Exploring the reliability, validity, and dimensionality of the 'Kieler kindergarten test for mathematics'

A. H. Van Hoogmoed, A. Van den Ham, A. Jordan, C. Duchhardt, E. H. Kroesbergen, A. Heinze

Abstract

Mathematical literacy consists of several content areas. However, in kindergarten, most assessments focus solely on the number area. It is yet unknown whether the content areas of mathematical literacy can already be distinguished in young children. The Kieler Kindergartentest for Mathematics (KiKi) is a German instrument that is developed to measure performance in five content areas of mathematics in kindergarten (4;0 – 6;6 years of age). The KiKi was translated into Dutch and administered in 244 children. IRT analyses were adopted to examine reliability and validity evidence for the instrument and to examine the dimensionality of mathematics in Dutch kindergarten. The reliability as well as convergent and discriminant validity evidence were generally good. A three-dimensional model with the dimensions *number* (including *measurement* and *change and relationships), space and shape*, and *data and chance* showed the best fit. This indicates that some but not all content areas can be distinguished in kindergarten. Implications for theory and educational practice are discussed.

Keywords: mathematics, kindergarten, dimensionality, assessment, IRT

1 Introduction

Children's mathematics development starts early in life, far before the start of formal schooling, and is foundational for their later mathematical literacy (Jordan, et al., 2009; Nguyen et al., 2016). According to OECD's Programme for International Student Assessment (PISA), mathematical literacy is students' ability to 'analyse, reason, and communicate ideas effectively as they pose, formulate, solve, and interpret mathematical problems' in various real-life situations (Organisation for Economic Co-operation and Development OECD, 2003). They divide mathematical literacy into four content areas; *quantity*, *space and shape*, *change and relationships*, and *uncertainty*. The content area *quantity* covers the processing and understanding of numbers and quantities that are presented in different ways, and dealing with relative sizes. The content area *space and shape* covers the understanding of shapes of objects, their representations, and relative positions. The content area *change and relationships* concerns dealing with relations and translations between different symbols, tables, and geometries. The content area *uncertainty* deals with data and chance, including statistics and probabilistic reasoning.

Similarly, the National Council for Teachers of Mathematics (NCTM, 2000) formulated five content areas, which have substantial overlap with those formulated by the OECD (see Table 1 for a comparison). Their content areas *sets, numbers, and operations (SNO)* and *measurement* cover the content area *quantity* of the OECD. The content area *algebra* is similar to *change and relationships*, *geometry* is similar to *space and shape*, and *data analysis and probability* is similar to *uncertainty*. In the Netherlands, the National Institute for Curriculum Development (Stichting Leerplan Ontwikkeling; SLO) develops curricula, commissioned by the Dutch government. They divide mathematics into four content

Table 1
Comparison of different divisions of mathematical content areas on the left.
On the right, the inclusion of the different content areas in the measurement instruments ENT-R and the KiKi

Content areas						***Measurement instruments***	
OECD		***NCTM***		***Dutch curriculum***		***ENT-R***	***KiKi***
1 & 2	Quantity	1	SNO (Sets, Numbers & Operations)	1	Numbers and Operations	Yes	Yes
		2	Measurement	2a	Measurement (together with Geometry)	Partly	Yes
3	Space and shape	3	Geometry	2b	Geometry (together with Measurement)	No	Yes
4	Change and relationships	4	Algebra	3	Proportions / Relationships (verhoudingen)	Partly	Yes
5	Uncertainty	5	Data analysis & probability	4	Data analysis (verbanden) (probability not in kindergarten)	No	Yes (data & chance)

areas: *numbers and operations*, *relationships*, *measurement and geometry*, and *data analysis*. The content area *measurement and geometry* can be divided into a sub-area *measurement*, and a sub-area *geometry* (Noteboom et al., 2017; van Groenestijn, et al., 2011). Also other countries in different parts of the world share a similar division into content areas, such as Hong Kong SAR, Botswana, Czech Republic, Georgia, and Israel (TIMSS, 2015). Thus, despite subtle differences in divisions and naming, internationally there is general consensus about the content areas of mathematics as a school subject (see Table 1).

Based on theory and empirical work, researchers have suggested that a distinction between different aspects of mathematical literacy is already present in early childhood (Clements & Sarama, 2007; 2018). However, no single source explicitly describes the scientific basis of the knowledge that children must acquire and at what age. The Dutch National Institute for Curriculum Development (in Dutch SLO) has published a content map of concepts and skills children need to acquire in the different content areas in kindergarten. For the content area *sets, numbers, and operations (SNO)*, children must understand the meaning of numbers (digits and number words) and be able to count flexibly. Moreover, they must acquire concepts such as 'before, after, further, smaller, bigger, smallest, biggest, equal' and know the rank number words. Also, they have to be able to count and estimate quantities, know the concepts 'more/less, most/least, equal, little/much, everything / nothing, approximate, how much'. At the end of kindergarten, they need to be able to add, subtract, and divide quantities by doing, know the concepts 'add, subtract, another one, take away, together, divide, fair, unfair', and be able to subitize. For the content area *measurement*, many concepts need to be acquired, such as 'length, circumference, surface area, volume, weight, day, time, hour, month'. Children need to apply these concepts to compare, order and contrast, and use informal instruments to measure. With regards to time, they need to be able to order events, name the days of the week, and reason about time. With regards to money, children also need to know that there are different coins and banknotes with different values, explore the roles of money, and reason about money. Of course, understanding of numbers is also necessary in the content areas *measurement*, but to a lesser extent than in the content area SNO. For the content area *space and shape* the goals are to orient oneself within space, for example by taking

viewpoints, draw simple maps, and reason about orientation in space; to construct things, for example by repeating patterns, build three-dimensional figures, fold paper etc; operate with shapes and figures, such as sorting objects, study geometric shapes, experiment with figures in mirrors, play with shadows. For the content area *change and relationships,* children need to reason proportionally with concrete objects, and reason about numerical proportions in daily situations. For the area *data and chance,* only data is included in the Dutch curriculum. Children need to be able to construct and use bar charts to order and compare.

Despite the content areas requiring different skills, most research in young children has focused solely on *sets, numbers, and operations (SNO*; e.g. Jordan et al., 2009; Toll et al., 2015; Xenidou-Dervou et al., 2017), and has shown that *SNO* is predictive of mathematical performance later in development. Research in other content areas has shown that performance in these areas also predicts later mathematical performance. For example, patterning ability and block design in kindergarten and the beginning of primary school, as a part of the content area *space and shape*, has been found to predict later math achievement (Burgoyne et al., 2019; Rittle-Johnson et al., 2019). Moreover, the development *change and relationships* has also been studied in children starting in kindergarten (e.g. Hurst & Cordes, 2018; Vanluydt et al., 2022), and has been shown to predict rational number knowledge in a sample of older children (McMullen et al., 2016). Also, *data analysis and probability* has been studied in young children (Nikiforidou et al., 2013), but, to our knowledge, performance in this area has not been studied in relation to later math achievement.

Despite the research in all content areas, in practice, assessment in kindergarten most often has a strong focus on quantity (i.e. *SNO* and *measurement*; e.g. Purpura et al., 2015). An instrument that has often been used to assess mathematics in Dutch kindergartens is the 'Cito test mathematics for kindergartners' (Koerhuis, 2010). This instrument aims to assess general mathematical skills in young children. It focuses on the content areas *SNO, measurement,* and *geometry* (Koerhuis & Keuning, 2011). However, this test has recently been withdrawn, since the Dutch government opposed to standardized testing in kindergarten in which children are compared to each other. Instead, Cito has now developed a new instrument based on observations which examines whether children meet the most important goals set by the Dutch National Institute for Curriculum Development.

A diagnostic instrument used to assess mathematics in kindergarten is the Early Numeracy Test – Revised (ENT-R; Van Luit & Van de Rijt, 2009), a Dutch instrument which has been translated into several languages and is used in Belgium, China, Finland, Germany, Greece, Slovenia, and the United Kingdom (Aunio et al., 2008; Torbeyns et al., 2002; Van de Rijt, et al., 2003). This diagnostic instrument consists of nine domains covering the content area *SNO*, and small parts of the content areas *measurement* and *change and relationships*. The areas *space and shape* and *data and chance* are not addressed by this instrument (see Table 1). The ENT-R is assumed to be a unidimensional measure of numeracy (Van de Rijt et al., 1999). However, some studies have also found two related subscales (Aunio et al., 2006), one measuring relational skills (i.e. organizing and comparing quantities) and one measuring counting skills.

