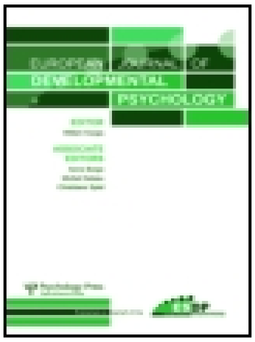


European Journal of Developmental Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pedp20

Psychometric properties of the Ani Banani Math Test

Dieuwer ten Braak & Ingunn Størksen

To cite this article: Dieuwer ten Braak & Ingunn Størksen (2021): Psychometric properties of the Ani Banani Math Test, European Journal of Developmental Psychology, DOI: 10.1080/17405629.2021.1879046

To link to this article: https://doi.org/10.1080/17405629.2021.1879046

Published online: 08 Feb 2021.

EUROPEAN JOURNAL OF DEVELOPMENTAL PSYCHOLOGY
https://doi.org/10.1080/17405629.2021.1879046

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

Psychometric properties of the Ani Banani Math Test

Dieuwer ten Braak and Ingunn Størksen

Norwegian Centre for Learning Environment and Behavioural Research in Education, University of Stavanger, Stavanger, Norway

ABSTRACT

This study assessed the psychometric properties of a digital early mathematics assessment, the Ani Banani Math Test (ABMT) in three samples ($N_1 = 243$, $N_2 = 691$, $N_3 = 1282$) in kindergarten and first grade (age range 4.67– 7.30). Confirmatory factor analyses showed that the ABMT appears to measure one general construct of children's informal mathematics development. MIMIC models showed that all items functioned similarly across age, socioeconomic status, and sex, except for two items which showed a bias towards boys and one item towards girls. The correlations with other mathematics assessments were strong, and significantly stronger than with executive function, working memory, and literacy measures, indicating concurrent and discriminant validity. The task was highly correlated with mathematics achievement five years later, indicating high predictive validity. Taken together, the ABMT appears to be a psychometrically valid research measure of children's overall early mathematics skills; however, caution should be taken when comparing mean scores for boys and girls.

ARTICLE HISTORY Received 27 September 2019; Accepted 15 January 2021

KEYWORDS Early childhood; mathematics; numeracy; geometry; play

Introduction

Early mathematics ability is a strong, if not the strongest, predictor of later mathematics (Duncan et al., 2007). This is likely due to its cumulative development (Sarama & Clements, 2009) where informal mathematics skills provide the basis for the development of formal mathematics (Purpura et al., 2013). It is therefore important to adequately assess and study these skills early in life. In this short report, we present the psychometric properties of a digital early mathematics assessment that was developed for use in kindergarten-aged children in play-based early childhood and care (ECEC) in Norway; the Ani Banani Math Test (ABMT; Størksen & Mosvold, 2013).

CONTACT Dieuwer ten Braak dieuwer.t.braak@uis.no Norwegian Centre for Learning Environment and Behavioural Research in Education, University of Stavanger, Postboks 8600 Forus 4036 Stavanger, Norway

The ABMT is founded on research (Clements et al., 2008) as well as theoretical and practical recommendations related to early childhood mathematics (Magne, 2003; Sarama & Clements, 2009). It was developed to reflect three areas of children's informal maths skills that are central to Norwegian ECEC – numeracy, geometry, and problem-solving (Magne, 2003). These areas are covered in the Norwegian Framework Plan and are considered important aspects of informal mathematics (Magne, 2003; Ministry of Education and Research, 2011, 2017). Empirically, there is strong evidence for the theory that mathematics skills develop cumulatively (Sarama & Clements, 2009), with early informal skills laying the foundation for the acquisition of later skills (Purpura et al., 2013). Informal numeracy skills are traditionally considered those skills that develop before formal schooling, through day-to-day situations and play and include numbering (e.g., counting), relations (e.g., set comparison), and simple arithmetic operations (e.g., addition with objects) (Purpura et al., 2013). Geometry (e.g., identifying shapes) is another important area of early mathematics (Sarama & Clements, 2009). Some argue that geometry is a gateway to certain higher-order mathematical and logical reasoning skills (Tatsuoka et al., 2004) and should receive more attention in early education (Clements & Sarama, 2011). Finally, problem-solving is a mathematical process interwoven with mathematical content and is also considered an important aspect of early childhood mathematics (Clements et al., 2003).

