Volume 25, Number 5, October 2017

European Early Childhood Education Research Journal

The Journal of the European Early Childhood Education Research Association

Routledge

ISSN: 1350-293X

Routledge
Taylor & Francis Group

European Early Childhood Education Research Journal

ISSN: 1350-293X (Print) 1752-1807 (Online) Journal homepage: http://www.tandfonline.com/loi/recr20

Early identification of children at risk for academic difficulties using standardized assessment: stability and predictive validity of preschool math and language scores

Niek Frans, Wendy J. Post, Mark Huisman, Ineke C. E. Oenema-Mostert, Anne L. Keegstra & Alexander E. M. G. Minnaert

To cite this article: Niek Frans, Wendy J. Post, Mark Huisman, Ineke C. E. Oenema-Mostert, Anne L. Keegstra & Alexander E. M. G. Minnaert (2017) Early identification of children at risk for academic difficulties using standardized assessment: stability and predictive validity of preschool math and language scores, European Early Childhood Education Research Journal, 25:5, 698-716, DOI: 10.1080/1350293X.2017.1356524

To link to this article: http://dx.doi.org/10.1080/1350293X.2017.1356524

Published online: 01 Aug 2017.

EUROPEAN EARLY CHILDHOOD EDUCATION RESEARCH JOURNAL, 2017
VOL. 25, NO. 5, 698–716
https://doi.org/10.1080/1350293X.2017.1356524

OPEN ACCESS

Early identification of children at risk for academic difficulties using standardized assessment: stability and predictive validity of preschool math and language scores

Niek Frans[a], Wendy J. Post[a], Mark Huisman[b], Ineke C. E. Oenema-Mostert[a,c], Anne L. Keegstra[d] and Alexander E. M. G. Minnaert[a]

[a]Faculty of Behavioral and Social Sciences, Department of Special Needs Education and Youth Care, University of Groningen, Groningen, the Netherlands; [b]Faculty of Behavioral and Social Sciences, Department of Sociology, University of Groningen, Groningen, the Netherlands; [c]School of Education, Stenden University of Applied Sciences, Leeuwarden, the Netherlands; [d]Otolaryngology Department, University Medical Center Groningen, Groningen, the Netherlands

ABSTRACT

Despite the claim by several researchers that variability in performance may complicate the identification of 'at-risk' children, variability in the academic performance of young children remains an undervalued area of research. The goal of this study is to examine the predictive validity for future scores and the score stability of two widely administered Dutch preschool tests. Specifically, the focus was on their suitability for identifying children that are at risk for academic difficulties. To evaluate at-risk identification using early standardized assessment, language and math scores were collected over a four-year period (*N* = 431). Score stability was evaluated by means of transition rates and score differences. Predictive validity was assessed using a mixed model. The majority of low-scoring children showed broad fluctuations in scores, although 12% to 17% did remain relatively stable in their scores. Correlations between preschool scores, and first- and second-grade language and math measurements were estimated at between .09 and .30. The longitudinal design of this study illustrates how assessment scores can fluctuate over time, which is a problem that may be inherent in this age group but one that warrants greater attention. This study provides a transparent evaluation of the suitability of assessments used for identifying children at risk for academic difficulties.

KEYWORDS
Stability; predictive validity; standardized testing; early childhood; education

Ideally, assessment instruments provide information that informs educators in their decisions about a child's instructional needs. An important function in this process, and one that is often ascribed to (standardized) assessment, is early identification of children who are deemed 'at risk' for academic or developmental problems (Heckman 2000; Snow 2006). In this respect, the general belief is that gathering objective information has as its merit the prevention of future academic problems by identifying these problems at an

CONTACT Niek Frans N.Frans@rug.nl Faculty of Behavioral and Social Sciences, Department of Special Needs Education and Youth Care, University of Groningen, Grote Rozenstraat 38, Groningen 9712 TJ, the Netherlands

early stage (Abu-Alhija 2007; Leseman 2004). Studies have shown that early intervention programs yield impressive results both from an economic and a social perspective (Heckman 2000), a finding that underscores the importance of early identification.

Although the potential benefits may be high, several scholars argue that the inherent variability in performance of a young child and among young children as a group (intra- and inter-individual variability), and the lack of stability in the way young children demonstrate their competence, does not allow for a reliable or valid assessment of current or future performance using standardized tests (Colpin et al. 2006; Gilliam and Frede 2012; Shepard et al. 1998). Indeed, studies indicate that, for many early assessment measurements, the relationship between test scores and future outcomes is consistently inadequate (Dockrell and Marshall 2015), widely varying (Kim and Suen 2003), or unclear (Heckman 2000; Snow 2006). The predictive validity for future outcomes, however, is imperative when using assessment measurements to inform decisions (Cronbach 1971).

Although large inter- and intra-individual variation can be problematic when trying to identify children for intervention purposes, both inter- and intra-individual variations have been largely neglected in studies of cognitive abilities (Siegler 2002; Zubrick, Taylor, and Christensen 2015). This is in stark contrast with the fields of motor, social, and emotional development, where developmental stability has received more attention. For example, a study by Darrah and Hodge (2003) concerning the stability of motor and communication abilities, shows that infants make large shifts in percentile rankings on standardized tests (Peabody Developmental Motor Scales, Communication Symbolic Behavior Scales). The majority of infants in their study showed unstable patterns in their scores over time: while a large portion of infants (61%) scored below the cut-off 16th percentile, most infants did so only once. These results show that, depending on the moment of assessment, decisions made based on any single score can lead to very different conclusions. Goorhuis and Schaerlaekens (2000) also indicate that normal variation in language development is often diagnosed as a developmental problem and treated accordingly. They plead for a more thorough distinction between normal variation and maturation, on the one hand, and developmental problems and disorders, on the other. From a neurological perspective, the sizable variation in early math and language skills is consistent with the rapid development of memory and attention processes that underlie these skills in early childhood (Fuchs et al. 2014; Geary 2006; Goorhuis and Schaerlaekens 2000; Shonkoff and Phillips 2000). Although this issue of stability may be inherent to the development of young children, it may hinder educational decisions based on any single assessment outcome, as scores are generally less reliable.