The content of the ENT-R largely aligns with the contents covered in kindergarten education in the Netherlands, in which the main focus is on *SNO*. However, the Dutch kindergarten curriculum also includes the content areas *measurement*, *space and shape*, *change and relationships,* and *data* (but not *chance*). As such, there is a misalignment between the curriculum and the main individual measure of mathematics in kindergarten in the Netherlands. Therefore, new assessment methods covering all content areas of mathematics are needed to gain insight into the needs of kindergartners in the area of mathematics.

One argument that might be raised against an assessment covering all content areas is

that the content areas cannot be differentiated yet at such a young age. The age differentiation hypothesis proposes that cognitive abilities start as a unified general ability and differentiate during development (Garrett, 1946; Li et al., 2004). This would mean that early in development, it might not be possible or necessary to distinguish between different cognitive abilities. As such, it would not be necessary to measure each of the mathematical content areas separately in kindergarten. However, next to evidence in favor of the hypothesis (Li et al., 2004), there is evidence against the age differentiation hypothesis (Bickley et al., 1995; Juan-Espinosa et al., 2000; Tucker-Drob, 2009). In the mathematical domain, recent research has shown that several content areas are already separable in preschoolers. Milburn and colleagues (2019) found four content areas when analyzing mathematical performance in 3-to-5.5-year-olds: *Numeracy* (consisting of numbers, operations, and relations), *geometry*, *patterning*, and *measurement*. Note that *data and chance* was not included in this study. Breauning and collleagues (2020) revealed a 4-factor model including *patterning and geometry, number sense, arithmetic,* and *data analysis and statistics* in 5-year-olds using factor analysis. Although other researchers have also examined the dimensionality of mathematics in young children, these focused on numeracy only, without including geometry and data analysis (Dierendonck et al., 2021; Hirsch et al., 2018).

To adequately measure the different content areas in mathematics in young children in kindergarten, the Kieler Kindergarten Test for Mathematics (Kieler Kindergartentest für Mathematik), shortened KiKi, was developed in Germany (Grüßing et al., 2013). The content areas that are represented in the test are *SNO*, *measurement*, *data and chance*, *change and relationships*, and *space and shape* (i.e. geometry). The items in the KiKi are generally found to be of good psychometric quality (Knopp et al., 2014). The test was developed for children between 4;0 and 6;6 years of age and is administered individually. The test consists of three versions with increasing difficulty. Linking items included in two or three versions of the test enable conversion of the scores to the same scale. As such, monitoring development and comparison between children who received different versions is possible.

Research on an adapted German version of the KiKi for young children (4;0 – 4;6 years) involving the content areas *SNO*, *change and relationships*, and *space and shape* showed a three-dimensional structure of the test (Jordan et al., 2015). In older kindergartners (mean age 5;7 years), the difficult version of the KiKi including all five domains was administered. The results pointed to a three-dimensional model with *SNO* as one factor, *space and shape / change and relationships* as a second factor, and *data and chance* as a third factor. The content area *measurement* did not fit the model, possibly due to covering a too wide range of concepts with only 5 items (Dunekacke et al., 2018). These studies show that children's performance in different content areas of mathematics can indeed be distinguished. However, no validity evidence has been gathered for the whole KiKi instrument (all content areas) within the full age range of 4;0 to 6;6 years of age. Moreover, the validity evidence was gathered in the German context, where kindergartens traditionally follow a social-pedagogic approach focusing mostly on children's general development by promoting, for example, the social competence instead of domain-specific competences like mathematical competence. This means that little instruction is given in the domain of mathematics. The Netherlands has more structured math-related opportunities to learn in early childhood education, mainly focusing on *SNO*, but also on *measurement*, *space and shape*, *change and relationships,* and *data*. The focus on specific domains and neglection of other domains could have an influence on the development and dimensionality of mathematical competences. Following this, for the Netherlands a more prominent distinction between *SNO* and other content areas could be expected

The goal of the current study was twofold. First, we examined the reliability and validity evidence of the Dutch version of the KiKi. Following the Standards of Educational and Psychological testing (AERA et al., 2014) we first studied the psychometric properties, of the KiKi. To examine validity evidence based on relations to other variables, relations between the dimensions of the final model and the ENT-R were examined after testing the dimensionality of the KiKi (AERA et al., 2014). The relations with working memory were also examined to study the discriminant validity evidence. Although both visual-spatial and verbal working memory are known to be related to mathematical achievement in kindergarteners (e.g. Friso-Van Den Bos et al., 2013), correlations between the KiKi and working memory should be substantially lower than correlations between the KiKi and ENT-R to show reasonable discriminant validity evidence. This would imply that the KiKi specifically measures mathematical achievement, instead of more general cognitive abilities.

The second goal of the study was to examine whether the content areas of mathematics can already be distinguished in kindergarten and therefore to evaluate validity evidence based on internal structure (AREA et al., 2014). Four models were tested against each other. If the age differentiation hypothesis would hold (Garrett, 1946), one would expect a one-dimensional model in these young children who have not yet entered formal education. A two-dimensional model was specified including the area *data and chance* versus the other content areas, related to the content areas mainly excluded versus included in the kindergarten curriculum in the Netherlands. A three-dimensional model was specified as *data and chance* versus *space and shape* versus *SNO/measurement/change and relationships*, since only the latter content areas are largely based on the processing of numbers, which is the main focus in education and testing in Dutch kindergartens. Also a five-dimensional model was specified, including all content areas separately, based on the content areas set by the OECD and NCTM. To gather strong evidence in favor of or against the age differentiation hypothesis (Garrett, 1946), differences in dimensionality between the different versions (i.e. different age groups) would be beneficial. However, our sample is not large enough for this kind of analyses. Therefore, if the final model has more than 1 dimension, we will examine the relations between the different dimensions per age group separately. If the age differentiation hypothesis holds, one would expect stronger correlations between the dimensions in easy version (younger children) as compared to the difficult version (older children).

2 Method

2.1 Participants

In total, 244 children from eight kindergartens in the southern and western part of the Netherlands participated in the study. The sample consisted of 133 boys and 111 girls, with a mean age of 5;1 years (sd = 0;8 years). Of these children, 108 were in the first grade of kindergarten, 102 in the second grade of kindergarten, and 30 were in a combined kindergarten group. For four children, information about grade was missing. According to their age, 74 children received the easy KiKi version (4;0-4;6 years), 95 received the medium version (4;7-5;6 years), and 75 received the difficult version (5;7-6;6 years). Next to the KiKi, the ENT-R was administered in 209 of the children, and working memory was measured in 150 of these 209 children.

2.2 Instruments

KiKi-NL

The Kieler Kindergartentest Mathematik (KiKi) is a standardized assessment procedure that measures the mathematical competence of children between the ages of four and six. The administration involves the use of a hand puppet named Kiki. The puppet is meant both as a comforting element, as well as an actor in several of the items. The items represent concrete situations for children which involve the use of mathematics. In order to take into account age-dependent cognitive development and the rapidly growing mathematical knowledge in kindergarten, the KiKi provides

three versions of the test instrument with different requirement levels. The easy version targets 4;0 to 4;6-year-olds, the medium version targets children between 4;7 and 5;6 years of age, and the difficult version targets children between 5;7 and 6;6 years of age. The three versions enable obtaining meaningful results for all children in the target age group within a test duration of 30 minutes. To ensure that the development of children's mathematical competence can also be tracked longitudinally - i.e. results in individual test cases are comparable on a scale - there is a larger intersection of common test items (linking items).The easy and the difficult version consist of 31 items, the medium version of 32 items divided over five content areas:

- Sets, Numbers & Operations (SNO)
- Measurement
- Space and shape
- Change and relationships
- Data and chance

The items in the area *SNO* cover quantity comparisons, number entries, number representations and counting. For the content area *measurement*, the focus is primarily on dealing with order relations among representatives of different sizes. For example, objects have to be arranged according to their length. The content area *space and shape* contains the handling of all kinds of flat or spatial configurations such as the identification of geometric forms, the analysis of simple properties as well as the recognition of different perspectives. In the content area *change and relationships*, the items are intended to determine whether the kindergarten children can use basic cognitive skills such as sorting and classifying to identify and continue geometric patterns in sequences. Elementary number relations (predecessor/successor) as well as simple proportionalities (e.g. longer path - more steps) and anti-proportionalities (more children - fewer sweets per child) must also be recognized in concrete situations. The section *data and chance* contains all situations where statistical data or randomness play a role. Accordingly, the KiKi will on the one hand check whether the children can already handle data, e.g. in the form of creating tally lists to display data. On the other hand, it also examines whether the children have a first understanding of the concept of probability. The contents of all items and the content areas and versions they belong to, can be found in Appendix A.