The ABMT differs from other maths assessments (e.g., Research-Based Early Maths Assessment, Clements et al., 2008; Utrecht Early Numeracy Test, Van Luit & Van de Rijt, 2009; Woodcock–Johnson Tests of Achievement, Woodcock et al., 2001) in several ways. First, the items are framed in a playful story-context. Child assessment in Norwegian ECEC is unusual and controversial. The playful and child-friendly design of the ABMT reduces scepticism in children, teachers, and parents. Second, the ABMT is designed for tablet use, which has several advantages over traditional paper-and-pencil assessments. Play is important in ECEC mathematics education (Hirsh-Pasek et al., 2009; Sarama & Clements, 2009) and tablets provide unique affordances for playful assessment (Lange & Meaney, 2013). The touch screen interface of tablets takes advantage of direct mediation through finger moves and gestures allowing children to produce and transform objects directly (Sinclair & SedaghatJou, 2013). Further, the use of technology enhances children's motivation in educational settings (Couse & Chen, 2010). The tablet automatically encodes results and sends encrypted data to a server which makes data collection time efficient. Finally, other tasks require considerable time to administer. The ABMT takes only 10 minutes to complete, thereby limiting test fatigue.

Even though the ABMT has been used in various research projects (e.g., Skoleklar, Agderprojekt, Lekbasert Læring), its factor structure, item functioning, and validity have not yet been investigated. It is unknown whether a single factor best represents the ABMT or whether the three aspects of numeracy, geometry, and problem-solving are reflected in separate factors. Moreover, it is necessary to determine how well scores on the ABMT generalize across children of different ages, sex, or socioeconomic status (SES). Test bias is serious as it may result in certain children scoring differently on the ABMT, despite the same level of true maths ability. Finally, a key factor in establishing the validity of the ABMT test scores is to provide evidence that the ABMT is a strong predictor of other maths assessments, but also shows discriminant validity with other constructs that are known to relate to mathematics, for example, executive function and working memory (e.g., Bull & Lee, 2014), phonological awareness (e.g., Simmons & Singleton, 2008), and vocabulary (e.g., Purpura et al., 2017).

The following research questions were asked:

(RQ1) What is the factor structure of the ABMT? Does a one, two, or three-factor model best fit the data? Although the task was developed to include aspects of three mathematical areas (problem-solving, geometry, and numeracy), we expected that the overlap in content across the items might not yield clearly defined factors.

(RQ2) Do the items of the ABMT function similarly across age, sex, and SES? This question was evaluated in an exploratory fashion.

(RQ3) Does the ABMT show concurrent, predictive, and discriminant validity? We expected the ABMT to relate more strongly to other mathematical assessments compared to related constructs (i.e., executive function, working memory, vocabulary, phonological awareness).

Method

Samples and procedures

This study is based on secondary analyses of three existing datasets. All samples were collected from different schools in both rural and urban areas of southern Norway. For Sample 1, data were derived from children

Table 1. Summary of the Sample Characteristics for Sample 1, Sample 2, and Sample 3.

	N	Number of ECEC/ schools	Missing Count/Percent	Male/female	Mean age (*SD*)	Age range
Sample 1						
Spring K	241	19	2/0.8%	122/119	5.78 (0.29)	5.29– 6.30
Spring G1	239	8	4/1.6%	122/117	6.78 (0.29)[b]	6.29– 7.30[b]
Fall G5[c]	160	8	83/34.2%	74/86	10.29 (0.29)[b]	9.79– 10.78[b]
Sample 2						
Fall K	664	71	27/3.9%	332/332	5.16 (0.26)	4.67– 5.67
Spring K	292[a]	35	20[a]/6.4%	141/151	5.99 (0.27)	5.50– 6.42
Sample 3						
Fall K	1199	96	83/6.5%	606/593	5.14 (0.28)	4.67– 5.67
Spring K	519[a]	47	75[a]/12.6%	259/260	5.93 (0.28)	5.42– 6.42

Note. K = kindergarten, G1 = 1st grade, G5 = 5th grade. [a] Data from Sample 2 and 3 in spring kindergarten only contains half of the sample at fall because data from the intervention group was excluded. [b] Approximation; no data on exact date of testing available, time-window was approximately 1 month. [c] No ABMT data, only data from national school assessments.

who participated in the research project 'Skoleklar' where all children from one municipality in the last year of ECEC (referred to as 'kindergarten') were invited to participate. The total sample[1] included 243 children (5.3% immigrant status[2]) who were assessed in the spring of the last year (2012; M_{age} = 5.78) of ECEC and in the spring of first grade (2013). In fall of fifth grade (2016) data from the national school assessments were collected from 160 children. For samples 2 and 3, data were derived from two cluster-randomized controlled trial intervention studies ('Agderprojekt' and 'Lekbasert Læring', respectively) where all ECECs in participating municipalities in southern Norway were invited to participate. In Sample 2, a total of 691 children (8.1% immigrant status) were assessed in fall (2016; M_{age} = 5.16) and spring (2017) of kindergarten. In Sample 3, a total of 1282 children (16.1% immigrant status) were assessed in fall (2017; M_{age} = 5.14) and spring (2018) of kindergarten. In the spring of kindergarten (post-test), solely data from the control groups were used. More details about the samples are given in Table 1.