Distinguishing between children at risk for academic problems and normal developmental variation requires the assessment outcome to be strongly indicative of the child's educational trajectory. Correlation coefficients are often reported in order to evaluate this property and justify the use of assessment instruments for screening and intervention purposes (Einarsdóttir, Björnsdóttir, and Símonardóttir 2016; Kim and Suen 2003). However, although correlations provide important information as to a test's average predictive validity for the entire range of scores, they might not adequately represent a test's adequacy in detecting children at risk for academic problems. Consequently, although correlation coefficients between early and later academic measurements are important, they might not adequately justify use of an assessment instrument for identification purposes.

Research into the predictive validity of early childhood assessments indicates that most early language measurements correlate only moderately with later test scores, while tests of early math skills fare only slightly better. For example, analyses of six data sets (N~10,000 teachers and 16,000 children) by Duncan et al. (2007) showed that preschool mathematics and language abilities at age five are significant predictors of later achievement, although the standardized coefficients of early language scores ($\beta = .17$) were considerably smaller than the coefficients of early math scores ($\beta = .34$). A replication study ($N = 1521$) by Romano et al. (2010) indicated slightly weaker correlation coefficients both between preschool and first/third-grade standardized mathematics assessments, and between preschool and first/third-grade language teacher/parent report measurements. A review study by La Paro and Pianta (2000) on the relationship between preschool academic and social assessments, and second-grade academic and social test scores, revealed that preschool academic assessments make only small to moderate contributions when it comes to predicting school success. The correlation coefficients collected from over 30 studies ranged between .08 and .78, with a mean correlation coefficient of .43 and .48 for first and second grade, respectively.

All three studies contributed correlation estimates over a large number of subjects and assessment instruments. Although similar correlation estimates were found for all three studies, they differ markedly in terms of the optimism of their conclusions. While Duncan et al. (2007) and Romano et al. (2010) stress the significance of early academic measurements as strong and important predictors of later math and reading scores, La Paro and Pianta (2000) conclude that ‘child-based assessment of skills will not accurately identify “high risk” children’ (p. 476). These statements illustrate how a focus on performance prediction in the general population can lead to more optimistic conclusions when compared to a focus on identification of children at risk for academic difficulties. Both interpretations are united in a paper by Dollaghan and Campbell (2009), who studied the relationship between several language measurements at ages three, four, and six ($N = 414$). Dollaghan and Campbell concluded that, while early language tests correlate moderately with later test scores on a group level (r between .35 and .77), on an individual level, low early language scores (defined as 1.5 *SD* below sample mean) were poor predictors of later language deficits. For the test with the highest correlation coefficient (PPVT-R), only 17% of low-scoring children remained consistently within this group over time.

The study by Dollaghan and Campbell (2009) shows the importance of combining group statistics, such as correlations with a more specific focus on individual scores over time, and, more specifically, a focus on the identification of children at risk for language and math difficulties. However, like the studies conducted by Duncan et al. (2007), Romano et al. (2010), and La Paro and Pianta (2000), the study by Dollaghan and Campbell is limited in this respect due to the focus on bivariate comparisons instead of longitudinal score trajectories. In addition, although the research by Dollaghan and Campbell examines whether early low scores result in an increased risk for later academic difficulties, little attention is paid to children who scored high on the early assessment but received low scores in subsequent years. Granted that falsely identifying children as being at risk (false positives) may be considered ineffective or even unethical, the occurrence of ‘false negatives’ may prove even more serious in assessment applications, since they would be indicative of children who did not receive the support needed. Finally, the analytical techniques used to mitigate the occurrence of missing data in these studies are prone to induce bias in the parameter estimates.

To summarize, although all these studies add valuable information, their utility in identifying children at risk for academic difficulties is limited by the lack of focus on this specific group. In addition, missing data can be handled using more effective and efficient methods that better limit bias caused by selective testing. Finally, all these studies share the implicit assumption that the score trajectories between measurements are stable, by restricting their comparisons to two measurement occasions instead of longitudinal score trajectories.

This study provides a new perspective on the evaluation of preschool assessment by combining group statistics with a specific focus on the individual score trajectories of low-performing children. In addition, the current study assesses the stability of scores over time, with special attention paid to children who score in the lower regions on early and/or later academic measurements. For the purpose of this study, score stability can be defined as the consistency between measurement occasions, as measured by the percentile ranking relative to the general population and to previously obtained scores.

This study aims to evaluate the utility of early standardized assessment for identifying children at risk for later academic difficulties. In this explorative study, we will analyze data originating from a Dutch educational context. Hence, the following research questions will be answered in terms of the preschool[1] tests used by the majority of Dutch elementary schools: 1. What is the degree of stability of language and math achievement scores? 2. What is their predictive value for future language and math scores?

Method

Population and sample

The target population consists of children in the first four years of Dutch primary school, which administered tests from the Student and Education Monitoring Program developed by Cito (*Leerling- en Onderwijs Volgsysteem*, LOVS). A selective sample of 18 Dutch regular primary schools has been used for this study. Within these schools, all children who started third grade in 2013 and were tested at least once have been included in the sample. On average, these children each took 5.8 language tests (SD 1.4) and 5.5 math tests (SD 1.3). Three children, who received special educational needs funding, were excluded from the sample, since these children were already known to be at risk for academic difficulties, and the low number of these children made generalization of study findings for this specific subpopulation difficult. The total sample consists of 431 children, with a mean age of 8 years and 2 months (SD 5.2 months) when the final test was administered.

The sample characteristics for the independent variables are given in Table 1. As shown, the sample contains roughly the same number of boys and girls and consists

Table 1. Sample demographics.

N	431
Girl (%)	52.90
Age (yr.; mo.)	8;2
Foreign background (%)	7.00
Low parent educ. (%)	4.40
Very low parent educ. (%)	1.90
Oldest (%)	58.70
Observed Language (%)	72.60
Observed Math (%)	68.80

primarily of native Dutch children (7% have a foreign background). Overall, around one-third of the measurements on the dependent variables are missing.