The content area *SNO* consists of 15 (easy) or 11 (medium and difficult) items, the other content areas are covered by four (easy version) or five (medium and difficult version) items. There are fourteen linking items that are included in all three versions of the test. Next to that, there are six linking items between the easy and medium version, and eleven linking items between the medium and difficult version. The linking items were selected from the different content areas by experts in the area of early mathematics, based on their suitability for children within a larger age range. The test includes three types of response formats: simple multiple-choice (MC, the participant has to choose the correct response option from several available response options), complex multiple-choice (CMC, a number of subtasks with two response options were presented), and short constructed response (SCR, required the participant to give an answer verbally or by manipulating material).

The items were developed at the Leibniz Institute for Science and Mathematics Education in Kiel, Germany, and subjected to an expert rating and initially tested in smaller case studies. Afterwards the items were translated into Dutch. The translation was validated through the method of back translation and checked for adequate language by clinical practitioners working with kindergarteners.

ENT-R

The ENT–R (Early numeracy test – revised) is a paper-and-pencil test measuring early numerical ability in children between 4 and 7 years of age. The test consists of a total of 45 items, which are divided over nine components. The components included in the test are concepts of comparison, classification, correspondence, seriation, use of numerals, synchronized and shortened

counting, general understanding of numbers, resultative counting, and estimation, which is similar to the content areas *SNO*, *measurement*, and *change and relationships*. The test is treated as one-dimensional and consists of two parallel versions (versions A and B). In this study, version A was administered, and the total score (0 – 45) was used in the analyses. This version shows a high reliability, Cronbach's alpha between .91 and .94 (Van de Rijt et al., 1999).

Working memory

Subtests of the Automated Working Memory Assessment (AWMA; Alloway, et al., 2008) were administered to measure verbal working memory and visual-spatial working memory. The subtest Word recall was used to measure verbal working memory. For visual-spatial working memory, Odd-One-Out and Dot Matrix were administered. Sum scores of each subtest were used in the analyses. The test-retest reliability of these subtests are .76 for word recall, .81 for Odd-One-Out, and .83 for dot matrix in children between 4;6 and 11;6 years of age (Alloway, Gathercole, & Pickering, 2006).

2.3 Procedure

Kindergartens were invited to participate in the research by mail. When schools decided to participate in the study, parents of the children received a letter about the research, and were asked to give written informed consent. After parents gave informed consent, children were tested individually in a separate room in their school. Tests were administered in two sessions. In one session, the KiKi was administered. In the other session, the ENT-R and AWMA were administered. Children were rewarded with stickers. Data were gathered between April 2014 and March 2015.

2.4 Analyses

Scaling model for the KiKi

Item and person parameters were estimated for the whole sample (all versions) using a partial credit model (PCM; Masters, 1982). Items were scored dichotomously or polytomously. Categories of polytomous variables with few responses were collapsed in the analyses to avoid possible estimation problems. For item Tot1_r (count as far as possible; SNO), LM4_r (who can see what; S&S) , and MS19a_r (Number of butterflies based on table; D&C) categories were collapsed. To estimate item and person parameters, a scoring of 0.5 points for each category of the polytomous items was applied, while for dichotomous items a scoring of 0 for an incorrect and 1 for the correct response was applied. Mathematical competences were estimated as plausible values (PV). For each student five PV's are drawn. Plausible values (PV) are a way to describe individual competencies at the group level. They allow unbiased estimates of population-level effects. A detailed review of the plausible values methodology is given in Mislevy (1991).

The psychometric properties of the KiKi items

In order to ensure appropriate psychometric properties, the quality of the test was examined in several analyses. All analyses were conducted for the whole test and for the different versions, respectively. Moreover, analyses were conducted for the different dimensional models (one-, two-, three, and five-dimensional model). Only the results of the one-dimensional model are presented. To ensure the appropriateness of scaling the three test versions on a common scale, we examined measurement invariance for the three test versions by adopting the minimum effect null hypothesis described in Fischer, Rohm, Gnambs, and Carstensen (2016). Therefore, we tested whether the item parameters derived in the linking subsample showed a non-negligible shift in item difficulties comparing the different test versions. For this purpose, we considered the common items of the test versions easy and medium, of the test versions medium and difficult and of the test versions easy and difficult; see Table 2). Decisions to exclude variables were based on all analyses. The data were analyzed in ConQuest (Wu, Adams, & Wilson, 2007).

The fit of the items to the partial credit model (Masters, 1982) was evaluated using two indices (see Pohl & Carstensen, 2012), i.e. weighted mean square (WMNSQ) and correlations with the total score. Items with a WMNSQ > 1.15 (and |*t*-value| > 6) were considered as having a noticeable item misfit, and items with a WMNSQ > 1.20 (and |*t*-value| > 8) were judged as having a considerable item misfit and were subsequently checked. Correlations of the item score with the total score (equal to the discrimination value as computed in ConQuest; Wu, Adams, & Wilson, 2007) greater than .30 were considered as good, greater than .20 as acceptable, and below .20 as problematic.

Evidence based on internal structure

In a first step, we tested whether the person-parameters (i.e. PV's) differed per test version (i.e. age group), using a univariate ANOVA with Bonferroni correction. In a second step, we examined the dimensionality of the test by specifying a five-dimensional model based on the five different content areas, a three-dimensional model which combined the content areas *Sets, Numbers & Operations*, *Change and Relationships* and *Measurement* into one dimension and a two-dimensional model which contained the *Data and Chance* content area on one dimension and a combination of all other content areas on the second dimension. These models were tested against a one-dimensional model. Each item was assigned to one content area (between-item-multidimensionality).

Monte Carlo estimation implemented in ConQuest (Wu et al., 2007) was used (nodes = 5,000) because of the multidimensionality of the models (Atanassov & Dimov, 2008; Pohl & Carstensen, 2012). The variances and correlations were computed. The models were compared on the comparative fit indices with the Akaike information criterion (AIC), consistent AIC (CAIC), and Bayesian information criterion (BIC). This was done using the value of the log-likelihood. Lower values on these information criteria indicate a better tradeoff between model fit (log-likelihood) and parsimony of the model (the number of parameters). Relations between the dimensions in the final model were computed per version (i.e. age group) to examine differences in relations between the dimensions per age group.

Evidence based on relations to other variables

The relations between the KiKi, the ENT-R and working memory were calculated using Pearson correlation. For the best fitting dimensional model for the KiKi, the PV's of each dimension were used to correlate to the total score of the ENT-R and working memory. For visual-spatial working memory, a mean score was calculated based on the standardized scores of Dot Matrix and Odd-One-Out.

3 Results

3.1 Psychometric properties

Here we present the psychometric properties of the KiKi for the one-dimensional model. In the analyses, the mean ability was constrained to be zero, assuming that the underlying latent trait is normally distributed in the population. The variance was estimated to be 1.479, indicating that the test differentiated reasonably well between subjects. The reliability of the test (EAP/PV reliability = 0.883) was good.

The extent to which the item difficulties and location parameters match the children's abilities is shown in Figure 1. The distribution of the estimated test takers' abilities is mapped onto the left side whereas the right side shows the distribution of item difficulties. The items covered a wide range of the ability distribution of test persons. However, there were no very difficult items, and a few items that were too easy. Consequently, low and medium abilities can be measured relatively precisely, while subjects with a high mathematical competence will have a larger standard error. In general, all content areas contained items with varying difficulty. Only in the content area *data & chance*, no items had a difficulty higher than 0.3.