Children were assessed individually by a trained research assistant in a single session. Testing took place within 1 month in Sample 1, and within 2 weeks in samples 2 and 3. Tasks were administered on a tablet. To ensure high-quality and uniform assessment, assistants were trained for 2 days focusing on child-friendly assessment, procedures to eliminate bias, and technical operation of the tablet. National school assessments

[1]The total sample includes children who participated in at least one wave of data collection.
[2]A child for whom both parents were born outside Scandinavia was coded as having an immigrant status.

were conducted by the teachers. All studies were approved by the Norwegian Social Science Data Service.

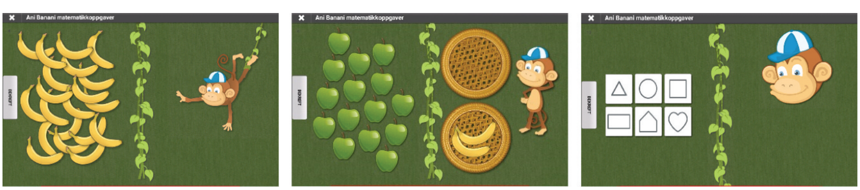
Measures

Ani Banani Math Test

The 18 items in the ABMT are embedded in playful contexts, including a monkey ('Ani Banani') and his everyday activities. Test items were constructed to embed elements of numeracy, geometry, and problem-solving. See Figure 1 for examples. To optimize the items, the task has been piloted in a previous study with 40 children. During this process, items that proved too difficult or otherwise did not work were adjusted to match the children's development (Størksen & Mosvold, 2013). See Table 4 for an overview of all items.

Preschool early numeracy scale

This task (PENS; Purpura et al., 2015), α =.90, is a brief early numeracy measure developed in the United States. It includes 24 non-interactive items regarding one-to-one counting, cardinality, counting subsets, subitizing, number comparison, set comparison, number order, numeral identification, set-to-numerals, story problems, number combinations, and verbal counting.



a. b. c.

Figure 1. Illustrations of three ABMT items for numeracy (a), problem-solving (b), and geometry (c). (a) 'Ani Banani is a little hungry today. Can you give him five bananas?' (b) 'Ani Banani has some bananas in one of his baskets. He would like twice as many apples as he has bananas. Can you give him twice as many apples in the other basket?' (c) 'Ani Banani likes to look at different shapes. Can you help him find the triangle?'.

National school maths and reading assessments

In first and fifth grades, mathematics and reading achievement were assessed with the national school assessments (NDET; Norwegian Directorate for Education and Training). Reliability estimates for the tests are all α > .8 (Norwegian Directorate for Education and Training, 2013, 2018).

Vocabulary

In Sample 1, vocabulary was assessed by the Norwegian Vocabulary Task, α = .84 (NVT; Størksen et al., 2013). Children were presented with 45 different pictures on a tablet screen and had to tell the name of the object depicted. In Sample 3, a shorter version of this task was used (20 items), α = .84.

Phonological awareness

Twelve-item blending task (NDET), α = .75 (Solheim et al., 2012). The target word was auditory presented in phonemes and children had to indicate the corresponding alternative from four images on a tablet screen.

Executive function

The Head-Toes-Knees-Shoulders task, α =.94 (HTKS; McClelland et al., 2014) is a measure of children's cognitive flexibility, working memory, and inhibition. Children were initially habituated to two different rules (touch your head/toes) and later had to inhibit this automatized response and replace it with the opposite (e.g., head = toes) and a different rule (e.g., head = knees).

Working memory

The Forward/Backward Digit Span subtest from the WISC-IV (Wechsler, 2003), α =.87 (Williams et al., 2003), was used as a measure of working memory. Children had to repeat a sequence of digits in the same and reversed order.

Covariates

Parents reported their highest obtained education, child sex, and age on a questionnaire. Maternal education (1 = junior high school, 2 = senior high school, 3 = 1– 2 years of college/university, 4 = 3 years of college/university, 5 > 3 years of college/university) was used as a proxy for SES.

Analyses

Missing data were generally low (see Table 1) except for school achievement data in fifth grade (Sample 1, RQ3) which had 34.2% missing data. Children with missing values in fifth grade had significantly lower mean scores on all predictors, except phonological awareness in kindergarten and EF in first grade. The proportion of children with an immigrant background, low SES, and boys was also higher in the 5th grade missing data group. For RQ3, the full information maximum likelihood estimator (FIML) was used. There were no systematic missing patterns in the other samples, except for immigrant status in Sample 3 where the proportion of immigrant children was slightly higher in children with missing values on the ABMT. For RQ1 and RQ2, the robust weighted least squares estimator (WLSMV) was used. This estimator is appropriate for categorical data (Muthén et al., 2015). All analyses were conducted with M*plus* Version 8 (Muthén & Muthén, 1998-2017). Raw scores were used for each measure (see Table 2).