Instruments

The instruments that are assessed in this study were developed by the Dutch Central Institute for Test Development (Cito). In conjunction with teacher observations, these instruments are designed to provide information for both identification/allocation and evaluative decisions (Koerhuis and Keuning 2011; Lansink and Hemker 2012). All items are formulated by a panel of assessment experts, teachers, and educational professionals, and are assessed using a one-parameter logistic model on large samples of elementary school children (Verhelst, Verstralen, and Eggen 1991). This Item Response Theory (IRT) model is identical to the two-parameter Birnbaum model, where the discrimination indices are estimated in advance by a weighted least squares algorithm and subsequently treated as known constants (Verhelst, Verstralen, and Eggen 1991). The construct validity of each test was examined through the fit of the items to the IRT model and correlations with an older version of the preschool test for the preschool and grade 1/2 tests respectively.

All the instruments were found to have satisfactory properties by the Dutch Committee for Test Materials (COTAN), an independent committee that evaluates test construction, quality of the materials, norms, reliability, and construct validity (COTAN 2011, 2013). The predictive validity for these instruments, however, has not been assessed. Two different versions of the preschool tests are currently in use: a version from 1996 and a revised version from 2009. Both versions measure the same construct, and previous studies indicate that the item banks, on which the instruments are based, correlate highly for both the language tests ($r = .92$; as found by Lansink and Hemker 2012) and the mathematics tests ($r = .99$; as found by Koerhuis and Keuning 2011).

The preschool language instruments (Lansink and Hemker 2012) are designed to measure receptive language ability. The instrument administered in the first year (equivalent to US preschool year, and abbreviated to Middle (M)1 and End (E)1 in this study) consists of 48 items with a maximum score of 97 designed to assess the child's receptive vocabulary, word definition skills, and understanding of written and spoken language. Phonological awareness and metalinguistic tasks are added to the second-year test (equivalent to US kindergarten year and abbreviated to M2 and E2), which consists of 60 items and a maximum score of 108. Reliability was assessed with Measurement Accuracy (Verhelst, Glas, and Verstralen 1995) and ranged from .84 to .89.

The language tests (De Wijs et al. 2010), administered in grades 1 and 2 (M3 to E4), consist of 50 items with a maximum score of 124 to 151 (see Table 2), which measures a child's ability to correctly spell a word and to recognize a wrongly spelled word. The tests consist of written assignments, though the module for the better spellers in second grade consists of multiple-choice items. Reliability for these instruments ranges between .90 and .94.

The mathematics preschool tests (Koerhuis and Keuning 2011) are designed to measure general early mathematics mastery. These instruments include 46 to 48 items (maximum score 106 and 137, respectively) that assess the child's numerical understanding; understanding of quantity; understanding of basic concepts related to location, length,

Table 2. Test score descriptives per measurement occasion.

	Language				Mathematics			
	Mean (SD)		*n* (%Tot)	Max[a]	Mean (SD)		*n* (%Tot)	Max[a]
M1	57.0	(9.31)	137 (32)	97	44.0	(9.68)	93 (22)	106
E1	61.9	(10.44)	185 (43)	97	50.3	(13.99)	105 (24)	106
M2	68.9	(9.33)	346 (80)	108	74.3	(17.88)	337 (78)	137
E2	75.4	(12.58)	172 (40)	108	79.9	(18.82)	177 (41)	137
M3	108.1	(5.22)	406 (94)	124	36.3	(15.01)	406 (94)	81
E3	115.4	(5.87)	412 (96)	135	47.7	(13.22)	406 (94)	88
M4	121.7	(6.97)	422 (98)	141	54.7	(14.48)	423 (98)	102
E4	123.5	(7.79)	424 (98)	151	64.7	(14.49)	426 (99)	109

[a] Theoretical maximum for reference purposes.

volume, weight, and time; and understanding of figures and simple symmetrical patterns. Reliability ranges between .87 and .91.

The grades 1 and 2 (M3 to E4) mathematics tests (Janssen et al. 2010) consist of 50 and 52 items with a maximum score between 81 and 109 (see Table 2), which are designed to measure applied math skills, including: number knowledge and basic operations (addition, subtraction, multiplication, and division); ratios, fractions, and percentages; and measurement, time, and money (these latter two are added in the second grade). Unlike the preschool tests, these tests consist of open-ended questions. Reliability ranges between .91 and .93.

Data collection and variables

Data was collected and anonymized by the school's secretary. Informed consent was given by the schoolboard to retrospectively retrieve data from a four-year period from the schools' student monitoring systems. As the data were retrieved from an existing database, the study did not interfere with the education of individual children. Furthermore, names were not collected and birthdates were rounded to the nearest month to ensure that data were not traceable to individual students. Ethical approval for this study was given by the University of Groningen Educational Sciences ethics committee.

A total of eight different measurements (language and mathematics tests) are taken, one in the middle (M) of each school year, and one during the end (E) of each school year. The Cito preschool tests are used for the first four measurements, with a dummy variable to indicate the test version. The last four measurements include the language and math scores in first and second grade. It is important to note that the preschool tests and first/second grade tests constitute two distinct (albeit related) measurements, with notably different scales for the continuous weighted (item-response function) scores (Koerhuis and Keuning 2011; Lansink and Hemker 2012). This means that absolute differences do not have the same meaning over the four-year period, which is why measurements of correlation need to be used. In addition, the continuous scores can be expressed in percentile quartiles[2] relative to the general population of children in Dutch elementary schools. All instruments are group-administered in two parts by the classroom teacher in roughly 20 to 40 minutes. Whenever a child had repeated a grade ($n = 85$), the second score was used when that child had been tested twice using the same test (occurred in 30.2% of the cases that repeated a grade), since testing effects were presumed to be negligible due to the long intervals between these two tests. In addition, several variables that are known to influence learning outcomes were measured, including: whether the child had a foreign background (i.e. a non-Dutch parent, *NNCA*; see Rovict n.d.), the gender of the child, whether the child was the first child in a family to attend a specific school, and the child's age in months at the start of third grade (OECD 2008). The educational level of the child's caregiver(s) was also measured in three categories in accordance with the 'educational burden.' This procedure is designed to assign extra school funding for children whose parents or caregivers had only graduated from the lowest track of high school education (US equivalent: ≤10th grade), and for children where at least one parent had discontinued his/her studies after primary school (DUO 2013). We will refer to these two categories as 'low parent education' and 'very low parent education,' respectively. The third category includes all other children, where at least one of their parents had finished his/her junior year in high school.

