International Journal of Educational Research 102 (2020) 101580

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

Reliability and validity evidence of the early numeracy test for identifying children at risk for mathematical learning difficulties

Heidi Hellstrand[a,*], Johan Korhonen[a], Pekka Räsänen[c,d], Karin Linnanmäki[a], Pirjo Aunio[b]

[a] Faculty of Education and Welfare Studies, Åbo Akademi University, Vasa, Finland
[b] Faculty of Educational Sciences, University of Helsinki, Helsinki, Finland
[c] Niilo Mäki Institute, University of Jyväskylä, Jyväskylä, Finland
[d] Social and Health Division, City of Helsinki, Finland

ARTICLE INFO

Keywords:
Assessment
Early numeracy
Mathematical learning difficulties
Screener
Validation

ABSTRACT

This study investigated reliability and validity evidence regarding the Early Numeracy test (EN-test) in a sample of 1139 Swedish-speaking children (587 girls) in kindergarten (n = 361), first grade (n = 321), and second grade (n = 457). Structural validity evidence was established through confirmatory factor analysis (CFA), which showed that a four-factor model fit the data significantly better than a one-factor or two-factor model. The known-group and cross-cultural validity were established through multigroup CFAs, finding that the four-factor model fit the gender, age and language groups equally well. Internal consistency for the test and sub-skills varied from good to excellent. The EN-test can be considered as an appropriate assessment to identify children at risk for mathematical learning difficulties.

1. Introduction

Early numeracy skills are vital for later learning in mathematics (Aunola, Leskinen, & Nurmi, 2006; Jordan, Fuchs, & Dyson, 2015). Early identification of children at risk is essential to prevent mathematical learning difficulties (Gersten, Clarke, Haymond, & Jordan, 2011), but it requires valid and reliable assessment tools (Aunio, 2019; Purpura & Lonigan, 2015). Screening tests are suggested to be a quick procedure to identify children at risk but should not stand alone as a basis for diagnostic decisions (Gersten et al., 2011).

There is a need to develop appropriate assessment tools that not only differentiate students but also give more detailed information regarding children's performance and development (Purpura & Lonigan, 2015). Several assessment points are also important to follow up with children's learning and to plan targeted instructions and interventions (Aunio, 2019). Curriculum-based measurement (CBM) has been suggested as a reliable and valid measure of children's performance, as it combines the advantages of standardized achievement tests and curriculum-based assessments developed by teachers, and it strives to minimize the gap between the measurement and instruction (Fuchs, 2016). Deno (1985) describes the criteria for applicable CBM: reliable and valid measures, simple and efficient for educators to use and understand, inexpensive to use, and enable repeated measurement to follow children's mathematical learning process. CBM can be accomplished either through a curriculum sampling approach or the Robust indicator method (Fuchs, 2016). In this study, the reliability and validity of the Early Numeracy test (EN-test) for identifying children at risk for mathematical learning difficulties were evaluated. The EN-test was developed following the Robust indicator method of designing

* Corresponding author.
E-mail address: heidi.hellstrand@abo.fi (H. Hellstrand).

https://doi.org/10.1016/j.ijer.2020.101580
Received 17 December 2019; Received in revised form 14 April 2020; Accepted 20 April 2020

CBM, meaning that the test relies on indicators representing the core competencies in early numeracy for this specific age group instead of relying on the curriculum (Aunio, 2019).

Currently, there is a wide range of tools used to assess the early numeracy individually or in groups. The strength of individual assessments is that the administrator can guide and monitor the performance during the assessment, whereas the group-administered assessment is generally less time consuming (Gregory, 2015). Many assessment tools either broadly measure mathematic skills and concepts or focus more deeply on a specific mathematical skill. Therefore, they lack substantial coverage of areas, such as geometry or measurement (Purpura & Lonigan, 2015). To underpin the value of the EN-test, we looked at early numeracy tests with good predictive value and assessments that are useable regardless national curriculum and focus at identifying children at risk for mathematical learning difficulties in kindergarten and first grades. We selected tests that are available and fulfil the criteria of CBM with Robust indicator method. Broadly measuring tests (e.g., Woodcock-Johnson achievement test) or tests for only psychologists were omitted from our review. A total of nine tests met the selection criteria (Table 1). Eight of the tests were individually administered and interview based. The Number Sets Test (Geary, Bailey, & Hoard, 2009) was the only group-administered test, though it mainly focused on fluency in a narrow area of early numeracy. All tests focused on early numeracy, and only the Research-based Early Mathematics Assessment (REMA; Clements, Sarama, & Liu, 2008) and Child Math Assessment (CMA; Starkey, Klein, & Wakeley, 2004) included both geometric and statistics. Only the Early Numeracy test (WENT; Wright, Marltand, & Stafford, 2006) separated the items into subcategories, the other tests provided only one general score (unidimensional approach). Some tests did not explicitly provide test items and the psychometric properties were only reported for six of the tests (REMA: Clements et al., 2008; Number Sets Test: Geary et al., 2009; Test of Early Mathematics Ability, TEMA-3: Ginsburg & Baroody, 2003; the Number Knowledge Test, NKT: Okamoto & Case, 1996; Number Sense Screener, NSS™: Jordan, Glutting, & Dyson, 2012; the Utrecht Test of Early Numeracy, ENT-test: Van Luit, Van de Rijt., & Pennings, 1994). Most of the standardized early numeracy measures have been developed in English language and mainly in the United States. In Europe, challenges have included national language differences and resources needed to standardize new assessment tools. As, a result, there is a need for reliability and validity evidence for local early numeracy tests.