Confirmatory factor analyses (CFA) were used to investigate the factor structure. Overall model fit was evaluated using the following criteria: RMSEA ≤ .06, CFI/TLI ≥ .95 (Hu & Bentler, 1999). Items 1 and 18 were considered fillers (99% of the children scored correctly) and omitted from all analyses. Residual covariances were included between items that were similar in wording and/or content and indicated areas of strain when not freely estimated (modification index >10.0). The number of residual covariances was kept as low as possible and kept similar across samples for consistency (see Table 4). Chi-square difference tests were performed using the DIFFTEST option in M*plus.*

To investigate whether the items functioned similarly across age, sex, and SES the three samples were merged, and multiple indicator multiple causes (MIMIC) models (Jöreskog & Goldberger, 1975) were estimated to assess uniform differential item functioning (DIF). The advantage of MIMIC over, e.g., multiple-groups CFA, is that continuous predictors can be accommodated which eliminates the need to impose categorical cut-

Table 2. Means and Standard Deviations for Observed Scores on the Assessments.

	Fall kindergarten		Spring kindergarten		Spring first grade		Fall fifth grade	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Sample 1								
ABMT			10.62	3.13	14.52	2.57		
HTKS			34.46	15.64	47.48	9.81		
Digit span			8.76	3.04	11.01	2.58		
Vocabulary			26.35	5.69	30.72	4.96		
PA			3.66	3.38	10.21	1.91		
NSMA[a]					42.76	7.43	50.88	9.72
NSRA[a]					96.12	10.16	49.89	9.91
Sample 2								
ABMT	7.00	2.85	9.94	3.12				
Sample 3								
ABMT	6.87	2.99	9.94	3.28				
Vocabulary[b]			12.58	4.28				
PENS			17.11	5.10				
Digit span			7.00	3.29				

Note. ABMT = Ani Banani Maths Test, HTKS = Head Toes Knees Shoulders, PA = Phonological awareness, NSMA = National School Maths Assessment, NSRA = National School Reading Assessment. PENS = Preschool Early Numeracy Screener. [a]The content and scale range of the NSMA and NSRA in first grade and fifth grade differ. [b] In Sample 3 a shorter version of the Norwegian Vocabulary Task was used.

offs (cf. MacCallum et al., 2002). DIF was assessed in an exploratory manner by regressing the latent ABMT variable on the covariate of interest and fixing all direct effects on the indicators to zero and then inspecting modification indices for any salient areas of strain (Brown, 2015).

Concurrent validity was determined by calculating the correlation between the observed score on the ABMT and other mathematical assessments at the same time point, whilst predictive validity was assessed by estimating the predictive power of ABMT for these assessments at a later time point. Discriminant validity was determined by testing the difference between two correlated correlations (dependent correlations obtained from the same sample) with one variable in common (Lee & Preacher, 2013) to show that the correlations between the ABMT and other maths assessments at concurrent time points were significantly higher than between the ABMT and related constructs. Only samples 1 and 3 were used for these analyses as Sample 2 did not contain other maths assessments.

Results

RQ1: factor structure

First, 1-factor models were estimated with the spring kindergarten data because the ABMT was originally designed for this age group. The models showed a good fit with the data (RMSEA ≤ .034, CFI ≥ .958, TLI ≥ .950) Next, 3-factor models (numeracy, geometry, and problem-solving) were estimated. Table 4 gives an overview of the constructs for each item. Although the 3-factor models showed an adequate fit (RMSEA ≤ .034, CFI ≥ .959, TLI ≥ .949), the correlation between the factors was often very high (r > .85), especially for numeracy and problem-solving, indicating poor discriminant validity between the latent dimensions (Brown, 2015). This model was therefore rejected and 2-factor models without a separate problem-solving factor were tested. The 2-factor models showed a good fit (RMSEA ≤ .034, CFI ≥ .961, TLI ≥ .953), and chi-square difference tests indicated a better fit of this model, compared to the 1-factor model. However, the correlation between the factors was still high (r = .704 – r = .961) and could not be classified as 'no problem' according to the confidence interval classification (CI_{CFA}(sys)) as described by Rönkkö and Cho (2020): upper limit < .8. The 1-factor model was therefore considered the best option. This solution showed a good fit at all time points as well as in the combined sample. Detailed results from the CFAs are presented in Table 3. In Table 4, the item parameters from spring kindergarten in the combined sample are presented.

RQ2: item functioning

MIMIC models in the combined sample indicated that age and SES were positive predictors of the latent factor (β = .241, p < .001 and β = .237, p < .001, respectively). None of the items showed DIF for these two covariates.[3] DIF was found for sex: items 3 and 17 indicating a bias towards boys, and item 11 towards girls. See Table 5. Sex was a non-significant predictor of the latent factor (β = .056, p = .134).

