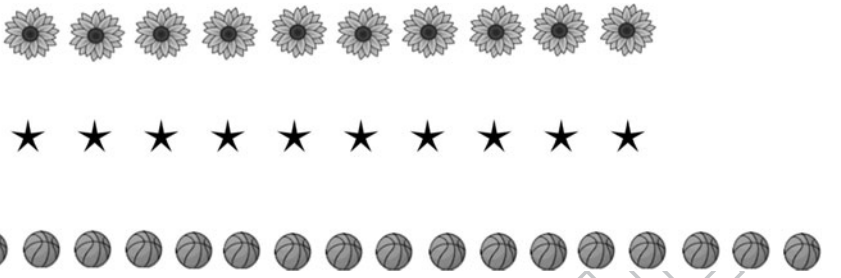
Q24 Table A1.

Item	Skill	% Correct	Alpha if item removed	Retained
1	Numerical relational	95.4%	.943	No
2	Numerical relational	76%	.942	Yes
3	Numerical relational	92.3 %	.943	Yes
4	Numerical relational	92.9 %	.942	Yes
5	Counting	82%	.942	Yes
6	Counting	88%	.942	Yes
7	Counting	75.1%	.942	Yes
8	Counting	84.2%	.943	Yes
9	Counting	48%	.942	Yes
10	Counting	45.6%	.942	Yes
11	Numerical relational	91.8%	.942	Yes
12	Numerical relational	51.6%	.942	Yes
13	Numerical relational	42.1%	.942	Yes
14	Numerical relational	39.6%	.942	Yes
15	Numerical relational	67.8%	.942	Yes
16	Numerical relational	21.3%	.942	Yes
17	Numerical relational	95.4%	.943	No
18	Numerical relational	50.5%	.942	Yes
19	Numerical relational	46.4%	.942	Yes
20	Numerical relational	42.7%	.942	Yes
21	Numerical relational	41%	.942	Yes
22	Numerical relational	37.7%	.943	Yes
23	Counting	77.9%	.943	Yes
24	Counting	74%	.942	Yes
25	Counting	54.1%	.942	Yes
26	Counting	80.9%	.942	Yes
27	Counting	71.6%	.941	Yes
28	Counting	49.5%	.941	Yes
29	Counting	68.9%	.941	Yes
30	Counting	56.8%	.941	Yes
31	Counting	42.9%	.941	Yes
32	Counting	58.2%	.941	Yes
33	Counting	36.1%	.941	Yes
34	Counting	35.5%	.942	Yes
35	Counting	84.2%	.942	Yes
36	Counting	74%	.942	Yes
37	Counting	64.5%	.942	Yes
38	Counting	85.5%	.941	Yes
39	Counting	71%	.941	Yes
40	Counting	48.9%	.942	Yes
41	Counting	87.7%	.942	Yes
42	Counting	83.6%	.941	Yes
43	Counting	61.2%	.941	Yes
44	Counting	41.3%	.941	Yes
45	Addition	72.7%	.941	Yes
46	Addition	67.8%	.941	Yes
47	Addition	63.4%	.941	Yes
48	Addition	71.3%	.941	Yes
49	Addition	55.2%	.941	Yes
50	Addition	35.2%	.941	Yes
51	Subtraction	28.7%	.941	Yes
52	Subtraction	24.6%	.941	Yes
53	Subtraction	20.8%	.941	Yes
54	Subtraction	20.2%	.941	No
55	Subtraction	19.4%	.941	Yes
56	Subtraction	18. 3%	.941	No

Appendix B: Sample Items for the Early Numeracy Screener

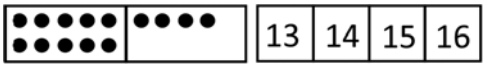
Q25 Note: The instructions are translated from Norwegian to English. Teacher instructions are displayed.

Counting skills:



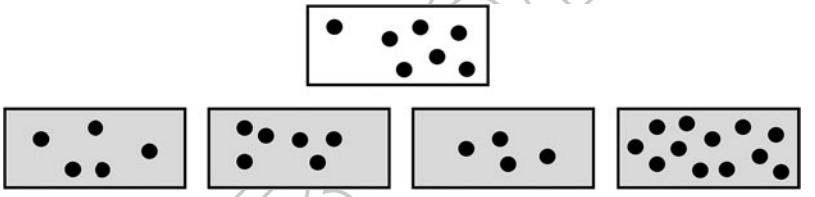
Tick on the second flower, tick on the sixth star, tick on the fourteenth ball.

Counting skills:



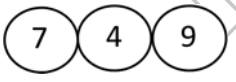
Beside the box with black dots, there are boxes with numbers in them. First, find out how many black dots there are all together, and tick the box that says how many black dots there is in the box.

Numerical relational skills:



There are black dots in the white box. Tick the grey box that has one black dot less than the white does.

Numerical relational skills:



Here are three numbers. Tick on the smallest number.

Basic arithmetic skills:

2 + 5 = _____

8 + 3 = _____

Here you have addition tasks. Solve as many of them as you can.