This study aimed to investigate the reliability and validity evidence of the EN-test for identifying children at risk for mathematical learning difficulties. The EN-test was developed following the Robust indicator method of designing CBM, where the test is constructed based on indicators of core competencies (instead of the curriculum) that are seen as good predictors of later mathematical learning. The validity of the test was examined by collecting evidence related to structural validity, known group validity, and cross-cultural validity. The reliability evidence of the test was analyzed based on internal consistency. The criteria chosen for reliability and validity was based on the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology (Terwee et al., 2017) and standards for educational and psychological testing (AERA, APA, NCME, 2014).

2. Method

2.1. Participants

A total of 1139 children (587 girls) was recruited from Swedish-speaking kindergartens and schools in Finland. The sample consisted of 361 children in kindergarten (185 girls), 321 children in first grade (162 girls), and 457 children in second grade (240 girls). The children ranged in age in kindergarten from 61 months to 85 months (M_{age} = 73.97; SD = 3.65), in first grade from 80 months to 111 months (M_{age} = 86.87; SD = 3.94), and in second grade from 91 months to 115 months (M_{age} = 98.38; SD = 3.65).

The sample consisted of children from different-sized schools in urban and rural areas of Swedish-speaking Finland.[1] Children were drawn from 23 kindergartens and 26 primary schools. Most of the children spoke Swedish as their native language (kindergarteners 87.8 %, first graders 80.7 %, and second graders 88.4 %). Of those who had a native language other than Swedish, the largest group were Finnish speaking (kindergarteners 11.9 %, first graders 17.1 %, and second graders 10.7 %). The teachers reported the children's language. All children attended neighborhood schools, and no special education classes were included in the study.

2.2. Measurements

The EN-test was constructed to identify children at risk for mathematical learning difficulties in kindergarten, and first and second grade (Koponen et al., 2011a, 2011b, 2011c). The test is based on the theoretical model of core numerical skills for learning mathematics in children aged 5–8, and focus on four skill groups: symbolic and non-symbolic number knowledge, understanding mathematical relations, counting skills, and basic skills in arithmetic (Aunio & Räsänen, 2015). A multi-professional and multilingual team constructed the EN-test within the LukiMat project, funded by the Finnish Ministry of Education and Culture.

[1] The education system and bilinguality in Finland. Compulsory formal education consists of nine years of comprehensive school, starting in August the year the child turns seven. To prepare children for formal schooling, they enter a one-year kindergarten. Teachers in comprehensive schools have a master's degree from university, and teachers in kindergarten have a bachelor's or master's degree. Finland has two official languages, Finnish and Swedish. The Swedish-speaking population is a minority (5.3%, according to Statistics Finland) in Finland, and they live mainly on the west coast and in southern parts of Finland. The Swedish-speaking population has the right to use and get services in their own language; their education is arranged in Swedish-speaking schools and kindergartens based on the national curriculum frameworks. Finland is a relatively ethnically homogeneous country, where only 7.3% of the whole population has another ethnicity than Finnish, and the Finnish school system provides relatively equal educational opportunities irrespective of the students' socio-economic background and place of residence (OECD, 2016).

Table 1
Overview of early numeracy tests.

Test	Author(s)	Main components	Language	Age group	Administration	Purpose	Items; reliability
Research-based Early Mathematics Assessment (REMA)	Clements et al., 2008	Object counting, subitizing, number comparison and sequencing, connecting numerals to quantities, number recognition, composition/decomposition, adding, subtracting, place value, congruence and construction of shapes, spatial imagery, measurement, and patterning.	English	3 – 8 years (full version), 4 – 5 years (short version)	Individually, interview based; 60 min/child	Screening and formative assessment	125 items, KR-20 = .98[a] (full version); 19 items, α = .71 – 79[b] (short version)
Number Sets Test	Geary et al., 2009	Fluency in identifying and processing quantities represented by numerals and object sets.	English	4 – 8 years	Group, paper-pen; 10 minutes	Screening	42 items, α = .88[c]
Test of Early Mathematics Ability, Third Edition (TEMA-3)	Ginsburg & Baroody, 2003	Numbering skills, number-comparison facility, numeral literacy, mastery of number facts, calculation skills, and understanding of concepts.	English	4 – 8 years	Individually; 40 min/child,	Norm-referenced measure, screening and diagnostic	Two parallel forms, á 72 items, α = .92[d]
The Number Knowledge Test (NKT)	Okamoto & Case, 1996	Comparing and quantifying sets and numbers, number sequence, basic arithmetic, and counting strategies.	English	4 – 10 years	Individually; 10 – 15 min/child	Screening	Unidimensional, a total score (0 – 32 points), α = .94[e]
Number Sense Screener (NSS™)	Jordan et al., 2012	Counting, number recognition, number comparisons, nonverbal calculations, story problems, and number combinations.	English, Turkish,	5 – 6 years	Individually, interview based; 15 – 20 min/child	Screening and diagnostic, base for instruction and intervention; Norm-referenced	26 items, six subareas, α = .85[f]
MARKO-D test	Ricken, Fritz, & Balzer, 2013	Counting numbers, mental number line, cardinality and decomposability, class inclusion and embeddedness, relationality, and units in numbers.	German, Four South African languages	4 – 8 years	Individually, interview based; 40 min/child	Screening and diagnostic	55 items; unidimensional
Child Math Assessment (CMA)	Starkey et al., 2004	Object counting, counting a subset, number order, number comparison, ordinal number terms, number reproduction, addition and subtraction with and without concrete objects, two-set addition, knowledge of shape names, shape matching, reasoning about triangle transformations, nonstandard measurement, pattern duplications and extension, and ordering series of objects.	English	3 – 5 years	Individually; 2 × 20 – 30 min/child	Broad assessment of skill level	16 tasks; unidimensional
The Utrecht Test of Early Numeracy (ENT-test)	Van Luit et al., 1994	Comparison, classification, one-to-one-correspondence, seriation, number words, counting, and understanding of numbers.	Dutch, English, Farsi, Finnish, Chinese	4 – 8 years	Individually, interview based; 20 – 30 min/child	Screening	40 items, unidimensional, α = .90[g]
The Early Numeracy Test (WENT)	Wright et al., 2006	Single-digit arithmetic from enumeration to more advanced strategies and base-10 strategies, forward and backward number word sequences, numeral identification, simultaneous numerical processes (i.e., spatial and finger patterns, subitizing, combining, and partitioning. grouped quantities).	English	4 – 9 years	Individually, interview based	Screening and diagnostic	Three subareas