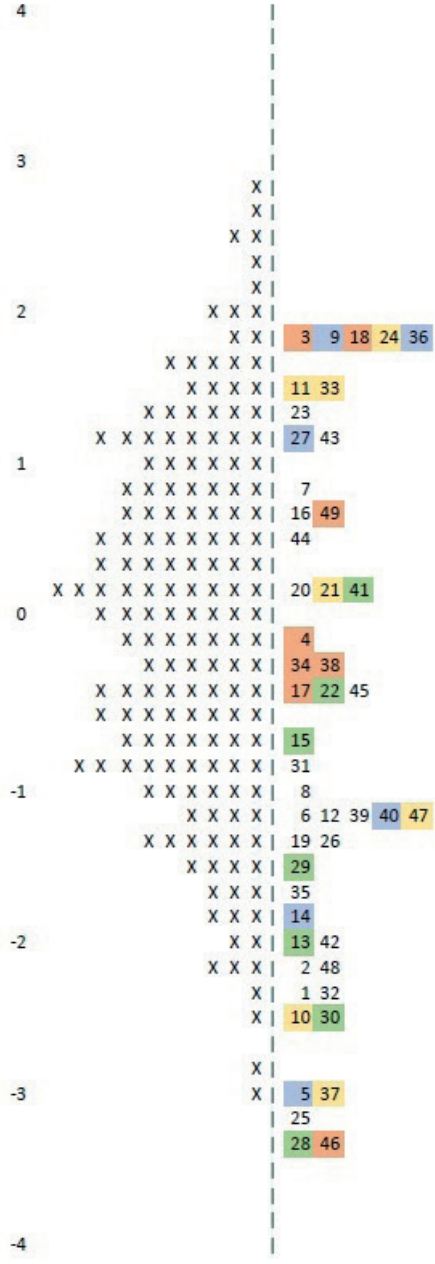

Figure 1. Test targeting. The distribution of person abilities in the sample is depicted on the left-hand side of the graph, with each 'X' representing 1.7 cases. The difficulty of the items is depicted on the right-hand side of the graph, with each number representing one item (corresponding to Table 2). Colors refer to the different content areas (white = SNO, red = Change & Relationships, blue = Space & Shape, yellow = Measurement, and green = Data & Chance)

Overall, 50 different items with different response formats were used (see Appendix A). Note that there was no multi-matrix design regarding the choice and the order of the items *within* a specific test booklet. Participants were assigned to a test version based on their age. After removal of one item (see below), six to eight items within each content area were included in the analyses, except for *SNO*, for which 21 items were included in the analyses.

The estimated item difficulties (for dichotomous variables) and location parameters (for the polytomous variables) are presented in the Appendix. The evaluation of the item fit was performed on the basis of the partial credit model, using the dichotomous and polytomous items. Overall, the item fit was good for the three versions for both the multidimensional models as well as the combined one-dimensional data. Therefore, only the values of the over-all scaling are presented below. Item S4_r has a negative item-total correlation (rit= -.14) and a problematic WMNSQ (WMNSQ=1.34). This might be due to the content of the item, focusing on chance and including a relatively difficult cover story. The other items in the content area *data and chance* mainly focus on data instead of chance. Therefore item S4_r was excluded from further analyses. Beyond that, values of WMNSQ were close to 1 with the lowest value being 0.76 (item Tot1_r) and the highest being 1.31 (item LM15_r). No item exhibited a *t*-value of the WMNSQ greater than |6|. All item characteristic curves showed a good fit of the items. Thus, there seemed to be no indication of severe item over- or underfit. However, since the sample is not that large, the *t*-values will not easily exceed |6| and therefore this criterion is not strictly used. If only based on the WMNSQ, three items exceed a WMNSQ of 1.20.

According to the WMNSQ, there were three items with noticeable misfit (WMNSQ>1.15; L10_r, Tot16_r, Tot27_r) and three items with considerable misfit (WMNSQ>1.20; MS2_r, LM15_r, Tot18_r). The items with misfit were all examined. Item L10_r has an acceptable item-total correlation and is one of the few easy items in this test.

Table 2
Differential Item Functioning Analyses between the Test Versions

	Easy vs. Medium			Medium vs. Difficult			Easy vs. Difficult		
Item name	**Δσ**	**$SE_{\Delta\sigma}$**	**F**	**Δσ**	**$SE_{\Delta\sigma}$**	**F**	**Δσ**	**$SE_{\Delta\sigma}$**	**F**
Tot1_r	0.00	0.25	0.00	-0.02	0.25	0.00	0.07	0.25	0.08
MS2_r				0.09	0.28	0.09			
Tot3_r	-0.02	0.21	0.01	0.05	0.21	0.06			
LM4_r	0.02	0.53	0.00				-0.01	0.21	0.00
LM5_r	0.04	0.25	0.02						
LM6_r	0.04	0.25	0.02						
LM7_r	0.03	0.32	0.01						
Tot8_r	-0.01	0.23	0.00	0.02	0.23	0.01	0.03	0.23	0.01
Tot9_r	0.00	0.26	0.00	0.01	0.27	0.00	0.04	0.27	0.03
MS10_r				0.03	0.26	0.01			
Tot11_r	-0.02	0.21	0.01	0.07	0.21	0.10	-0.04	0.21	0.03
LM12_r	0.05	0.24	0.03						
MS13_r				0.05	0.24	0.05			
Tot14_r	-0.02	0.21	0.01	0.06	0.21	0.07	-0.02	0.21	0.01
LM15_r	0.05	0.24	0.04						
Tot16_r	-0.03	0.25	0.02	0.09	0.25	0.11	-0.06	0.25	0.06
MS17_r				0.01	0.30	0.00			
Tot18_r	-0.03	0.22	0.02	0.08	0.22	0.11	-0.05	0.22	0.05
MS19_r				0.00	0.32	0.00			
MS19a_r				0.00	0.37	0.00			
Tot20_r	-0.01	0.22	0.00	0.03	0.22	0.02	0.02	0.22	0.01
MS21_r				0.08	0.26	0.09			
Tot22_r	-0.01	0.21	0.01	0.05	0.21	0.05	0.00	0.21	0.00
MS23_r				0.08	0.28	0.09			
MS24_r				0.04	0.25	0.02			
Tot25_r	-0.01	0.22	0.00	0.03	0.23	0.01	0.02	0.23	0.01
Tot26_r	-0.01	0.23	0.00	0.02	0.23	0.01	0.02	0.23	0.01
Tot27_r	-0.02	0.21	0.01	0.06	0.21	0.08	-0.02	0.21	0.01
Tot29_r	-0.01	0.21	0.00	0.04	0.21	0.04	0.00	0.21	0.00
MS30_r				0.03	0.35	0.01			
MS31_r				0.06	0.24	0.07			

Note. Δσ = Difference in item difficulty parameters; SEΔσ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test (see Fischer et al., 2016). The critical value for the minimum effects hypothesis using an α of .05 is F0154 (1, 167, 2.61) = 10.84 for the easy and medium version and F2154 (1, 170, 2.63) = 10.87 for the test with medium and high level of difficulty and F0154 (1, 147, 2.3) = 10.21 for the test with medium and high level of difficulty. A nonsignificant test indicates measurement invariance.

Therefore, it is desirable to keep this item. Items Tot18_r and Tot27_r have an acceptable respective good item-total correlation and lie on the high end of the item scale. As such, it is desirable to keep these items, too. Item Tot16_r just misses the benchmark for an acceptable item-total correlation in the overall one-dimensional and the overall three-dimensional scaling (0.19). We decided to keep this item because it is one of the few very difficult items and because this item shows good WMNSQ values for the easy and the moderate booklet (WMNSQ= 1.12/1.07) and has only a noticeable misfit in the difficult booklet (WMNSQ= 1.19) in the separate scaling of the three booklets. Item MS2_r has a noticeable misfit and a problematic item-total correlation. However, it is one of few very difficult items and shows, in contrast to the one-dimensional model displayed here, good WMNSQ in the (final) three-dimensional scaling (WMNSQ= 1.07). Item LM15_r is the only item with considerable misfit regarding the WMNSQ and the item-total correlation. In the (final) three-dimensional scaling this item has a good WMNSQ (WMNSQ=1.14) and a still problematic item-total correlation of r=.12. Regardless of these somewhat problematic values, we decided to keep this item because it is the only link item between the easy and the moderate booklet in the domain Data and Chance. To examine measurement invariance between test versions, we considered the common items of the test versions easy and medium, of the test versions medium and difficult and of the test versions easy and difficult. Adopting the minimum effect null hypothesis described in Fischer et al. (2016) the examinations identified no significant DIF (inspecting the differences in item difficulties between the test versions and the respective tests for measurement invariance based on the Wald statistic; see Table 2). Thus, overall, there was no pronounced DIF with regard to the different test versions, indicating that the three versions can be scaled on a common scale.