Since schools are not obligated to test at every measurement occasion during the preschool years, missing observations were likely to arise. Statistical analyses were therefore used to compensate for the occurrence of bias in parameter estimates due to missing test scores.

Statistical analyses

First, sample descriptives are presented for the demographic variables. In addition, means and standard deviations for the language and math scores are calculated for each measurement occasion, along with the number of observations. For explorative purposes, correlations between the measurement occasions for the language and math tests are calculated using pairwise deletion to treat missing data. For subsequent analyses, missing data is handled using multiple imputation (Rubin 1987), because deleting cases with missing data from the analyses is generally wasteful of information and is known to generate biased parameter estimates if the causes of missingness are excluded from the analyses (Allison 2009; Graham 2009; Van Buuren and Groothuis-Oudshoorn 2011). Multiple imputation procedures work by replacing each missing value m times with an adequate estimate based on the available information in the dataset and an added random residual (Graham 2009). These estimates consist of simulated random draws from the posterior missing data distribution and result in m different datasets, which are subsequently analyzed separately to obtain m parameter estimates. The results of these analyses are pooled using Rubin's rules (Rubin 1987) to create unbiased parameter estimates and standard errors when the MAR assumption holds (Graham 2012). According to Graham (2009, 2012), MI estimates are generally superior to older methods even when the MAR assumption is violated.

A major benefit of any MI technique is that it separates the estimation of missing values from the actual analyses. Essentially, MI works with an imputation model that uses all available information to impute missing data, which may be different from the substantive model that only uses variables of substantive interest to the researcher. This means that the imputation model can be larger than the substantive model by including auxiliary variables, which makes a MAR assumption more tenable in comparison to ML models that often only include variables that are part of the analysis model (Graham 2012).

To identify the missing data mechanisms and determine possible sources of bias, the mean scores in each pattern of missing data are visualized and compared to the complete case scores. Missing observations are multiply-imputed, using the R package MICE V2.0 (Van Buuren and Groothuis-Oudshoorn 2011). Multiple Imputation by Chained Equations (MICE) is an imputation technique that specifies a separate univariate imputation model for each partially observed variable (Van Buuren and Groothuis-Oudshoorn 2011; White, Royston, and Wood 2011). This makes it an extremely versatile technique that allows for the imputation of both normally and non-normally distributed variables. In addition, the software has been adapted to impute hierarchically structured data. All available variables are used in the imputation models (both as predictors and outcomes) to generate 50 complete data sets, which should be sufficient to alleviate relative efficiency and power problems (see Graham 2009, 2012). The variables include all demographic variables as well as available (continuous) language and math scores. The categorical scores based on percentile groups are not included in the imputation models but are derived from the continuous scores after imputation.

The stability of the individual scoring sequences is determined by analyzing the percentile groups with the TraMineR package for sequence analyses (Gabadinho et al. 2011). Each child who achieved a score below the 25th percentile is grouped according to two events, namely a switch from a ≤25th percentile score in the preschool years to a >50th percentile score in first/second grade, or a switch from a >50th percentile score in preschool to a ≤25th percentile score in first/second grade. Both of these events signify score changes larger than 25 percentile points between preschool tests and first/second grade tests, and are therefore flagged as 'large switches.' The test version will be taken into account, since the classification into percentile groups differs for the two versions of the preschool tests. In addition to the occurrence of these events for individual children, the conditional probability of shifting percentile groups between two consecutive measurements (i.e. transition rates) is also calculated for the entire sample.

Finally, the predictive validity is assessed with a multilevel model fitted to the imputed data. After a fully multivariate model is constructed, the model is reapplied to the imputed datasets, where the covariance matrix of the different measurement occasions allows for estimation of the between-test correlations. Any child demographics that show a significant relationship with the predicted score in the original data are added to the model as fixed effects. Parameter significance testing is done following the procedure described in Snijders and Bosker (2012, pp. 94–95), with a significance level set at .05.

Results

Sample descriptives

As shown in Table 2, the majority of the missing observations occur in the first two years (M1to E2). Roughly a quarter (language) to a third (math) of the missing observations in the first year are missing because the schools chose not to administer the test at this point. Although the schools that did not administer these tests did not differ significantly in terms of the percentage of students from low-educated or one-parent households or the age of the children attending the school, those schools that did not administer the first-year language tests contained relatively few students with a foreign background (~2% vs. 10%, $p < .05$). In addition, the mean scores and standard deviations for each measurement occasion are shown in Table 2 for both language and mathematics tests. The large discrepancy between E2 and M3 is a result of the two different measurement scales used for the preschool tests and first/second grade tests.

Missing data

Figures 1 and 2 show the sequences of mean scores (±2 SE) for each pattern of missing observations on the language and math tests, respectively. For each figure, the box labels indicate the identification of the missing data pattern, where a 0 indicates an observed score and a 1 indicates a missing value. For example, in the upper right box in Figure 1, the identification ‘0:1:0:1:0:0:0:0’ indicates the following pattern: observed M1, missing E1, observed M2, missing E2, observed M3, E3, M4, E4. Inside each box, the number of children with that particular pattern of observed scores and a plot of their mean scores are presented. Nineteen missing data patterns that occur more than once were found in the language scores, and eighteen in the math scores (range $n = 2$ to 100). The mean language scores only differ slightly between the patterns with missing values and the complete cases ($n = 31$). For example, slightly higher mean scores in the preschool measurements are seen for the middle box in the second row of Figure 1 ($n = 64$), whereas the last box on the first row ($n = 55$) shows slightly higher scores for both the mean preschool tests and subsequent tests.