Note. Reliability reported in: [a]Clements et al., 2008; [b]Weiland et al., 2012; [c]Geary et al., 2009; [d]Ginsburg & Baroody, 2003; [e]Gersten et al., 2007; [f]Jordan et al., 2012; [g]Aunio et al., 2006.

Table 2
Number of items, reliability coefficients and content in the early numeracy test.

	Kindergarten			First grade			Second grade		
	items	α	corrected item-total correlation	items	α	corrected item-total correlation	items	α	corrected item-total correlation
Symbolic and non-symbolic number knowledge (NK) Comparing magnitudes, approximate counting	8	.87	[.33, .75]	8	.70	[.16, .49]	8	.75	[.22, .61]
Understanding mathematical relations (MR) Seriation, comparison, classification, one-to-one correspondence, basic arithmetic principles (additive composition, commutativity, associativity, inversion), understanding mathematical symbols, place-value, and base-ten system	16	.79	[.18, .59]	16	.71	[.17, .47]	8	.93	[.71, .79]
Counting skills (CS) Counting up and back, skip, count from given number, number identification, recognition and writing, counting numerosity of a set, and counting part of a whole	16	.84	[.19, .63]	16	.86	[.35, .55]	8	.81	[.18, .78]
Basic skills in arithmetic (BA) Addition and subtraction in verbal story problems and with symbols	8	.65	[.25, 44]	16	.81	[.09, .66]	40	.94	[.15, .72]
	48	.92	[.16 – .57]	56	.91	[.15 – .55]	64	.95	[.13 – .66]

Note. α = Cronbach's alpha.

Table 3
Summary of the participants performance in the early numeracy test by gender and age groups.

		Age (months)	Total	NK	MR	CS	BA
		N	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Kindergarten							
All		361	73.97 (3.65)	39.27 (7.36)	7.03 (1.82)	13.13 (2.88)	6.73 (1.46)
Gender							
	girls	185	74.05 (3.59)	39.58 (7.49)	7.14 (1.76)	13.09 (2.99)	6.78 (1.46)
	boys	176	73.88 (3.73)	38.94 (7.22)	6.91 (1.89)	13.17 (2.78)	6.68 (1.48)
Age							
	younger	193	71.05 (2.02)	38.34 (7.39)	6.83 (2.01)	12.88 (2.87)	6.54 (1.54)
	older	168	77.32 (1.73)	40.33 (7.22)	7.25 (1.56)	13.42 (2.88)	6.95 (1.34)
First grade							
All		321	86.87 (3.94)	45.60 (7.90)	6.26 (1.84)	12.69 (3.32)	13.38 (2.70)
Gender							
	girls	162	86.80 (3.94)	45.10 (7.92)	6.20 (1.85)	12.65 (3.10)	13.30 (2.70)
	boys	159	86.94 (3.94)	46.11 (7.87)	6.31 (1.82)	12.74 (3.55)	13.47 (2.70)
Age							
	younger	163	83.76 (2.00)	44.57 (7.76)	6.06 (1.90)	12.48 (3.28)	13.00 (2.70)
	older	158	90.07 (2.67)	46.67 (7.92)	6.50 (1.73)	12.92 (3.36)	13.78 (2.64)
Second grade							
All		457	98.38 (3.65)	39.20 (10.61)	7.07 (1.51)	7.11 (1.61)	19.17 (7.23)
Gender							
	girls	240	98.54 (3.74)	38.02 (9.68)	6.92 (1.55)	6.95 (1.69)	18.40 (6.42)
	boys	217	98.20 (3.56)	40.51 (11.43)	7.24 (1.44)	7.29 (1.50)	20.01 (7.96)
Age							
	younger	235	95.44 (1.88)	37.34 (10.58)	6.90 (1.64)	6.95 (1.71)	17.84 (6.96)
	older	222	101.49 (2.22)	41.18 (10.31)	7.25 (1.33)	7.28 (1.48)	20.58 (7.26)

Note. NK = Symbolic and non-symbolic number knowledge; MR = Understanding mathematical relations; CS = Counting skills; BA = Basic skills in arithmetic.

The EN-test is available for kindergarten, first grade, and second grade. This paper-and-pencil test was conducted in groups, with teachers giving verbal instructions. The teacher used a manual with written detailed instructions for each task. In kindergarten and first grade, eight items measuring verbal counting skills were administered individually. In second grade, 20 addition and 20 subtraction tasks were measured within a two-minute time limit to assess arithmetic fluency with small numbers. Example tasks are presented in the supplementary materials (Table A1). The total number of items, the number of items measuring each sub-skill, and the number range within tasks differed between the grades. There were 48 items in kindergarten, 56 items in first grade, and 64 items in second grade covering the four skill groups (Table 2). The descriptive statistics of the participants' performance in the EN-test are presented in Table 3.