3.2 Evidence related to internal structure

First, we examined whether the ability estimates on the KiKi were higher for older children (i.e. children who had received a more difficult version of the test). The results show that there was indeed a general difference between the scores of the children tested with the easy version, mean PV's ranging between -0.984 (SE = .120) and -0.905 (SE = .125), the medium version, mean PV's ranging between -0.117 (SE = .099) and 0.015 (SE = .091) , and the difficult version, mean PV's ranging between 0.925 (SE = .100) and 1.067 (SE = .110), all $F(2, 241) > 66.59$, $p < .001$, $\eta_p^2 > .356$. Also, the differences between the easy and medium version, and medium and difficult version were significant (all $p < .001$), showing that older children, who are assumed to have higher abilities, indeed score better on the KiKi when scaling all versions together.

Second, we examined the dimensionality of the test. The correlations among the five dimensions were rather high and varied

Table 3
Results of Five-Dimensional Scaling. Variance of the Dimensions are Depicted in the Diagonal, Correlations are given in the Off-Diagonal

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Dim 1: Sets, Numbers & Operations (21 items)	(3.224)				
Dim 2: Space and shape (7 items)	0.839	(0.581)			
Dim 3: Change and relationships (7 items)	0.955	0.877	(2.219)		
Dim 4: Measurement (7 items)	0.856	0.831	0.933	(1.07)	
Dim 5: Data and chance (7 items)	0.855	0.764	0.876	0.833	(0.954)

Table 4
Model fit statistics of model with different numbers of dimensions

Model	Deviance	Number of parameters	AIC	BIC	CAIC
Unidimensional	8469.32	53	8575	8761	8814
2 dimensional	8438.56	55	8549	8741	8796
3 dimensional	8385.60	58	8502	8704	8762
5 dimensional	8346.90	67	8481	8715	8782

Table 5
Correlations between the dimensions per version

Version	Age group (yrs)	Correlation number domain – space & shape	Correlation number domain – data & chance	Correlation space & shape – data & chance
Easy	4;0 – 4;6	.614	.558	.347
Medium	4;7 – 5;6	.825	.817	.670
Difficult	5;7 – 6;6	.834	.863	.896

Table 6
Correlations between the dimensions of the KiKi, the ENT-R and working memory

	1.	2.	3.	4.	5.	6.
1. KiKi - number dimension	-					
2. KiKi - space and shape	.886	-				
3. KiKi - data and chance	.870	.802	-			
4. ENT-R	.804	.733	.725	-		
5. verbal working memory	.558	.512	.502	.518	-	
6. visual-spatial working memory	.601	.516	.559	.638	.430	-

between .76 and .96 (see Table 3). Especially, the dimensions *SNO*, *change and relationships,* and *measurement* have a high proportion of shared variance, supporting the three-dimensional model. Model fit of the different models is shown in Table 4. According to the BIC and CAIC fit-indices, the three-dimensional model describes the data best. Therefore, the three-dimensional model is used in subsequent analyses. The EAP/PV reliability of all dimensions is good (.71 – .89).

Correlations between the 3 dimensions of the final model were calculated separately for each version (see Table 5). The results show that the correlations between the dimensions are moderate to high for all versions but also show a unique proportion of variance for each of the dimensions. The correlations are higher for the medium and difficult version than for the easy version.

3.3 Evidence related to other variables

Person parameters from the three-dimensional model were correlated to the scores of the ENT-R to test for convergent validity evidence and to working memory to test for discriminant evidence. Results are reported in Table 6.

Relations between the dimensions of the KiKi were high. Moreover, relations between the KiKi and ENT-R are all higher than r = .72, indicating good convergent validity evidence. As could be expected from the

content of the ENT-R, the correlation with the number dimension is the highest (r = 0.82). Relations between the dimensions of the KiKi and working memory were much lower, and similar to the relation between the ENT-R and working memory. Comparison of correlations from dependent samples (Eid et al., 2011, as implemented in Lenhard & Lenhard, 2014) on the 132 participants with complete data revealed that the correlations of the dimensions of the KiKi with the ENT-R were indeed significantly higher than the correlations with the working memory scores (all $Z > 5.15$, $p < .001$). The reasonable discriminant validity evidence indicates that different concepts are measured by the KiKi and AWMA.

4 Discussion

The KiKi is a relatively new instrument that measures all five content areas of mathematics in kindergarten. The aim of this study was twofold. First, we examined reliability and validity evidence of the KiKi in the Dutch context. The results showed that the EAP/PV reliability of the test is generally good. Only one item needed to be removed from the analysis because of negative discrimination. In addition, there were only few items with low discrimination. The items with low discrimination were mainly easy items, which are good for motivating young children. The test is especially good at differentiating between children scoring around the mean and below. For high performers, the test is less reliable due to the lack of (very) difficult items. The test would therefore be suitable for formative assessment, since it indicates in which content areas additional instruction is needed for which children. Moreover, the test could be used for cross-sectional or longitudinal examination of mathematical competence in kindergarten, or the evaluation of mathematics interventions in kindergarten. In its current form, the test is less suitable for summative assessment, since no norms are available. However, this is in line with the recent view on education of young children by the Dutch government who also disfavors norm-based testing in kindergarten.

The second aim of the study was to examine whether the five content areas of mathematics could already be differentiated in kindergarten by testing the dimensionality of the KiKi. The results showed evidence for the three-dimensional model with *data and chance* versus *space and shape* versus the other numerical-based content areas (*SNO; change and relationships; measurement*). Although there was evidence for the five-dimensional model as well, two of the three correlations between the dimensions *SNO*, *change and relationships,* and *measurement* were above $r = .90$. Moreover, all three content areas heavily rely on children's ability to give meaning to numbers. This contrasts with the content areas *data and chance* and *space and shape* which less heavily rely on knowledge of symbolic numbers. Moreover, the ENT-R also measures *SNO*, and parts of *measurement* and *change and relationships*. This instrument is also considered to be a unidimensional measure of numeracy (Van de Rijt et al., 1999). Based on the current results, it is suggested that in kindergarten numeracy (*SNO*), including *measurement* and *change and relationships* might still be seen as one dimension, with *data and chance* and *space and shape* being additional dimensions that can be measured in kindergarten. However, the correlations between these dimensions and the numeracy-dimension were also quite high. This is probably because even *data and chance* requires a basic ability to deal with numbers. As such, performance in most domains is likely to be constrained by the ability in numeracy (mostly counting). Within the domain of *data and chance,* the items targeting *chance* (vs *data*) showed slightly poorer item fits. This suggests that the content area *data and chance* may consist of two separable domains of *data* and *chance*. This would also align better with the division by the Dutch National Insitute for Curriculum Development (SLO), which does include *data*, but not *chance*. However, the small number of items on both *data* and *chance* did not allow us to examine whether two different dimensions would fit better. Future research including more items on *data and chance* may elucidate on this question.

The results regarding the dimensionality differ from previous results of the German difficult version of the KiKi in which *space and shape* and *change and relationships* were combined in one dimension, next to the dimension *SNO* and the dimension *data and chance* (Dunekacke et al., 2018). This might relate to differences in the educational system (in the German kindergarten tradition, mathematics belongs to the less important educational goals), or differences in the samples. The German version only targeted the older children, measured with the difficult version of the test. In the current study, we used all three different versions. As such, in total there were more items representing each content area, which may have led to more reliable results. Future research examining the dimensionality with the exact same version of the KiKi in different samples in different countries may reveal differences between countries with different educational systems. The current results suggest that it is advisable to analyze the results three-dimensionally in the Dutch educational context.