The math scores show that the preschool scores for the complete cases are generally much lower than the observed scores for cases with missing values. For example, the boxes in the second ($n = 24$) and fourth ($n = 100$) columns of the second row of

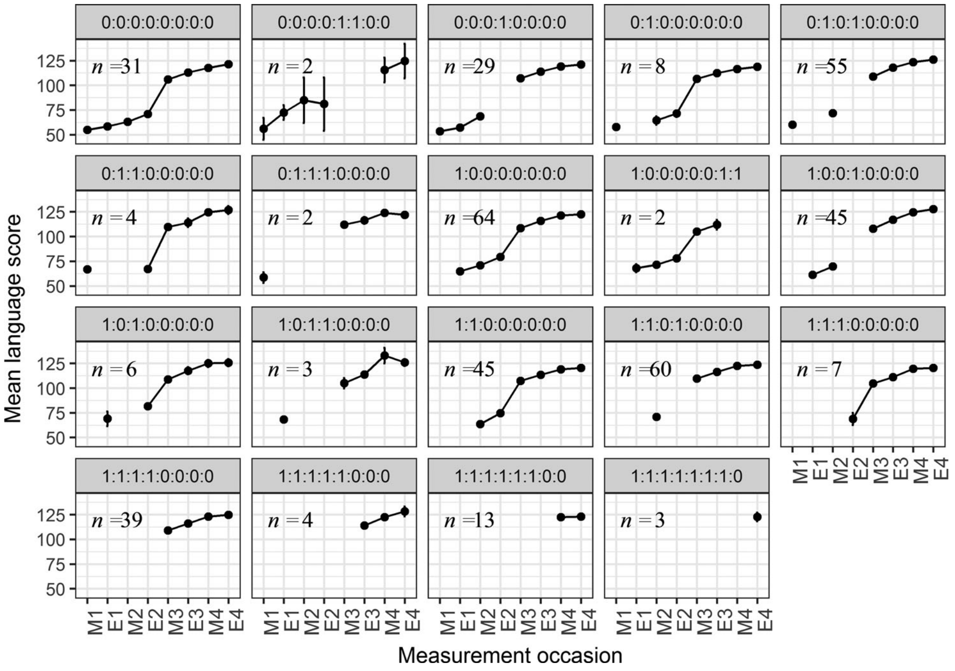

Figure 1. Mean scores for the language tests (*y*-axes) per measurement occasion (*x*-axes), split by missing data pattern (headers, 0 = observed, 1 = missing).

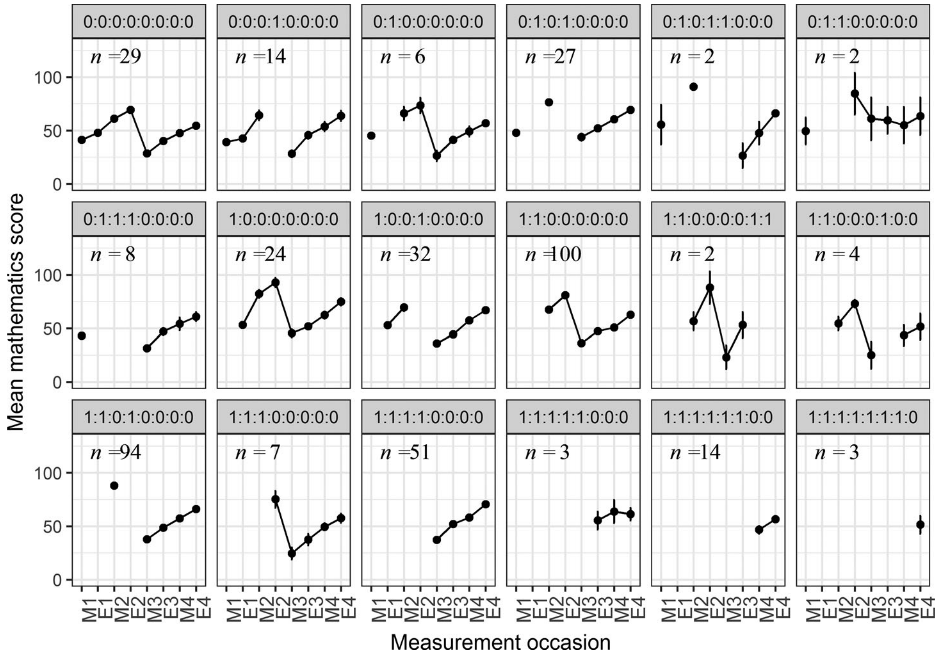

Figure 2. Mean scores for the mathematics tests (*y*-axes) per measurement occasion (*x*-axes), split by missing data pattern (headers, 0 = observed, 1 = missing).

Figure 2 show much higher mean scores on the preschool tests compared to the complete cases ($n = 29$). Furthermore, the subsequent test scores appear to be much higher, on average, for these missing data patterns. This difference in subsequent scores can also be seen for the third box on the last row ($n = 51$), which has no measurements for the preschool tests. These results indicate that missingness seems to be related to the language and math scores of the child. Specifically, children with higher scores at later measurements are more likely to have missing data in preschool years, which is indicative of selective testing by schools. This relationship between later scores and missingness in preschool is likely to bias parameter estimates if it is not accounted for in further analyses. By including this relationship in the imputation model, this bias can be mitigated.

Score stability

The score sequences of children who scored ≤25th percentile at any point during the four-year period were grouped according to the description shown in Table 3 for both language ($n = 143$) and math scores ($n = 101$). Contrary to other analyses, missing scores in these analyses are not imputed but assigned a 'missing' category, because grouping is done according to characteristics of individual sequences. Conditional on their scores in preschool and their subsequent scores in first/second grade, children would receive the label *Up*, *Down*, *Fluctuating*, *Missing*, or *Stable*. Most of these children, labeled as 'Down,' switched from one or more above-average scores in preschool years to one or more ≤25th percentile scores in first/second grade. On average, this group scored

Table 3. Conditions used to cluster children that received at least one ≤25th percentile score on the language ($n = 143$) or math ($n = 101$) tests, and percentages in each group.