2.3. Procedure

Data collection took place in the beginning of the academic year. Permission for children to participate in this study was obtained in writing from their parents. The study followed the ethical guidelines of Åbo Akademi University. The testing took place in an ordinary classroom setting during one or two lessons. The tests were administered by classroom or special education teachers trained by the researchers, and the teachers received detailed instruction regarding how to execute the tests. Finnish teachers have a university degree and are trained to conduct assessments. The first author and trained research assistants corrected and coded the children's answers. Correct answers were awarded one point, while wrong or empty answers yielded zero points.

2.4. Data analyses

Structural validity of the EN-test was tested through confirmatory factor analyses (CFA). CFAs were conducted separately for kindergarten, first grade, and second grade. Due to categorical data, the parameters of the models were estimated using the weighted least squares means and variance estimation (WLSMV). The goodness of the model fit was evaluated using the chi-square test, the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the root mean square error of approximation (RMSEA).

Known-group validity was tested by investigating measurement invariance across gender (girls vs. boys) and age (median split: younger vs. older) using multigroup CFA. A model that imposed no invariance constraints but assumed the same factor structure in both groups (configural invariance) was compared with a model that constrained all factor loadings and item thresholds to be equal across groups (scalar invariance).

Cross-cultural validity was tested by investigating measurement invariance across the Swedish and Finnish language versions of the EN-test using multigroup CFA. The data for the Finnish sample (kindergarten $n = 563$; first grade, $n = 462$; and second grade, $n = 622$) was retrieved from another study.

Internal consistency reliability was calculated using Cronbach's alpha coefficient to assess how well all the items (kindergarten,

Table 4
Goodness-of-Fit Indicators of the Model for alternative CFAs in kindergarten, first grade and second grade.

	χ^2	*df*	CFI	TLI	RMSEA	Δ χ^2	*p*	ΔCFI	ΔRMSEA
Kindergarten (48 items)									
Four-factor	1401.180	1074	.935	.932	.029				
Three-factor	1530.963	1077	.910	.906	.034	48.953	< .001	.025	.005
One-factor	1684.776	1080	.881	.875	.039	139.090	< .001	.054	.010
First grade (56 items)									
Four-factor	2033.487	1478	.885	.880	.034				
Three-factor	2133.735	1481	.865	.860	.037	63.403	< .001	.020	.003
One-factor	2499.482	1484	.790	.782	.046	245.388	< .001	.095	.012
Second grade (64 items)									
Four-factor	2664.914	1946	.967	.966	.028				
Three-factor	2747.181	1949	.964	.962	.030	64.870	< .001	.003	.002
One-factor	3393.660	1952	.934	.932	.040	358.644	< .001	.033	.012

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation.

first grade, and second grade) are related to each other and measures the same construct. A value higher than 0.70 is recommended for Cronbach's alpha coefficient.

3. Results

3.1. Structural validity

To address the structural validity, CFAs were conducted on the EN-test for three different age groups. A four-factor model, a three-factor model, and one-factor model were fitted to the data for each age group. The four-factor model reflected the core group model: symbolic and non-symbolic number knowledge (NK), understanding mathematical relations (MR), counting skills (CS), and basic skills in arithmetic (BA). This model was compared to a one-factor model representing an overall numerical skills factor model and a three-factor model where NK and MR were combined into one factor, which is consistent with the Krajewski and Schneider (2009) model. When comparing nested models, a change of more than .01 in CFI and .015 in RMSEA indicates that the more complex model (four-factor model) fits the data better than the more parsimonious model (the one-factor model and the three-factor model; Chen, 2007).

3.1.1. Kindergarten

In kindergarten, the overall goodness-of-fit indices of the test suggested that the four-factor model fit the data well: χ^2 (1074) = 1401.18, $p < .001$; CFI = .94; TLI = .93; RMSEA = .03. The four-factor model also described the data significantly better than the one-factor model: ΔCFI = .05; ΔRMSEA = .01; $\Delta\chi^2$ (6) = 139.09, $p < .001$ and the three-factor model: ΔCFI = .03; ΔRMSEA = .01; $\Delta\chi^2$ (3) = 48.95, $p < .001$. The fit indices of the different models are presented in Table 4. The correlations between the latent factors ranged from .60 to .89, indicating a strong correlation between the skill groups. The strongest correlation (.89) was found between the latent variables of CS and BA. All factor loadings were found to be significant ($p < .001$).

3.1.2. First grade

In first grade, the overall goodness-of-fit indices suggested that the four-factor model fit the data adequately: χ^2 (1478) = 2033.49, $p < .001$; CFI = .89; TLI = .88; RMSEA = .03. The four-factor model also described the data significantly better than the one-factor model: ΔCFI = .10; ΔRMSEA = .01; $\Delta\chi^2$ (6) = 245.39, $p < .001$ and the three-factor model: ΔCFI = .02; ΔRMSEA = .00; $\Delta\chi^2$ (3) = 63.40, $p < .001$ (Table 4). The correlations between the latent factors ranged from moderate to strong between the skill groups (.43–.61). The strongest correlations were found between CS and NK (.61) and between CS and BA (.60). The correlations among the other latent variables were more similar in terms of strength. All factor loadings were also found to be significant ($p < .001$).