Our results partly align with the results of Milburn and colleagues (2019), who showed that the four content areas *SNO* (including *relationships*), *measurement*, *geometry*, and *patterning* were already separable in preschoolers between 3.5 and 5.5 years of age. In their study, *data and chance* was not included. Here we have shown that this content area is also separable from the other content areas already between 4 and 6 years of age. However, our results did not show separate content areas for *SNO, measurement*, and *change and relationships*. This may be due to differences in sample, or the relatively small number of items included in the areas *measurement* and *change and relationships* in our study.

The results of the current study and the other studies showing multi-dimensionality of mathematical competence in kindergarten (Dunekacke et al., 2018; Milburn et al., 2019) provide some evidence against the age differentiation hypothesis (Garrett, 1946; Li et al., 2004). Since the participating children do not yet attend formal education, one would expect little differentiation between the different content areas yet, based on the age differentiation hypothesis. However, our sample was too small to test for differences in dimensionality between the different age groups. Correlations between the three dimensions in the final model for the different age groups were higher for the more difficult versions as compared to the easier versions. These results go against the age-differentiation hypothesis, based on which decreasing correlations would be expected with age and version. This is in line with earlier research showing evidence against the age differentiation hypothesis (Bickleyet al., 1995; Juan-Espinosa et al., 2000; Tucker-Drob, 2009).

The three dimensions of the KiKi were related to the ENT-R to examine convergent validity evidence. All dimensions showed high correlations with the ENT-R. As expected, the correlation was highest for the number-dimension. This can be explained by the core focus on numeracy in the ENT-R (Van de Rijt et al., 1999). To test for the discriminant validity evidence of the KiKi, the dimensions were correlated with working memory. The correlations between the KiKi and working memory measures were significantly lower than the correlations between the KiKi and ENT-R, and similar to correlations between the ENT-R and working memory. This shows that the discriminant validity evidence of the KiKi is also good. As such, the KiKi has the potential to become an important addition to the currently used assessment instruments for mathematics in kindergarten. However, the validity evidence for the different dimensions is limited, since most content areas are only covered by four or five items per version. Also, the fixed order in which the items were administered may have affected the difficulty estimates. Moreover, no norms are currently available. Therefore, adaptations and additional research would be necessary to enable the use of the test for diagnostic reasons.

The results on the dimensionality align with other studies showing separable content areas in mathematics (Breauning et al., 2020; Dunekacke et al., 2018; Milburn et al., 2019).

Recent research has shown that the factor structure of math abilities in young children is quite stable over time (Jordan, Kaplan, Nabors Olah, & Locuniak, 2006; Breauning et al., 2020). Different domains measured in kindergarten have also been shown to predict mathematical performance in sixth grade (Hirsch et al., 2018). These results have both scientific and societal implications. The most important scientific implication is that early numeracy is not a single construct. Instead, a more differentiated view on early numeracy should be aimed for, especially since intra-individual variability between different math domains seems to be present (Dowker, 2008; Hirsch et al., 2018).

The combination of a relatively stable structure of math abilities and the predictive value of multiple math abilities (i.e. content areas) for later mathematics also has societal implications, for both assessment and teaching. With regards to assessment, it emphasizes the importance ofassessing multiple content areas already at an early age. The KiKi is the first instrument for Dutch kindergarten that could fulfil this need for early assessment, for example for formative assessment or assessment of the effectiveness of interventions.

When the KiKi can be implemented in kindergarten, this will allow teachers to adjust their teaching to the content areas in which the children have most difficulty. Teachers can include this content in the interactive lessons they provide to the whole group, and depending on the specific topic, use guided play in the different play areas in the classroom. For instance, if the results of the KiKi suggest that the content area *space and shape* requires more attention, the teacher could introduce perspective taking in the group lessons related to the current theme the children are working on. For example, if this theme is 'knights', one could imagine showing a castle to the children from different perspectives. From which viewpoints were the knights able to defend the castle? What parts of the surroundings could they not see? Could they see each other to communicate? After discussing these questions, guided play could be deployed to help children practice, for example when they are playing with building blocks in the classroom.

Currently, the Dutch kindergarten curriculum primarily focuses on *SNO*, but also involves *measurement*, *space and shape*, *change and relationships*, and *data* (but not *chance*). To cover the full range of mathematical competences needed to attain mathematical literacy (OECD, 2003), *chance* may be included in the kindergarten curriculum as well. However, the main focus should remain on the domain of *SNO*, since a basic understanding of numbers is necessary to be successful in parts of other content areas as well. Also, future research is needed to examine whether performance in *data and chance* is also predictive of later mathematics achievement, like performance in the other content areas (e.g. Burgoyne et al., 2019; Rittle-Johnson, et al., 2019).

To conclude, we were able to present validity evidence for the KiKi, but also revealed some problems. Showing reliable test score interpretations using PV's and a three-dimensional structure, we can conclude that the KiKi test is an effective instrument measuring a broad concept of mathematical competence in four to six year old children within 30 minutes. Therefore, the KiKi has additional value over the already existing tests for mathematics in kindergarten, which often mainly focus on *sets, numbers, and operations*.

References

Alloway, T. P., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2008). Evaluating the validity of the automated working memory assessment. *Educational Psychology, 28*(7), 725-734.

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable?. *Child development*, 77(6), 1698-1716.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, USA: American Educational Research Association.

Aunio, P., Aubrey, C., Godfrey, R., Pan, Y., & Liu, Y. (2008). Children's early numeracy in England, Finland and people's republic of China. *International Journal of Early Years Education, 16*(3), 203-221.

Aunio, P., Niemivirta, M., Hautamäki, J., Van Luit, J. E., Shi, J., & Zhang, M. (2006). Young children's number sense in China and Finland. *Scandinavian Journal of Educational Research, 50*(5), 483-502.

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence, 20*(3), 309-328.

Braeuning, D., Ribner, A., Moeller, K., & Blair, C. (2020). The multifactorial nature of early numeracy and its stability. *Frontiers in Psychology*, *11*, 2965.

Burgoyne, K., Malone, S., Lervag, A., & Hulme, C. (2019). Pattern understanding is a predictor of early reading and arithmetic skills. *Early Childhood Research Quarterly*, *49*, 69-80.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles–Results from Germany. *Research on PISA* (pp. 199-213). Springer.

Dierendonck, C., de Chambrier, A. F., Fagnant, A., Luxembourger, C., Tinnes-Vigne, M., & Poncelet, D. (2021). Investigating the dimensionality of early numeracy using the bifactor exploratory structural equation modeling framework. *Frontiers in psychology*, *12*, 2195.

Dunekacke, S., Grüßing, M., & Heinze, A. (2018). Is considering numerical competence sufficient? the structure of 6-year-old preschool children's mathematical competence. In C. Benz, A. S. Steinweg, H. Gasteiger, P. Schöner, H. Vollmuth & J. Zöllner (Eds.), *Mathematics education in the early years. results from the POEM 3 conference, 2016* (pp. S145-S157). Cham, Switzerland: Springer International. doi:10.1007/978-3-319-78220-1_8

Eid, M. Gollwitzer, M., & Schmitt, M. (2011). *Statistik und Forschungsmethoden Lehrbuch.* Weinheum (Germany): Beltz.

Friso-Van den Bos, I., Van der Ven, S. H., Kroesbergen, E. H., & Van Luit, J. E. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational research review*, *10*, 29-44.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. (2016). *Linking the Data of the Competence Tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Garrett, H. E. (1946). A developmental theory of intelligence. *American Psychologist, 1*(9), 372.

Grüßing, M., Heinze, A., Duchhardt, C., Ehmke, T., Knopp, E., & Neumann, I. (2013). KiKi–Kieler kindergartentest mathematik zur erfassung mathematischer kompetenz von vier-bis sechsjährigen kindern im vorschulalter. *Diagnostik Mathematischer Kompetenzen,* 67-80.

Hirsch, S., Lambert, K., Coppens, K., & Moeller, K. (2018). Basic numerical competences in large-scale assessment data: Structure and long-term relevance. *Journal of experimental child psychology*, *167*, 32-48.