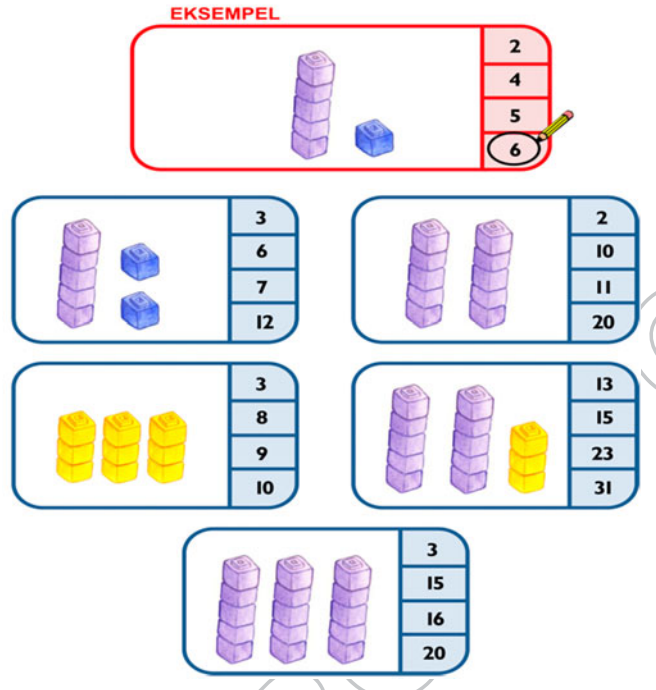
7 - 4 = _____

11 - 3 = _____

Here you have subtraction tasks. Solve as many of them as you can.

Q26 Appendix C

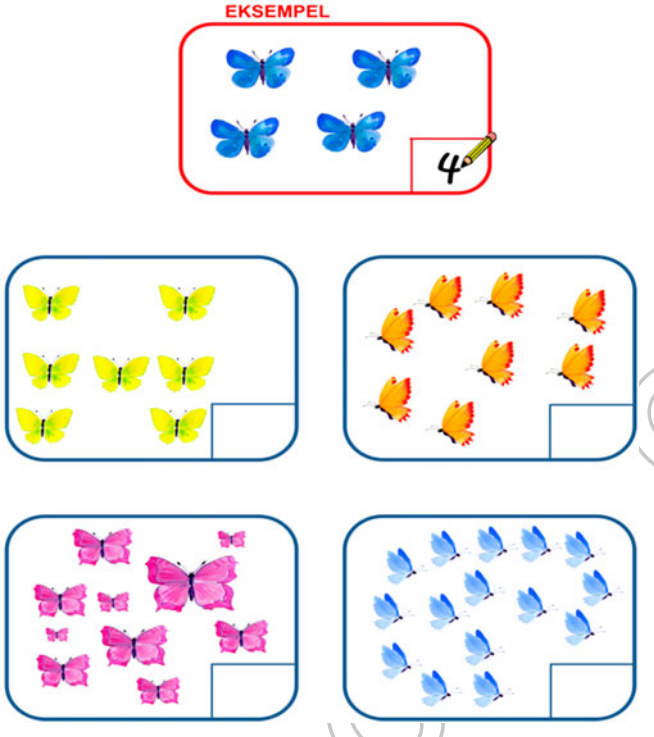
Q5 Note: The instructions are translated from Norwegian to English. Teacher instructions are displayed.
"How many", time limit: 1.5 min



On this page you are going to find out how many bricks there are in each task. Look at the example (point). There are six bricks. That is why there is a circle around the number six. Now you are going to find out how many bricks there are in the other tasks.

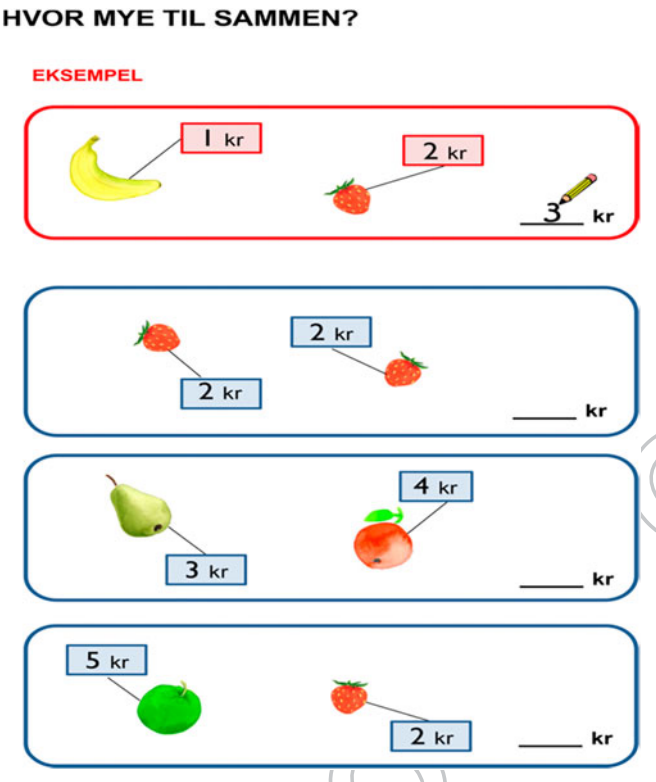
"How many", time limit: 1.5 min

HVOR MANGE?



Here you need to find out how many butterflies there are in each task and write the amount in the little square. Look at the example (point). There it is four butterflies. That is why the number four is written in the small square. Now you are going to do the rest of the tasks on this page.

"How much all together", time limit: 1.5 min
Note: krone(r) (kr = NOK)



On this page you should imagine that you are buying things and you need to find out how much you need to pay. The price tag tells how much each thing costs. Look at the example. The banana costs 1 krone and the strawberry costs 2 kroner. Then we need to pay 3 kroner in total, which is why it is written two on the line (point). Now you are going to find out how much you have to pay for the stuff in the other tasks.