3.1.3. Second grade

In second grade, the overall goodness-of-fit indices showed that the four-factor model displayed excellent model fit: χ^2 (1946) = 2664.91, $p < .001$; CFI = .97; TLI = .97; RMSEA = .03. The four-factor model also described the data significantly better than the one-factor model: ΔCFI = .03; ΔRMSEA = .01; $\Delta\chi^2$ (6) = 358.64, $p < .001$ and the three-factor model: ΔCFI = .00; ΔRMSEA = .00; $\Delta\chi^2$ (3) = 64.87, $p < .001$. The fit indices of the different models are presented in Table 4. The correlations between the latent factors ranged from moderate to strong (.53–.74) between the skill groups. The strongest correlation (.74) was found between CS and NK. The correlations among the other latent variables were more similar in terms of strength. All factor loadings were significant ($p < .001$). Supplementary materials (Table B1, B2, and B3) provides detailed information of correlations, factor loadings and residuals for the models.

Table 5
Goodness-of-Fit Indicators of the Model for known-group and cross-cultural validity in kindergarten, first grade and second grade.

		χ^2	*df*	CFI	TLI	RMSEA	Δ χ^2	*p*	ΔCFI	ΔRMSEA
Kindergarten										
Gender										
	Configural	2507.099	2148	.926	.922	.030				
	Scalar	2543.973	2188	.927.	.924	.030	57.567	.036	.001	.000
Age										
	Configural	2514.830	2148	.922	.918	.031				
	Scalar	2538.264	2188	.925	.923	.030	40.638	.442	.003	.001
Language group										
	Configural	2916.703	2148	. 960	.958	.028				
	Scalar	3014.280	2188	.957	.955	.029	130.106	< .001	.003	.001
First grade										
Gender										
	Configural	3479.005	2956	.892	.888	.032				
	Scalar	3512.254	3004	.895	.893	.033	64.775	.054	.003	.001
Age										
	Configural	3422.119	2956	.887	.882	.032				
	Scalar	3452.545	3004	.891	.888	.031	54.006	.256	.004	.001
Language group										
	Configural	4373.262	2956	.908	.904	.035				
	Scalar	4431.480	3004	.907	.905	.035	123.220	< .001	.001	.000
Second grade										
Gender										
	Configural	3924.789	3178	.961	.959	.032				
	Scalar	3978.095	3228	.961	.960	.032	67.798	.048	.000	.000
Age										
	Configural	3799.282	3066	.964	.962	.032				
	Scalar	3863.275	3115	.963	.962	.032	103.477	< .001	.001	.000
Language group										
	Configural	3411.107	2338	.985	.984	.029				
	Scalar	3466.365	2380	.984	.984	.029	69.365	.005	.001	.000

Note. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation.

3.2. Known-group validity

3.2.1. Kindergarten

In Kindergarten, the multigroup CFAs for gender and age showed that constraining the factor loadings and item thresholds to equality did not significantly worsen the fit of the models, thus supporting measurement invariance (Table 5).

3.2.2. First grade

In first grade, the multigroup CFAs for gender and age showed that constraining the factor loadings and item thresholds to equality did not significantly worsen the fit of the models, thus supporting measurement invariance (Table 5).

3.2.3. Second grade

In second grade, the four-factor model with the four latent variables did not converge. Therefore, item descriptive statistics were examined to identify too easy (solving rate > 95 %) or too difficult (solving rate < 5%) items in BA. Consequently, in the analyses with age, six items (BA 1, 2, 3, 19, 20, and 21), and with gender, seven items (BA 1, 2, 3, 4 20, 21, and 22), were discarded. After discarding these items, the models were found to be invariant across gender and age (Table 5).

3.3. Cross-cultural validity

3.3.1. Kindergarten

In Kindergarten, the multigroup CFAs for the Swedish and Finnish EN-test showed that constraining the factor loadings and item thresholds to equality did not significantly worsen the fit of the models, thus supporting measurement invariance (Table 5).

3.3.2. First grade

In first grade, the multigroup CFAs for the Swedish and Finnish EN-test showed that constraining the factor loadings and item thresholds to equality did not significantly worsen the fit of the models, thus supporting measurement invariance (Table 5).

3.3.3. Second grade

In second grade, the four-factor model with the four latent variables did not converge, because of specific items in BA. Therefore, item descriptive statistics were examined to identify too easy (solving rate > 95 %) or too difficult (solving rate < 5%) items in BA.

Consequently, in the analyses, 14 items (BA 1–4, 17–22, 24, and 38–40), were discarded. After discarding these items, the four-factor model were found to be invariant across language versions (Table 5).

3.4. Internal consistency

3.4.1. Kindergarten

In kindergarten, the internal consistency was excellent for the total score (α = .92). The corrected item-total correlation coefficients ranged from .16 to .57, indicating no need for removing items. For the four skill groups Cronbach's alpha ranged from .65 to .87.

3.4.2. First grade

In first grade, the internal consistency was excellent for the total score (α = .91). The corrected item-total correlation coefficients ranged from .15 to .55, indicating no need for removing items. For the four skill groups Cronbach's alpha ranged from .70 to .86.

3.4.3. Second grade

In second grade, the internal consistency was excellent for the total score (α = .95). The corrected item-total correlation coefficients ranged from .13 to .66, indicating no need for removing items. For the four skill groups Cronbach's alpha ranged from .75 to .94. Cronbach's alphas are presented in Table 2.