RQ3: validity

Zero-order correlations between the observed score on the ABMT and other related constructs in samples 1 (N = 233) and 3 (n = 518) are presented in Table 6. The correlations between the ABMT and other maths measures at the same time point were all strong (r > .50). The ABMT also showed strong correlations (r > .50) with maths measures at a later time-point. As shown in Table 7, a test of correlated correlations showed that the concurrent correlations between the ABMT and other

[3] Results were similar when SES was coded as dummy variables

Table 3. Fit Indices for Confirmatory Factor Models on the ABMT Scores in All Samples.

	RMSEA	CFI	TLI	WRMR	χ^2 (*df*)	*p*	*r* between factors	95% CI *r*	χ^2_{diff} (*df* = 1)	*p*
Sample 1										
Spring kindergarten (*N*= 241)										
1-factor model	.034	.958	.950	.880	129.152 (101)	.0309	-			
2-factor model	.033	.961	.953	.870	126.679 (100)	.0370	.848 (N-G)	0.679– 1.017	2.407	.1208
3-factor model	.034	.959	.949	.866	125.861 (98)	.0304	1.263 (N-PS)[b]			
							.985 (G-PS)			
							.858 (N-G)			
Spring first grade (*N*= 239)										
1-factor model[a]	.019	.980	.976	.780	94.545 (87)	.2721	-			
2-factor model[a]	.020	.978	.974	.779	94.320 (86)	.2529	.961 (N-G)	0.76– 1.162	.137	.7109
3-factor model[a]	.020	.980	.975	.767	91.78 (84)	.2631	.813 (N-PS)			
							.744 (G-PS)			
							.985 (N-G)			
Sample 2										
Fall kindergarten (*N*= 664)										
1-factor model	.025	.975	.971	.930	141.263 (101)	.0051	-			
2-factor model	.020	.983	.980	.873	127.717 (100)	.0322	.789 (N-G)	0.669– 0.910	9.170	.0025
3-factor model	.021	.982	.978	.869	126.778 (98)	.0268	1.197 (N-PS)[b]			
							1.061 (G-PS)[b]			
							.780 (N-G)			
Spring kindergarten (*n*= 292)[c]										
1-factor model	.032	.972	.967	.886	131.313 (101)	.0230	-			
2-factor model	.024	.984	.981	.823	116.932 (100)	.1186	.704 (N-G)	0.554– 0.853	10.335	.0013
3-factor model	.025	.984	.980	.821	115.845 (98)	.1054	.899 (N-PS)			
							.622 (G-PS)			
							.708 (N-G)			
Sample 3										
Fall kindergarten (*N*= 1199)										
1-factor model	.034	.963	.957	1.214	243.591 (101)	< .001	-			
2-factor model	.034	.965	.958	1.193	237.370 (100)	< .001	.902 (N-G)	0.823– 0.982	5.079	.0242

(*Continued*)

Table 3. (Continued).

	RMSEA	CFI	TLI	WRMR	χ^2 (*df*)	*p*	*r* between factors	95% CI *r*	χ^2_{diff} (*df* = 1)	*p*
3-factor model	.033	.968	.961	1.152	223.236 (98)	< .001	.642 (N-PS) .799 (G-PS) .882 (N-G)			
Spring kindergarten (*n*= 519)[c]										
1-factor model	.033	.974	.969	.962	156.545 (101)	.0003				
2-factor model	.030	.978	.974	.923	145.980 (100)	.0019	.841 (N-G)	0.739– 0.944	7.790	.0053
3-factor model	.030	.978	.973	.914	143.839 (98)	.0018	.850 (N-PS) .771 (G-PS) .831 (N-G)			
Combined sample										
Spring kindergarten (*N*= 1052)										
1-factor model	.033	.970	.965	1.145	219.858 (101)	< .001	-			
2-factor model	.030	.977	.972	1.063	191.990 (100)	< .001	.804 (N-G)	0.728– 0.881	19.846	< .001
3-factor model	.030	.976	.971	1.062	192.077 (98)	< .001	.950 (N-PS) .758 (G-PS) .807 (N-G)			

Note. The χ^2 difference test was performed on the 1- and 2-factor models with the DIFFTEST option in Mplus Version 8. [a] In first grade, item 4 was omitted because it resulted in empty cells. [b] The latent variable covariance matrix (psi) was not positive definite for this model due to a correlation greater than one between two latent factors and results for this model should not be interpreted. [c] Data from Samples 2 and 3 in spring kindergarten only contains half of the sample at fall because data from the intervention group was excluded. N = numeracy, G = geometry, PS = problem-solving. WRMR = weighted root mean square residual; CI = confidence interval.

Table 4. Intended Constructs for the Multiple Factor Models and Psychometric Information from the 1-Factor Model in the Spring Kindergarten in the Combined Sample.