		%	
Group label	Definition	Language	Math
Down	Child moves from a >50th percentile score in preschool to a ≤25th percentile score in first/second grade at least once, and child does not score ≤25th percentile within preschool.	47	35
Up	Child moves from a ≤25th percentile score in preschool to a >50th percentile score in first/second grade at least once, and child does not score >50th percentile within preschool.	8	10
Fluctuating	Child has both >50th percentile and ≤25th percentile scores in preschool, but would otherwise be categorized as either Up or Down.	25	30
Stable	Child does not switch from a >50th percentile score in preschool to a ≤25th score in first/second grade or vice versa. That is, no large fluctuations within preschool, or between preschool and first/second grade.	12	17
Missing	Child has no observed values in preschool or subsequent years useable for categorizing.	8	9

below the 25th percentile in the first/second grade on 1.7 and 1.6 out of 4 measurement occasions for language and mathematics, respectively ($Mdn = 1$).

A relatively small group, labeled 'Up,' made an inverse switch from one or more ≤25th percentile scores in preschool years, to one or more above-average scores in first/second grade. This group scored above the 50th percentile in first/second grade on an average of 2.6 and 2.2 out of 4 measurement occasions for language and mathematics respectively ($Mdn = 2$), and did not score ≤25th percentile in first/second grade.

The second largest group of children, labeled as 'Fluctuating,' showed one or more >50th percentile- and ≤25th percentile scores within the preschool years. In contrast, only a small group showed no large fluctuations in scores between the preschool years and first/second grade, instead remaining more or less in the lower scores. It is worth mentioning that this group is relatively larger for those children tested with the new version of the test at M2, as compared to children tested with the old version of the test. A child that was tested with the new version of the mathematics test at M2 is 5.5 times more likely to belong to the 'Stable' group than a child tested with the old version. For language, a child tested with the new version is 1.8 times more likely to belong to the stable group. On the other hand, these children are 3.2 and 1.7 times less likely to belong to the 'Down' group for the language and math tests, respectively.

Table 4 shows the transition rates between two consecutive scores for the entire sample. Each cell gives the proportion of children that moved from a percentile group at time t (columns) to a percentile group on the next measurement time $t + 1$ (rows) along with the estimated standard error in brackets. As shown by the diagonal transition rates (i.e. children that remain in the same percentile group), the >75th percentile scores are generally the most stable, whereas the other scores tend to show stability rates that are around half as large. In addition, it is also apparent that children are more likely to increase in score than they are to decrease or stay within the same score.

When focusing on the occurrence of the two large switches mentioned in Table 3 (i.e. Down and Up), one can see that, between two consecutive measurements, children with a ≤25th percentile score switch to an above-average score around 33% and 48% of the time for language and math, respectively. The switch from an above-average score to a ≤25th percentile score, however, occurs around 12% and 29% between two consecutive

Table 4. Estimated transition rates and standard errors for language and mathematics.

		Percentile score at time = *t*							
Language		100–75		75–50		50–25		25–0	
	100–75	.61	(.024)	.42	(.026)	.19	(.021)	.16	(.022)
Perc. score at time = *t* + 1	75–50	.22	(.021)	.29	(.024)	.27	(.023)	.17	(.022)
	50–25	.12	(.016)	.22	(.021)	.40	(.024)	.27	(.024)
	25–0	.06	(.012)	.06	(.013)	.14	(.018)	.40	(.028)
Mathematics									
	100–75	.61	(.024)	.36	(.024)	.17	(.021)	.31	(.026)
Perc. score at time = *t* + 1	75–50	.19	(.019)	.31	(.023)	.34	(.025)	.17	(.020)
	50–25	.07	(.013)	.18	(.019)	.30	(.023)	.19	(.021)
	25–0	.14	(.018)	.15	(.019)	.19	(.021)	.33	(.025)

measurements. Although children are most likely to switch from above-average scores in preschool to ≤25th percentile scores in subsequent years, the reverse is more likely to occur between two consecutive measurements. When focusing only on the new version of the test, the largest difference in language transition rates is .02 compared to the transition rates over both versions; for mathematics the maximum difference is .06.

Predictive value

In order to compensate for any bias due to missing data, and/or confounding variables, the values below the diagonal were estimated with a multilevel model on the multiply imputed (MI) scores. The multilevel model controls for any fixed factors that show a significant effect in the dataset without imputations. These fixed effects included the parent-education variable, gender, test version, and foreign background. For the sake of completeness, the fixed effect coefficients and standard errors of the model for language scores are included in the first two columns of Table 5. As in a linear regression model, these coefficients indicate the average test score of children for every measurement time as well as the average effect of included variables on these scores. After imputation, the only effect that remained statistically significant was gender: on average, girls score 1.9 points higher than boys. Measurement occasion was included in the model using a fully multivariate model, as illustrated in Snijders and Bosker (2012, pp. 255–260). For example, an average-scoring Dutch boy of whom at least one parent finished the junior year of high school scores an estimated 62.77 on measurement occasion M1 when tested with an old version of the language test.

Table 6 shows the estimated correlation matrix for the language scores. The above diagonal values are the correlation estimates of the observed data, where missing values were handled using pairwise deletion. The pairwise deletion correlations are generally highest within the first and second grade (M3 to E4), and within the preschool years (M1 to E2), with most correlations in the range of .50 to .70. Between the preschool years and first/second grade, correlations are markedly smaller, ranging between .11 and .32. Similarly, the MI estimated correlations are highest in the first and second grades, and overall lowest between the preschool years and first/second grade. The last estimates mentioned range between .09 and .30, with an average correlation of .20, which is similar to the pairwise estimates.

For preschool years, there are large differences between correlations based on multiple imputations and correlations based on pairwise deletion. Most correlations drop in

Table 5. Fixed effects and standard errors for multilevel models on imputed data.