4. Discussion

This study aimed to investigate the reliability and validity evidence of the early numeracy test for identifying children at risk for mathematical learning difficulties in Swedish-speaking kindergartens and schools in Finland. The EN-test was found to have adequate structural validity, known-group validity, cross-cultural validity, and internal consistency. The empirical data support the aim of measuring four early numeracy skills described in the model (Aunio & Räsänen, 2015): symbolic and non-symbolic number knowledge, understanding mathematical relations, counting skills, and basic skills in arithmetic in the three age groups in kindergarten, first grade and second grade.

Describing and interpreting sub-skills as separate factors, gives researchers and educators possibilities to more specifically examine areas of strength and weakness in children's performance, which is a prerequisite for planning and conducting targeted instructions (Purpura & Lonigan, 2015). Different sub-skills in the EN-test provides a more valuable spectrum of information about how well children perform than what can be offered in more narrower scales with an unidimensional approach (e.g., MARKO-D test, Ricken et al., 2013; CMA, Starkey et al., 2004). Some tests originally thought to be unidimensional have also been found to have an underlying multi-factorial structure. Aunio, Hautamäki, Heiskari, and Van Luit, (2006) examined the factor structure of the Finnish ENT-test and found support for a two-factor structure. In contrast, Ryoo et al. (2015) found support for a six-factor structure in TEMA-3. It is also in line with frameworks that describe the conceptual structure of early numeracy and mathematical development in early school years using a multi-factorial approach (e.g. Krajewski & Schneider, 2009; Sarama & Clements, 2009; Steffe, 1992; Wright, Martland, & Stafford, 2006). Furthermore, longitudinal (Aunio & Niemivirta, 2010; Geary et al., 2018; Jordan, Resnick, Rodrigues, Hansen, & Dyson, 2017) and cross-cultural studies (Aunio et al., 2006; Aunio, Korhonen, Bashash, & Khoshbakht, 2014; Cankaya & LeFevre, 2016; Rodic et al., 2015) have shown that different sub-skills of early numeracy are differentially related to later learning of mathematical skills.

The known-group validity confirmed measurement invariance for the configural and scalar models across gender and age groups, suggesting that the EN-test is comparable and measures the same constructs across girls and boys, and younger and older children within grade. In the current study, the factor structure was the same in kindergarten, first grade and second grade. However, the correlations were higher in kindergarten than in first and second grade, indicating that the skills become more separated in higher grades. A change in the strength of associations between factors reflects differences in early numeracy development (Ryoo et al., 2015). Children's understanding of numeracy is expected to change with age and experience, and it is not only the age when the children start their formal schooling that is of importance, but also variation in the early learning environment affects children's early numeracy skills (Cankaya & LeFevre, 2016; Lee & Aunio, 2018).

The cross-cultural validity display that the EN-test works equally well in the two language groups, supporting the aim of the EN-test development, that is, to provide evidence-based assessment for Swedish and Finnish kindergarten and schools in Finland to identify children with mathematical learning difficulties. When exploring the conceptual structure of early numeracy in studies a multi-factorial structure has been supported in diverse, international samples (e.g. Iran, Aunio et al., 2014; Norway, Lopez-Pedersen, Mononen, Korhonen, Aunio, & Melby-Lervåg, 2020; South Africa, Aunio et al., 2019). The EN-test is based on a developmental model of early numeracy skills, and not on the curriculum. Additionally, the cross-cultural validity evidence indicates that the EN-test could also be used in other Nordic countries or Estonia, countries with similar cultural, language and educational context. However, when adapting a test to a different cultural, language and educational context, a validation of the test is needed, as many aspects affect the development of children's early numeracy skills. More studies from very different types of educational cultures with a different starting-age of school going are needed for better understanding the relationships between the test-structures and skills-structures.

The Cronbach's alpha values indicated good internal consistency with this specific sample, in line with the reviewed tests that had published levels for internal consistency (REMA: Clements et al., 2008; Number sets test: Weiland et al., 2012; TEMA-3: Ginsburg & Baroody, 2003; NKT: Gersten, Clarke, & Jordan, 2007; NSS™: Jordan et al., 2012; ENT-test: Aunio, Hautamäki, Heiskari, & vain Luit,

2006).

Although our findings represent important additional research on the assessment of children's early numeracy skills to identify children at risk of mathematical learning difficulties, some limitations must be noted. First, even though the empirical data supported the structure of the four distinct factors (NK, MR, CS and BA) across the three age groups, the factors were highly related. In the EN-test, symbolic and non-symbolic number knowledge mainly included tasks measuring understanding of magnitudes, represented with number symbols, and similar skills were required in the tasks in both mathematical relations (number knowledge) and counting skills (number identification, recognition and writing, and number sequence). This similarity in the tasks and the conceptually overlapping represent concerns when trying to separate the different items into sub-skills and may be seen in the high correlation between factors. Secondly, the psychometric properties of the EN-test could have been strengthened with additional validity and reliability testing. As this was a cross-sectional study, we were not able to investigate the test-retest reliability. It was not possible to test the concurrent validity of the EN-test by using another early numeracy test as no such test was available in Swedish for this particular group of children. Inter-rater reliability was not tested, but the test is constructed to be easy to administer and objectively scored, as detailed information is provided in the test handbooks. A longitudinal study would have provided the opportunity to test the predictive validity and give information regarding how the test works overtime (Gersten et al., 2011).