No.	Construct	Item detail	% correct	Loadings 1-factor solution	Thresholds
1	N/A (filler)	Select biggest milkshake	99.1	N/A	N/A
18	N/A (filler)	Divide equal number of apples	98.5	N/A	N/A
4	Numeracy	Give 5 banana's	89.3	0.707	−1.240
9	Numeracy	Put 5 plates on the table	84.9	0.627	−1.032
12[a]	Geometry	Find triangle	83.2	0.453	−0.960
13[a]	Geometry	Find triangle again	81.0	0.424	−0.879
8	Numeracy	How many bricks	64.4	0.441	−0.368
7	Numeracy	How many cars	60.8	0.314	−0.275
14	Numeracy/Problem-solving *	Select monkey with most marbles	58.4	0.443	−0.211
15[b]	Geometry	Copy a pattern	57.3	0.622	−0.185
5	Numeracy	Give 17 banana's	45.4	0.645	0.115
10	Numeracy	Put more plates so there's place for 7	38.7	0.617	0.286
16[b]	Geometry	Copy a pattern	37.5	0.556	0.320
11	Geometry	Complete a puzzle	35.1	0.451	0.383
3 [c]	Numeracy	Count to 50	27.8	0.741	0.589
2	Numeracy/Problem-solving *	Select next smallest milkshake	22.4	0.449	0.758
17 [c]	Numeracy	Count backwards from 15	14.6	0.664	1.052
6	Numeracy/Problem-solving *	Give twice as many apples	11.4	0.502	1.204

Note. [abc] freely estimated correlations between error terms. * Numeracy in the 2-factor solution and problem-solving in the 3-factor solution

maths assessments were all significantly higher than the correlations between the ABMT and all other constructs.

Discussion

This study assessed the psychometric properties of the ABMT (Størksen & Mosvold, 2013). Although the ABMT did seem to tap at least two highly related aspects of children's early mathematics skills: numeracy and geometry, the measure failed to reliably distinguish between these mathematics areas. Rather, the well-fitting and most parsimony 1-factor model suggests that the ABMT measures one general construct of children's informal mathematics development. In future versions of the ABMT, items that more distinctively tap into the areas of numeracy, geometry, and problem-solving should be added. Item functioning appeared robust across age and SES. When planning to compare ABMT means between the sexes, it is advised to exclude items with signs of DIF as three items showed indications of bias in the present study. In future versions of the ABMT, these items should be considered for revision. ABMT scores

Table 5. Results from the MIMIC Models in the Combined Sample.

Covariate	RMSEA	CFI	TLI	WRMR	χ^2 (df)	*p*	Item nr.	M.I. >10	StdYX E.P.C.
Age	.032	.968	.962	1.144	239.268 (116)	< .001	N/A	-	N/A
Maternal education	.033	.964	.958	1.157	243.002 (116)	< .001	N/A	-	N/A
Sex (female = 1, male = 0)	.040	.952	.944	1.305	311.534 (116)	< .001	11	24.509	.190
Item 11 free	.038	.958	.950	1.250	287.228 (115)	< .001	17	20.000	−.211
Item 11, 17 free	.036	.963	.956	1.203	267.194 (114)	< .001	3	24.728	−.208
Item 11, 17, 3 free	.033	.969	.962	1.143	242.156 (113)	< .001	N/A	-	N/A

Note. Paths from the covariate to the indicator were freed one by one. M.I. = Modification Indices; StdYX E.P.C. = standardized expected parameter change.

Table 6. Zero-Order Correlations between the Observed ABMT Scores and Other Measures in Samples 1 and 3.

			T1					T2							G5		T2	
	Measure	*N*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	PENS	*N*
1.	ABMT T1	241	-							.708		.521	.453				.529	560
2.	HTKS T1	241	.477	-														
3.	Digit span T1	241	.597	.516	-													
4.	Vocabulary T1	241	.456	.328	.468	-												
5.	PA T1	240	.400	.381	.407	.408	-											
6.	NSMA T2	233	.587	.403	.434	.259	.211	-										
7.	NSRA T2	233	.484	.399	.435	.335	.293	.666	-									
8.	ABMT T2	239	.669	.480	.575	.395	.405	.685	.532	-		.545	.483				.648	519
9.	HTKS T2	239	.436	.380	.392	.316	.195	.547	.418	.454	-							
10.	Digit span T2	239	.500	.368	.647	.414	.353	.421	.437	.488	.357	-	.395				.559	518
11.	Vocabulary T2	239	.424	.306	.445	.816	.347	.224	.337	.394	.298	.417	-				.514	518
12.	PA T2	233	.344	.305	.364	.397	.397	.326	.649	.366	.311	.368	.350	-				
13.	NSMA G5	160	.601	.305	.542	.366	.263	.661	.479	.621	.467	.532	.345	.291	-			
14.	NSRA G5	159	.511	.337	.550	.479	.338	.516	.492	.521	.389	.479	.424	.314	.694	-		

Note. Correlations for Sample 1 are presented below the diagonal. Correlations for Sample 3 are presented above the diagonal. T1 = spring kindergarten in Sample 1 and fall kindergarten in Sample 3, T2 = first grade in Sample 1 and spring kindergarten in Sample 3, G5 = fifth grade. ABMT = Ani Banani Math Test, HTKS = Head Toes Knees Shoulders, PA = Phonological awareness, NSMA = National School Math Assessment, NSRA = National School Reading Assessment. PENS = Preschool Early Numeracy Screener. Sample size for PENS = 518.