Hurst, M. A. & Cordes, S. (2018). Attending to Relations. *Developmental Psychology, 54* (3), 428-439. doi: 10.1037/dev0000440.

Jordan, A., Duchhardt, C., Heinze, A., Tresp, T., & Grüßing, M. (2015). Mehr als numerische basiskompetenzen? zur dimensionalität und struktur mathematischer kompetenz von kindergartenkindern. *Psychologie in Erziehung Und Unterricht, 62*(3), 205-217.

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

Jordan, N. C., Kaplan, D., Nabors Oláh, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child development*, *77*(1), 153-175.

Juan-Espinosa, M., Garcı´a, L. F., Colom, R., & Abad, F. J. (2000). Testing the age related differentiation hypothesis through the wechsler's scales. *Personality and Individual Differences, 29*(6), 1069-1075.

Knopp, E., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Neumann, I. (2014). Von Mengen, Zahlen und Operationen bis hin zu Daten und Zufall – Erprobung eines Itempools zum Erfassen der mathematischen Kompetenz von Kindergartenkintern. *Zeitschrift für Grundschulforschung, 7*(1), 20-34.

Koerhuis, I. (2010). Rekenen voor kleuters. Arnhem: Cito.

Koerhuis, I., & Keuning, J. (2011). *Wetenschappelijke verantwoording van de toetsen Rekenen voor kleuters*. Arnhem: Cito.

Lenhard, W. & Lenhard, A. (2014). *Hypothesis Tests for Comparing Correlations*. Bibergau (Germany): Psychometrica. DOI: 10.13140/RG.2.1.2954.1367. Retrieved from: https://www.psychometrica.de/correlation.html.

Li, S., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science, 15*(3), 155-163.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

McMullen, J., Hannula-Sormunen, M. M., Laakkonen, E., & Lehtinen, E. (2016). Spontaneous focusing on quantitative relations as a predictor of the development of rational number conceptual knowledge. *Journal of Educational Psychology*, *108*(6), 857.

Milburn, T. F., Lonigan, C. J., DeFlorio, L., & Klein, A. (2019). Dimensionality of preschoolers' informal mathematical abilities. *Early Childhood Research Quarterly, 47*, 487-495. doi:10.1016/j.ecresq.2018.07.006

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177–196.

NCTM. (2000). *Principles and standards for school mathematics*. Reston, VA: National council of teachers of mathematics.

Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., & Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly, 36*, 550-560.

Nikiforidou, Z., Pange, J., & Chadjipadelis, T. (2013). Intuitive and informal knowledge in preschoolers' development of probabilistic thinking. *International Journal of Early Childhood, 45*(3), 347-357.

Noteboom, A., Aartsen, A., & Lit, S. (2017). *Tussendoelen rekenen-wiskunde voor het primair onderwijs. uitwerkingen van rekendoelen voor groep 2 tot en met 8 op weg naar streefniveau 1S.* Enschede: SLO.

OECD (Ed.). (2003). *The PISA 2003 assessment framework - mathematics, reading, science and problem solving knowledge and skills.* OECD publishing.

Pohl, S. & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests (NEPS Working Paper No. 14)*. Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

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.

Rittle-Johnson, B., Zippert, E. L., & Boice, K. L. (2019). The roles of patterning and spatial skills in early mathematics development. *Early Childhood Research Quarterly, 46*, 166-178.

TIMSS (2015). *TIMSS 2015 encyclopedia*. https://timssandpirls.bc.edu/timss2015/encyclopedia/countries/

Toll, S. W., Van Viersen, S., Kroesbergen, E. H., & Van Luit, J. E. (2015). The development of (non-) symbolic comparison skills throughout kindergarten and their relations with basic mathematical skills. *Learning and Individual Differences, 38*, 10-17.

Torbeyns, J., Van den Noortgate, W., Ghesquière, P., Verschaffel, L., Van de Rijt, B. A. M., & Van Luit, J. E. (2002). Development of early numeracy in 5-to 7-year-old children: A comparison between flanders and the netherlands. *Educational Research and Evaluation, 8*(3), 249-275.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental psychology*, *45*(4), 1097.

Vanluydt, E. , Verschaffel, L. & Van Dooren, W. (9000). The Early Development of Proportional Reasoning. *Journal of Educational Psychology, Publish Ahead of Print* , doi:

10.1037/edu0000734.
van de Rijt, B. A. M., Godfrey, R., Aubrey, C., van Luit, J. E. H., Ghesquière, P., Torbeyns, J., . . . Tzouriadou, M. (2003). The development of early numeracy in Europe. *Journal of Early Childhood Research, 1*(2), 155-180. doi:10.1177/1476718X030012002
van de Rijt, B., Van Luit, J., & Pennings, A. H. (1999). The construction of the Utrecht early mathematical competence scales. *Educational and Psychological Measurement, 59*(2), 289-309.
van Groenestijn, M., Janssen, C., & Borghouts, C. (2011). *Protocol ernstige RekenWiskundeproblemen en dyscalculie (ERWD): BAO, SBO, SO.* Van Gorcum.
van Luit, H., & van de Rijt, B. (2009). *Utrechtse getalbegrip toets-revised.* Graviant Educatieve Uitgaven.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.
Wu, M., Adams, R. J., & Wilson, M. R. (2007). ACER conquest version 2.0: Generalised item modelling software [computer software]. Camberwell, Victoria: ACER Press.
Xenidou-Dervou, I., Molenaar, D., Ansari, D., van der Schoot, M., & van Lieshout, E. C. (2017). Non-symbolic and symbolic magnitude comparison skills as longitudinal predictors of mathematical achievement. *Learning and Instruction, 50*, 1-13.

Auteurs

Anne van Hoogmoed is universitair docent aan het Behavioural Science Institute van de Radboud Universiteit. Ten tijde van uitvoering van het onderzoek was zij verbonden aan de Universiteit Utrecht en de Rijksuniversiteit Groningen. **Ann-Katrin van den Ham** is universitair docent aan de Universiteit van Hamburg. Ten tijde van uitvoering van het onderzoek was zij verbonden aan de IPN-Leibniz Institute for Science and Mathematics Education, Kiel, Duitsland. **Anne-Katrin Jordan** is hoogleraar muziektherapie en muziekonderwijs aan de MSH Medical School Hamburg. Ten tijde van uitvoering van het onderzoek was zij verbonden aan de IPN-Leibniz Institute for Science and Mathematics Education, Kiel, Duitsland.

Christoph Duchhardt is universitair docent didactiek van de wiskunde aan het Universiteit Bremen. Ten tijde van uitvoering van het onderzoek was hij verbonden aan de IPN-Leibniz Institute for Science and Mathematics Education, Kiel, Duitsland.

Evelyn Kroesbergen is hoogleraar orthopedagogiek en verbonden aan de Radboud Universiteit Nijmegen en de Universiteit Utrecht. **Aiso Heinze** is hoogleraar didactiek rekenen-wiskunde bij het IPN-Leibniz Institute for Science and Mathematics Education, Kiel, Duitsland. *Correspondentieadres:* Anne van Hoogmoed, Radboud Universiteit Nijmegen, Behavioural Science Institute, Postbus 9104, 6500 HE Nijmegen. Email: anne.vanhoogmoed@ru.nl

Samenvatting

Een exploratief onderzoek naar de betrouwbaarheid, validiteit en dimensionaliteit van de 'Kieler kindergarten test for mathematics'

Gecijferdheid of wiskundige geletterdheid bestaat uit verschillende domeinen. In de kleuterklassen (groep 1 en 2) is toetsing echter enkel gericht op het domein getallen en bewerkingen. Het is nog onduidelijk of de verschillende domeinen van gecijferdheid al onderscheiden kunnen worden bij kleuters. De 'Kieler Kindergarten Test für Mathematik' (KiKi) is een Duits instrument dat ontwikkeld is om vaardigheden in vijf domeinen van gecijferdheid te meten bij kleuters van 4 tot 6;6 jaar oud. De KiKi is vertaald in het Nederlands en bij 244 kinderen afgenomen. Met behulp van IRT-analyses is de betrouwbaarheid en de validiteit van het instrument onderzocht en gekeken naar de dimensionaliteit van gecijferdheid bij kleuters. De betrouwbaarheid en de convergente en discriminante validiteit bleken over het algemeen goed. Een driedimensionaal model met de dimensies Getallen en bewerkingen inclusief meten en verhoudingen, Ruimte en Vormen, en Data en Kans was het meest passend. Deze resultaten laten zien dat sommige, maar nog niet alle, domeinen van gecijferdheid onderscheiden kunnen worden bij kleuters. Implicaties hiervan voor theorie en de onderwijspraktijk worden besproken.