Model language scores			Model mathematic scores		
Fixed effects	Coefficient (SE)		Fixed effects	Coefficient (SE)	
Measurement M1	62.77	(1.918)	Measurement M1	50.93	(3.347)
Measurement E1	63.27	(1.533)	Measurement E1	49.63	(3.693)
Measurement M2	71.75	(1.343)	Measurement M2	73.77	(2.534)
Measurement E2	83.38	(1.458)	Measurement E2	70.32	(2.867)
Measurement M3	106.92	(1.488)	Measurement M3	49.30	(3.906)
Measurement E3	114.33	(1.502)	Measurement E3	60.01	(3.965)
Measurement M4	120.78	(1.470)	Measurement M4	67.20	(4.064)
Measurement E4	122.82	(1.492)	Measurement E4	77.37	(4.055)
Low parent educ.	−3.40	(2.395)	Repeated a grade	−7.18	(6.019)
Very low parent educ.	−3.41	(3.112)	Foreign background	−9.34	(2.754)*
Gender (Girl)	1.91	(0.754)*	Single parent	−5.48	(4.557)
Foreign background	−1.01	(1.463)	New test	−11.78	(4.747)*
New test	−0.01	(1.973)	E1 × Repeated a grade	7.91	(7.702)
			M2 × Repeated a grade	11.70	(10.427)
			E2 × Repeated a grade	9.75	(11.579)
			M3 × Repeated a grade	14.73	(7.436)*
			E3 × Repeated a grade	15.70	(6.714)*
			M4 × Repeated a grade	19.84	(7.398)*
			E4 × Repeated a grade	10.51	(7.437)

*Significant coefficient at $p < .05$.

magnitude by about half and range between .29 and .47. Since most of the missing data occurs within the preschool years, large differences are more likely to occur there. Further interpretation of these results is provided in the discussion.

Table 7 shows a similar table for the mathematics scores as for the language scores. The multilevel model of the MI estimated math scores included the fixed effects from test version, foreign background, single-parent household, and the interaction effect of measurement occasion and grade retention. Both the interaction effect between time and repeating a grade and the negative effect of foreign background remained significant after imputation. The coefficients and their standard errors are shown in the third and fourth columns of Table 6. Similar to the language scores, the mathematics pairwise deletion correlations are largest within the first and second grades (M3 to E4), and lowest between the preschool years and the first/second grade.

The correlations between the preschool years and first/second grade have estimated values between .16 and .39, when missing data are handled with pairwise deletion, and between .13 and .30 for the imputed data. Again, the correlations within the preschool

Table 6. Observed correlations of language scores (pairwise deletion, above diagonal) and MI estimated language correlations from multilevel model (below diagonal).

	M1	E1	M2	E2	M3	E3	M4	E4
Language M1	–	.65	.43	.53	.21	.11	.13	.14
Language E1	.29	–	.59	.63	.22	.18	.14	.15
Language M2	.42	.31	–	.78	.17	.23	.21	.18
Language E2	.34	.30	.47	–	.32	.32	.28	.19
Spelling M3	.17	.24	.18	.23	–	.60	.53	.49
Spelling E3	.12	.21	.19	.26	.60	–	.66	.63
Spelling M4	.14	.16	.25	.30	.52	.62	–	.81
Spelling E4	.09	.14	.21	.27	.46	.59	.79	–

Table 7. Observed correlations of math scores (pairwise deletion, above diagonal) and MI estimated mathematics correlations from multilevel model (below diagonal).

	M1	E1	M2	E2	M3	E3	M4	E4
Mathematics M1	–	.51	.49	.54	.37	.21	.25	.39
Mathematics E1	.15	–	.17	.45	.23	.16	.18	.17
Mathematics M2	.57	.09	–	.63	.38	.27	.30	.28
Mathematics E2	.34	.22	.40	–	.37	.31	.35	.36
Mathematics M3	.21	.26	.29	.24	–	.68	.60	.63
Mathematics E3	.16	.30	.21	.19	.60	–	.61	.61
Mathematics M4	.15	.28	.24	.19	.58	.61	–	.77
Mathematics E4	.13	.25	.21	.19	.58	.59	.77	–

years show higher estimates, where the imputed data result in lower correlates (.22 on average) than missing data with pairwise deletion (.29 on average).

Discussion

This study was set up to evaluate the utility of early standardized assessment for identifying children at risk for later academic difficulties. In line with the study by Dollaghan and Campbell (2009), the results showed that only a small proportion of students identified as at risk remained in this group in consecutive years, while a large group showed wildly fluctuating scores and a small group moved from bottom-range to above-average scores. Moreover, the results indicate that a large number of low-achieving children are not identified as such in preschool years. These 'false negatives' receive little attention in the study by Dollaghan and Campbell (2009), but failure to identify children at risk for academic difficulties might constitute a more serious problem than wrongly identifying children as 'at risk.' The overall transition rates reveal that, while the top scores (>75th percentile) are generally stable over time, the other scores show far lower consistency. The higher stability of the top scores could be a result of the generally higher probability of progressing to a better score, as opposed to the probability of regressing to a lower score.

Both the language and math imputed preschool scores show small to moderate correlations in the range of .09 and .30 with first/second grade achievement. These correlations are slightly lower than the average correlations found by La Paro and Pianta (2000) and Duncan et al. (2007). However, the differences between the coefficients are small and fall within the range of values that are included in both meta-analyses. The low correlations within the preschool years, and between preschool and first/second grade, might be indicative of the large intra-individual variation in the test scores of these children, especially since the correlations between first and second grade appear to be much stronger.

A noteworthy change between the imputed and pairwise deletion correlations is the drop in correlation magnitude within the preschool years. These differences can be explained by the combination of the large number of missing values within the preschool years, and bias in the pairwise deletion correlations due to selective testing. Indeed, Figures 1 and 2 indicate that high-performing children are less frequently tested in preschool. In addition, schools that did not administer the test in preschool had fewer children with a foreign background. This could imply that observed values in preschool are downward biased (i.e. contain more low-scoring children), meaning that deletion would lead to biased results. Indeed, subsequent analyses by Duncan et al. (2007), using multiple imputation, show a similar drop in coefficients. Unfortunately, the variables used in the multiple imputation procedure, the number of iterations and imputations, and the convergence of the parameter estimates in their study are unclear.