Table 7. Results from the Test of Correlated Correlations between the Observed ABMT sand the National Math Assessment and the PENS versus other Measure scores.

	N	Measure	Correlation with ABMT(j)	r_{kh}	z-score
			r_{jk}		
Sample 1	233	NSMA(k)	.685		
			r_{jh}		
	233	NSRA(h)	.532	.666	3.83***
	239	HTKS(h)	.454	.547	4.88***
	239	Digit span(h)	.488	.421	3.80***
	239	Vocabulary(h)	.394	.224	4.74***
	233	PA(h)	.366	.326	5.46***
			r_{jk}		
Sample 3	518	PENS(k)	.648		
			r_{jh}		
	518	Digit span(h)	.545	.559	3.33***
	518	Vocabulary(h)	.483	.514	4.96***

Note. Sample 1 includes data from first grade. Sample 3 includes data from spring kindergarten. In Sample 1, N = 233 was used for the z-test. In Sample 3, N = 518 was used for the z-test. r_{jk} and r_{jh} = the two correlation coefficients to be compared, r_{kh} = correlation of the unshared variables, ABMT = Ani Banani Math Test, PA = Phonological awareness, NSMA = National School Math Assessment, NSRA = National School Reading Assessment. PENS = Preschool Early Numeracy Screener.
*** $p < .001$

showed good concurrent validity with the school assessments and the PENS (Purpura et al., 2015) and discriminant validity when contrasted with measures of executive function, working memory, phonological awareness, and vocabulary. The task showed good predictive validity for children's maths achievement five years later. Items showed moderate to high factor loadings indicating that children's maths skills explain a significant proportion of the variance in the level of underlying maths skills needed to score correctly on an item.

Limitations

At present, the ABMT is intended for use in research only and not as a diagnostic tool for teachers. Further studies are necessary to investigate whether the ABMT can be used to identify children at risk of mathematics difficulties. Also, available data varied by sample, and sample sizes varied across analyses. There was substantial missingness on the fifth-grade data in Sample 1 which was related to low performance on earlier skills, sex, SES, and immigrant status. This may have biassed some estimates.

Conclusions

Despite these limitations, and acknowledging the DIF for sex, the ABMT appears to be a psychometrically valid, playful research measure of kindergarten-aged children's early mathematics. The ABMT gives researchers a tool to assess children's early mathematics and investigate how these skills predict later outcomes. This can provide policymakers and teachers with crucial information on the importance of early mathematics in ECEC. Another practical application of the ABMT is in evaluations of ECEC interventions. Recent findings indicate that the ABMT is sensitive to intervention effects (Rege et al., 2019). The ABMT is likely to be adequate for use in ECEC outside of Norway as well, especially if children are situated in play-based ECEC, are not used to being in test-situations or when motivation problems are expected to arise.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Research Council of Norway [203326,237973].

ORCID

Dieuwer ten Braak http://orcid.org/0000-0003-3200-0198

References

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. The Guildford Press.

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*(1), 36–41. 10.1111/cdep.12059.

Clements, D. H., & Sarama, J. (2011). Early childhood teacher education: The case of geometry. *Journal of Mathematics Teacher Education*, *14*(2), 133–148. 10.1007/s10857-011-9173-0.

Clements, D. H., Sarama, J., DiBiase, A. E., & DiBiase, A.-M. (2003). *Engaging young children in mathematics: standards for early childhood mathematics education*. Routledge. http://ebookcentral.proquest.com/lib/uisbib/detail.action?docID=957204

Clements, D. H., Sarama, J. H., & Liu, X. F. 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. 10.1080/01443410701777272.

Couse, L. J., & Chen, D. W. (2010). A tablet computer for young children? exploring its viability for early childhood education. *Journal of Research on Technology in Education*, *43*(1), 75–98. 10.1080/15391523.2010.10782562.

Duncan, G. J., Claessens, A., Huston, A. C., Pagani, L. S., Engel, M., Sexton, H., Japel, C., Dowsett, C. J., Magnuson, K., Klebanov, P., Feinstein, L., Brooks-Gunn, J., & Duckworth, K. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. 10.1037/0012-1649.43.6.1428.