Kernwoorden: wiskunde, kleuteronderwijs, dimensionaliteit, assessment, IRT

Appendix A

Item number	Item name	Version	Content Area	Description	Difficulty	*SE*	WMNSQ	*t*	rit
1	Tot1_r	E, M, D	SNO	Count as far as possible	-2.434	0.21	0.76	-2.4	0.56
2	L2_r	E	SNO	More black or white disks?	-2.224	0.29	1.13	1	0.24
3	MS2_r	M, D	C & R	4 shapes: one is hidden, only partly visible: which one is hidden?	1.904	0.2	1.24	2.2	0.11
4	Tot3_r	E, M, D	C & R	Hands behind fence: how many children are there?	-0.028	0.15	0.91	-1.5	0.51
5	LM4_r	E, M	S & S	Scene setup: who can see what? (perspective taking)	-3.007	0.27	1.04	0.3	0.30
6	LM5_r	E, M	SNO	Spot the (verbal) counting error	-1.258	0.18	0.95	-0.7	0.40
7	S5_r	D	SNO	What comes before 25?	0.846	0.25	0.9	-1.2	0.59
8	LM6_r	E, M	SNO	Ruler with missing numbers: which number belongs here?	-1.131	0.18	0.99	-0.2	0.43
9	S6_r	D	S & S	5 triangles with colored corners in different rotations. Why does one not belong? (colors are changed)	1.808	0.26	1.06	0.5	0.30
10	LM7_r	E, M	M	Measuring 2 children: who is taller?	-2.522	0.23	1.06	0.5	0.21
11	S7_r	D	M	2 lawns. Sheep wants most grass. Decide which lawn to use using measuring tile	1.576	0.29	0.98	-0.2	0.39
12	Tot8_r	E, M, D	SNO	Order dice	-1.314	0.16	0.77	-3.1	0.55
13	Tot9_r	E, M, D	D & C	Read from table with non-symbolic numbers of animals. How many chicken? How many chicken and cows together	-2.006	0.19	1.05	0.5	0.35
14	L10_r	E	S & S	Construct square of puzzle pieces (on top of example)	-1.843	0.27	1.19	1.6	0.23
15	MS10_r	M, D	D & C	Read from table with non-symbolic numbers of animals. How many pigs and cows in total?	-0.677	0.19	0.94	-0.7	0.52
16	Tot11_r	E, M, D	SNO	Cover 4 stones, add 3: how many?	0.666	0.15	0.97	-0.5	0.44

17	LM12_r	E, M	C & R	Divide 6 cookies over 2 people and over 3 people: when do you get more?	-0.472	0.17	1.07	1	0.32
18	S12_r	D	C & R	One egg carton contains 6 eggs. KiKi wants 18. How many cartons? (mc question)	1.878	0.27	0.98	-0.1	0.43
19	L13_r	E	SNO	Which number comes after 6?	-1.352	0.26	0.87	-1.5	0.61
20	MS13_r	M, D	SNO	Lotte has 5 sweets. She gives 2 to Max. How many does she have left?	0.282	0.17	0.93	-1	0.51
21	Tot14_r	E, M, D	M	Chocolate pieces: put in order from small to large	0.173	0.15	0.96	-0.7	0.45
22	LM15_r	E, M	D & C	3 jars with blue and brown candy: Want brown candy, drawn blindfolded: which one to choose?	-0.413	0.17	1.31	4.3	0.12
23	S15_r	D	SNO	Bounded number line (0-10) with a mark in the middle: which number goes here?	1.411	0.25	0.94	-0.7	0.52
24	Tot16_r	E, M, D	M	Chocolatebar: how many cholocate pieces fit in?	1.802	0.18	1.18	1.8	0.19
25	L17_r	E	SNO	Subitizing 3 dots	-3.157	0.36	0.89	-0.4	0.49
26	MS17_r	M, D	SNO	Subitizing 4 dots	-1.377	0.21	0.95	-0.4	0.46
27	Tot18_r	E, M, D	S & S	Choose correct photo based on viewpoint	1.137	0.16	1.23	3	0.24
28	L20_r	E	D & C	Read off amount of collected balls. Who has won the game?	-3.294	0.38	1.14	0.6	0.14
29	MS19_r	M, D	D & C	2 kinds of butterflies are presented after each other: put a tally for each presented butterfly	-1.617	0.23	1.04	0.3	0.33
30	MS19a_r	M, D	D & C	(Based on tally list) How many butterflies of each kind appeared?	-2.534	0.39	0.92	-0.7	0.39
31	Tot20_r	E, M, D	SNO	Count backwards from 8	-0.982	0.16	0.86	-2.1	0.52
32	L21_r	E	SNO	Different objects on sheet: of which object are 2?	-2.321	0.29	0.8	-1.4	0.64
33	MS21_r	M, D	M	Running contest: read off stopwatches. Who won the contest?	1.523	0.19	1.05	0.6	0.38
34	Tot22_r	E, M, D	C & R	Construct stairway pattern with disks from memory (after viewing example)	-0.324	0.15	0.85	-2.5	0.53
35	L23_r	E	SNO	Matching numbers (1 – 4) with amount of sticks	-1.698	0.27	0.82	-1.8	0.63
36	MS23_r	M, D	S & S	Program robot to go to location	1.86	0.2	0.99	-0.1	0.39

37	L24_r	E	M	4 rods: this one is the shortest, which one is the longest?	-3.031	0.35	1.13	0.6	0.15
38	MS24_r	M, D	C & R	4 sandwiches, 8 pieces of cheese. Each sandwich needs 2 pieces of cheese. How many can you make?	-0.286	0.18	0.89	-1.6	0.46
39	Tot25_r	E, M, D	SNO	Show 8 fingers	-1.184	0.16	0.95	-0.6	0.44
40	Tot26_r	E, M, D	S & S	Find 3 triangles on drawing	-1.201	0.16	1.08	1	0.37
41	Tot27_r	E, M, D	D & C	4 wheels of fortune: where is the highest chance to win?	0.293	0.15	1.18	2.9	0.33
42	L28_r	E	SNO	Divide 9 disks over 3 people	-2.073	0.28	1.15	1.1	0.26
43	M28_r	M	SNO	Which number comes before 13?	1.251	0.26	1	0	0.41
44	S28_r	D	SNO	Count in two's: 2, 4……..(until 12)	0.528	0.26	0.89	-1.2	0.59
45	Tot29_r	E, M, D	SNO	6 children, 4 plates: how many plates are missing?	-0.428	0.15	0.92	-1.3	0.54
46	L30_r	E	C & R	Patterning (a,b,a,b,a,b): repeat example from memory	-3.294	0.38	1.02	0.1	0.35
47	MS30_r	M, D	M	2 lines: decide which one is longer using sticks	-1.222	0.25	1.03	0.3	0.27
48	L31_r	E	SNO	Put exactly 4 candles on the cake	-2.153	0.29	0.84	-1.3	0.60
49	MS31_r	M, D	C & R	Patterning (a,b,b,c,c,c,d,d,d,d,d): repeat example from memory	0.712	0.17	1.02	0.3	0.44
deleted	S4r	D	D & C	Three containers with bone black stone and different numbers of black stones: which one to choose for a white stone?					

Note: E = easy, M = medium, D = difficult. SNO = sets, numbers , and operations, C & R = change and relationships, M = measurement, S & S = space and shape, D & C = data and chance. Difficulty = Item difficulty / location parameter, SE = Standard error of item difficulty / location parameter, WMNSQ = Weighted mean square, t = t-value for WMNSQ, r_{it} = Corrected item-total correlation. For the dichotomous items, the item-total correlation corresponds to the point-biserial correlation between the correct response and the total score; for polytomous items it corresponds to the product-moment correlation between the corresponding categories and the total score (discrimination value as computed in ConQuest).