Numerical skills are important in everyday life and poor achievement can have educational and vocational consequences (Jordan et al., 2015; Korhonen, Linnanmäki, & Aunio, 2014). Children who do not develop foundational numerical skills in their early school years are at risk of encountering mathematical difficulties (Clements & Sarama, 2014; Jordan et al., 2015). Identifying children at risk for learning difficulties is the first step in supporting them (Penner, Buckland, & Moes, 2019). The results of our study implied that all four core numerical skills could be assessed with the EN-test. We agree with Clements and Sarama (2014) and Desoete, Ceulemans, Roeyers, and Huylebroeck, (2009), both of which indicated that educators need evidence-based guidance to improve their knowledge regarding what to focus on in mathematics, which is especially important in preventing mathematical learning difficulties.

The EN-test makes it possible to simply and efficiently assess children's early numeracy skills. The results can be interpreted both with the overall test score and based on the four core skills. The EN-test was found to be reliable and valid for identifying children at risk for mathematical learning difficulties in kindergarten, first grade, and second grade.

Funding

This research was supported by a grant from the Högskolestiftelsen i Österbotten and the Swedish Cultural Foundation in Finland (12/3632). The data collection was administered during the LukiMat-project, funded by the Ministry of Education and Culture in Finland. The first author was supported by a research grant from The Board of the Graduate School at Åbo Akademi University.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijer.2020.101580.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.) (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Aunio, P. (2019). Early numeracy skills learning and learning difficulties—Evidence-based assessment and interventions. In (1st edition). D. Geary, D. Berch, & K. M. Koepke (Vol. Eds.), *Cognitive foundations for improving mathematical learning: Vol. 5*, (pp. 195–214). . https://doi.org/10.1016/B978-0-12-815952-1.00008-6.

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy skills. *Learning and Individual Differences, 20*, 427–435. https://doi.org/10.1016/j.lindif.2010.06.003.

Aunio, P., & Räsänen, P. (2015). Core numerical skills for learning arithmetic in children aged five to eight years – A working model for educators. *European Early Childhood Education Research Journal, 24*(5), 684–704. https://doi.org/10.1080/1350293X.2014.996424.

Aunio, P., Hautamäki, J., Heiskari, P., & Van Luit, J. E. H. (2006). The early numeracy test in finnish: Children's norms. *Scandinavian Journal of Psychology, 47*(5), 369–378. https://doi.org/10.1111/j.1467-9450.2006.00538.x.

Aunio, P., Korhonen, J., Bashash, L., & Khoshbakht, F. (2014). Children's early numeracy in Finland and Iran. *International Journal of Early Years Education, 22*(4), 423–440. https://doi.org/10.1080/09669760.2014.988208.

Aunio, P., Korhonen, J., Ragpot, L., Törmänen, M., Mononen, R., & Henning, E. (2019). Multi-factorial approach to early numeracy — The effects of cognitive skills, language factors and kindergarten attendance on early numeracy performance of South African first graders. *International Journal of Educational Research, 97*, 65–76. https://doi.org/10.1016/j.ijer.2019.06.011.

Aunola, K., Leskinen, E., & Nurmi, J. E. (2006). Developmental dynamics between mathematical performance, task motivation, and teachers' goals during the transition to primary school. *The British Journal of Educational Psychology, 76*(Pt 1), 21–40. https://doi.org/10.1348/000709905X51608.

Cankaya, O., & LeFevre, J. A. (2016). The home numeracy environment: What do cross-cultural comparisons tell us about how to scaffold young children's mathematical skills? In B. Blevins-Knabe, & A. Austin (Eds.). *Early childhood mathematics skill development in the home environment*https://doi.org/10.1007/978-3-319-43974-7_6.

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

Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math: The learning trajectories approach.* New York, NY: Routledge.

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. https://doi.org/10.1080/01443410701777272.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303.

Desoete, A., Ceulemans, A., Roeyers, H., & Huylebroeck, A. (2009). Subitizing or counting as possible screening variables for learning disabilities in mathematics education or learning? *Educational Research Review, 4*(1), 55–66. https://doi.org/10.1016/j.edurev.2008.11.003.

Fuchs, L. S. (2016). Curriculum-based measurement as the emerging alternative: Three decades later. *Learning Disabilities Research and Practice, 32*(1), 5–7. https://doi.