Hirsh-Pasek, K., Michnick Golinkoff, R., Berk, L. E., & Singer, D. G. (2009). *A mandate for playful learning in preschool*. Oxford University Press, Inc.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit Indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling-a Multidisciplinary Journal*, *6*(1), 1–55. 10.1080/10705519909540118.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639. 10.2307/2285946.

Lange, T., & Meaney, T. (2013). *Ipads and mathematical play: A new kind of sandpit for young children*. European Society for Research in Mathematics Education. http://cerme8.metu.edu.tr/wgpapers/WG13/WG13_Lange_Meaney%20.pdf

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]*. Retrieved from September. http://quantpsy.org

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19. 10.1037/1082-989X.7.1.19.

Magne, O. (2003). *Barn oppdager matematikk - Aktiviteter for barn i barnehage og skole [Children discover mathematics - Activities for children i daycare and school]*. Info Vest Forlag.

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in Psychology*, *5*(599), 599. 10.3389/fpsyg.2014.00599.

Ministry of Education and Research. (2011). *Framework Plan for the Content and Tasks of Kindergartens*. Oslo: Ministry of Education and Research.

Ministry of Education and Research. (2017). *Framework Plan for the Content and Tasks of Kindergartens*. Oslo. https://www.udir.no/globalassets/filer/barnehage/rammeplan/framework-plan-for-kindergartens2-2017.pdf

Muthén, B., Muthén, L., & Asparouhov, T. (2015). *Estimator choices with categorical outcomes*.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Muthén & Muthén.

Norwegian Directorate for Education and Training. (2018). *Rammeverk for kartleggingsprøver på 1.-4. trinn*. https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-kartleggingsprover-pa-1.-4.-trinn/#

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. 10.1037/a0031753.

Purpura, D. J., Logan, J. A. R., Hassinger-Das, B., & Napoli, A. R. (2017). Why do early mathematics skills predict later reading? the role of mathematical language. *Developmental Psychology*, *53*(9), 1633–1642. https://doi.org/10.1037/dev0000375

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review*, *44*(1), 41–59. 10.17105/spr44-1.41-59.

Rege, M., Størksen, I., Solli, I., Kalil, A., McClelland, M., Lenes, R., Lunde, S., Breive, S., Carlsen, M., Erfjord, I., & Hundeland, P. (2019). Promoting child development in a universal preschool system: A field experiment impressum. *Economics Division Working Papers, 2019*, 7775.

Rönkkö, M., & Cho, E. (2020). An Updated Guideline for Assessing Discriminant Validity. *Organizational Research Methods*. https://doi.org/10.1177/1094428120968614

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research*. Routledge.

Simmons, F. R., & Singleton, C. (2008). Do weak phonological representations impact on arithmetic development? A review of research into arithmetic and dyslexia. *Dyslexia*, *14*(2), 77–94. 10.1002/dys.341.

Sinclair, N., & SedaghatJou, M. (2013). *Finger counting and adding with touch counts*. Retrieved on from January 15, 2013. https://www.academia.edu/7823550/Finger_Counting_And_Adding_With_TouchCounts

Solheim, O. J., Brønnick, K. S., & Walgermo, B. R. (2012). *Kartlegging av leseferdighet 1-3-trinn*. Analyse av resultater og prøvekvalitet i perioden 2008-2013.

Størksen, I., Ellingsen, I. T., Tvedt, M. S., & Idsøe, E. M. C. (2013). Norsk vokabulartest (NVT) for barn i overgangen mellom barnehage og skole: Psykometrisk vurdering av en nettbrettbasert test. *Spesialpedagogikks forskningsdel 04/13,* 40–54. Utdanningsforbundet.

Størksen, I., & Mosvold, R. (2013). Assessing early math skills with tablet computers: development of the Ani Banani math test (ABMT) for young children. program seminar arranged by UTDANNING2020. Oslo: The Norwegian Research Council, March 18th 2013.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 Countries. *American Educational Research Journal*, *41*(4), 901–926. 10.3102/00028312041004901.

The Norwegian Directorate for Education and Training. (2013). *Nasjonale prøver 2013 - fagmiljøenes analyse av prøvene*. https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Nasjonale-prover-2013—fagmiljoenes-analyse-av-provene/

Van Luit, J. E. H., & Van de Rijt, B. A. M. (2009). *Utrechtse Getalbegrip toets [Utrecht early numeracy test]*. Graviant.

Wechsler, D. (2003). *The Wechsler intelligence scale for children—fourth edition*. Pearson.

Williams, P. E., Weiss, L. G., & Rolfhus, E. L. (2003). *WISC_IV Technical report #2 Psychometric Properties*. https://images.pearsonclinical.com/images/pdf/wisciv/WISCIVTechReport2.pdf

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson tests of achievement*. Riverside.