org/10.1111/ldrp.12127.
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. https://doi.org/10.1177/0734282908330592.
Geary, D. C., vanMarle, K., Chu, F. W., Rouder, J., Hoard, M. K., & Nugent, L. (2018). Early conceptual understanding of cardinality predicts superior school-entry number-system knowledge. *Psychological Science, 29*(2), 191–205. https://doi.org/10.1177/0956797617729817.
Gersten, R., Clarke, B., & Jordan, N. (2007). *Screening for mathematics difficulties in K-3 students.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.
Gersten, R., Clarke, B., Haymond, K., & Jordan, N. (2011). *Screening for mathematics difficulties in K-3 students* (2nd edition). Portsmouth, NH: RMC Research Corporation, Center on Instruction.
Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early math achievement* (3rd ed.). Austin, TX: Pro-Ed.
Gregory, R. J. (2015). *Psychological testing: History, principles and applications* (7th edition). Harlow: Pearson.
Jordan, N. C., Glutting, J., & Dyson, N. (2012). *Number Sense Screener™ (NSS™). User´s guide, K-1* (research edition). Baltimore, MD: Brookes.
Jordan, N. C., Fuchs, L. S., & Dyson, N. (2015). Early number competencies and mathematical learning: Individual variation, screening, and intervention. In R. C. Kadosh, & A. Dowker (Eds.). *The Oxford handbook of numerical cognition* (pp. 1079–1098). . https://doi.org/10.1093/oxfordhb/9780199642342.013.010.
Jordan, N. C., Resnick, I., Rodrigues, J., Hansen, N., & Dyson, N. (2017). Delaware longitudinal study of fraction learning: Implications for helping children with mathematics difficulties. *Journal of Learning Disabilities, 50*(6), 621–630. https://doi.org/10.1177/0022219416662033.
Koponen, T., Salminen, J., Aunio, P., Polet, J., & Hellstrand, H. (2011a). *LukiMat - Bedömning av lärandet: Identifiering av stödbehov i matematik i förskola. Handbok [LukiMat – Assessment for Learning: Identifying Children in Need of Support in Mathematics in Kindergarten. Handbook].* http://www.lukimat.fi/lukimat-bedomning-av-larandet/material/identifiering-av-stodbehov/forskola/f-mat-handbok.
Koponen, T., Salminen, J., Aunio, P., Polet, J., & Hellstrand, H. (2011b). *LukiMat - Bedömning av lärandet: Identifiering av stödbehov i matematik i årskurs 1. Handbok [LukiMat – Assessment for Learning: Identifying Children in Need of Support in Mathematics in First Grade. Handbook].* http://www.lukimat.fi/lukimat-bedomning-av-larandet/material/identifiering-av-stodbehov/ak-1/1-mat-handbok.
Koponen, T., Salminen, J., Aunio, P., Polet, J., & Hellstrand, H. (2011c). *LukiMat - Bedömning av lärandet: Identifiering av stödbehov i matematik i årskurs 2. Handbok [LukiMat – Assessment for Learning: Identifying Children in Need of Support in Mathematics in Second Grade. Handbook].* http://www.lukimat.fi/lukimat-bedomning-av-larandet/material/identifiering-av-stodbehov/ak-2/2-mat-handbok.
Korhonen, J., Linnanmäki, K., & Aunio, P. (2014). Learning difficulties, academic well-being and educational dropout: A person-centred approach. *Learning and Individual Differences, 31*, 1–10. https://doi.org/10.1016/j.lindif.2013.12.011.
Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction, 19*(6), 513–526. https://doi.org/10.1016/j.learninstruc.2008.10.002.
Lee, K., & Aunio, P. (2018). Individual differences in math achievement: Finland and Singapore. In J. W. Vasbinder, & B. Gulyás (Vol. Eds.), *Cultural patterns and neurocognitive circuits II: East-west connections: Vol. 5*, (pp. 169–186). . https://doi.org/10.1142/9789813230484_0008.
Lopez-Pedersen, A., Mononen, R., Korhonen, J., Aunio, P., & Melby-Lervåg, M. (2020). Validation of an early numeracy screener for first graders. *Scandinavian Journal of Educational Research.* https://doi.org/10.1080/00313831.2019.1705901.
OECD (2016). *Socio-economic status, student performance and students' attitudes towards science. PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishinghttps://doi.org/10.1787/9789264266490-10-en.
Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development, 61*, 27–58. https://doi.org/10.1111/j.1540–5834.tb1996.00536.x.
Penner, M., Buckland, C., & Moes, M. (2019). Early identification of, and interventions for, kindergarten students at risk for mathematics difficulties. In K. M. Robinson, H. P. Osansa, & D. Kotsopoulus (Eds.). *Mathematical learning and cognition in early childhood* (pp. 57–78). . https://doi.org/10.1007/978-3-030-12895-1_5.
Purpura, D. J., & Lonigan, C. J. (2015). Early numeracy assessment: The development of the preschool early numeracy scales. *Early Education and Development, 26*(2), 286–313. https://doi.org/10.1080/10409289.2015.991084.
Ricken, G., Fritz, A., & Balzer, L. (2013). *MARKO-D – Matematik und Rechnen Test zur Erfassung von Konzepten im Vorshualter [MARKO-D- mathematics and arithmetic test for assessing concepts at preschool age].* Göttingen, Germany: Hogrefe.
Rodic, M., Zhou, X., Tikhomirova, T., Wei, W., Malykh, S., Ismatulina, V., et al. (2015). Cross-cultural investigation into cognitive underpinnings of individual differences in early arithmetic. *Developmental Science, 18*(1), 165–174. https://doi.org/10.1111/desc.12204.
Ryoo, J. H., Molfese, V. J., Brown, E. T., Karp, K. S., Welch, G. W., & Bovaird, J. A. (2015). Examining factor structures on the Test of Early Mathematics Ability—3: A longitudinal approach. *Learning and Individual Differences, 41*, 21–29. https://doi.org/10.1016/j.lindif.2015.06.003.
Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children.* Routledge.
Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly, 19*, 99–120. https://doi.org/10.1016/j.ecresq.2004.01.002.
Steffe, L. P. (1992). Schemes of action and operation involving composite units. *Learning and Individual Differences, 4*, 259–309.
Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., De Vet, H. C. W., Westerman, M. J., Patrick, D. L., et al. (2017). COSMIN standards and criteria for evaluating the content validity of health-related Patient-Reported Outcome Measures: A Delphi study. *Quality of Life Research, 27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0.
Van Luit, J. E. H., Van de Rijt, B. A. M., & Pennings, A. H. (1994). *Utrechtse Getalbegrip Toets [Utrecht Early Mathematical Competence Scales].* Doetinchem, the Netherlands: Graviant.
Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Hirokazu, Y. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten and kindergarten mathematics measure. *An International Journal of Experimental Educational Psychology, 32*(3), 311–333. https://doi.org/10.1080/01443410.2011.654190.
Wright, R. J., Martland, J., & Stafford, A. K. (2006). *Early Numeracy. Assessment for teaching & intervention.* London: SAGE Publications.