While the preschool tests are said to measure the prerequisites for later language and mathematics skills (Koerhuis and Keuning 2011; Lansink and Hemker 2012), the test developers do not claim that these tests measure exactly the same (underlying) construct as the first/second grade mathematics and language tests. The different IRT scales used for the preschool and first/second grade tests also mean that absolute differences in scores are meaningless, which is why interpretation was restricted to the correlation measurements and percentile groups. As a result of the differences in construct, a perfect correlation could never be expected. However, given that we are dealing with interconnected constructs, the small size of the correlation coefficients does raise the question: What information does a low score on these preschool measurements actually convey? When considering identification/allocation as a goal for standardized tests, the results of this study indicate that there is a large group that might not be being provided with the help they need. In contrast, a smaller group may be receiving an intervention that could well be unnecessary. Adding to the complexity of the situation is the fact that the influence that language has on mathematics education might be playing a considerable role in terms of determining the predictive validity of the math test (Van Eerde 2009). This makes it difficult to view math as a completely separate construct from language.

Limitations and recommendations

Missing data appears to be a reoccurring obstacle in many studies on assessment in early childhood education. Selective testing by teachers or schools can be a major problem for internal validity when this is not dealt with in an adequate manner. Dummy coding with mean imputation (Duncan et al. 2007), excluding the dependent variable from the imputation procedure (Romano et al. 2010), and pairwise/listwise deletion (Dollaghan and Campbell 2009; La Paro and Pianta 2000) have all been shown to bias the coefficient estimates and/or standard errors (e.g. Allison 2009; Graham 2009). Through careful handling of missing observations, any threat to the internal validity of the current study is presumably limited. However, the loss of information does result in an increase in standard errors, thereby leading to a loss of power and increased uncertainty about any statistical parameters.

To assure unbiased results for the imputation model and multilevel model, all model assumptions were thoroughly checked. Notably, a relatively large proportion of children were imputed as having been tested with the new version of the test. However, since both versions were developed to measure the same latent construct, and since previous research has indicated that the correlations between the item banks of the new and old versions of the tests are very high (Koerhuis and Keuning 2011; Lansink and Hemker 2012), it is unlikely that the inclusion of two different versions had any large influence on the correlation estimates. Some differences can be seen in the stability measurements. Across the entire sample, the new version of the test does appear to be somewhat better at identifying children at risk for academic difficulties, although the differences are small. Additionally, while the nesting of measurements within children was adequately handled by the imputation model, it was not possible to include the nesting of children within schools in the imputation software. Since there was no significant variation between schools remaining in the language and mathematics models, once the fixed effects were added, exclusion of school level in the imputation model would most likely have had very little impact on the results.

Because the sample was selectively chosen from a specific region in the Netherlands, the results might not be representative for the entire Dutch population. For instance, the sample contains relatively few children with a foreign background (7.0%), who have been shown to perform relatively poorly on achievement tests (Centraal Bureau voor de Statistiek 2012; OECD 2008). This might limit the external validity of the results. In addition, a small group of children who repeated a grade were tested multiple times with the same test (~2% of the total number of measurements in the sample). Because only a few scores were observed twice, and because differences between the first and second tests were often small, the effects of this on the observed parameters are presumably negligible. Indeed, correlations with pairwise deletion showed minimal differences when using the first or second observations. This study illustrates how a longitudinal analysis of assessment data provides a more complete picture of how assessment scores develop over time. In addition, this method supports the analyses of academic achievement stability, an area that undoubtedly deserves more attention, especially in a target population, where this stability is under much scrutiny. Finally, the current study provides a better, more transparent evaluation of the suitability of assessment use for identification purposes.

The results of this study show that early childhood educators should be careful in their interpretation of test scores and take into account that there might be a wide margin of error when it comes to the early identification of children. In addition, when an educator has concerns about a child's academic development, these concerns might be better validated by means of tests specifically designed to identify children in the tails of the score distribution, rather than tests designed for a more general population. Though standardized tests that are normed on a national population might have a known predictive validity for the entire distribution of scores, this does not mean that these tests are adequate measurements for identifying children in the tails of this distribution. The selective testing of low-scoring students in the early years suggests that teachers/schools are generally more concerned about the test performance of these students. This could indicate use of these instruments as a diagnostic and evaluative tool but may also reflect an increased concern with low-scoring students 'making the cut.' Although diagnostic tools are just as important on the high-achieving end of the spectrum, the infrequent testing of high-scoring children suggests that the latter is more likely to be the case.

The authors recognize the importance of early assessment and the potential benefits of a norm referenced measure. However, a thorough evaluation of the underlying assumptions in assessment-based decision making is necessary to identify the limitations of an instrument. The results of this study suggest that the amount of variability in early development makes it difficult to base decisions about the child's educational trajectory on a single assessment outcome and may underpin the need for frequent assessments using multiple sources in the identification of children at risk for academic difficulties.

Notes

1. Since the terms in the Dutch educational system differ from the US terminology, the term 'preschool' in this paper is used to define both the pre-K and K classes, which correspond to the start of mandatory Dutch education known as groups 1 and 2 or *kleutergroepen* (age 4–6). The terms 'first grade' and 'second grade' are reserved for groups 3 and 4, respectively (age 6–8), at which point education becomes more formal in the Dutch system.
2. The original test uses five percentile groups. To facilitate interpretation, the two lowest percentile groups were combined to create four percentile quartiles.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Funding for this research was provided by the University of Groningen through The Netherlands Organization for Scientific Research (NWO) Graduate Program. The NWO had no influence in the study or the decision to publish.

References

Abu-Alhija, F. N. 2007. "Large-scale Testing: Benefits and Pitfalls." *Studies in Educational Evaluation* 33 (1): 50–68. doi:10.1016/j.stueduc.2007.01.005.

Allison, P. D. 2009. "Missing Data." In *The SAGE Handbook of Quantitative Methods in Psychology*, edited by R. E. Millsap and A. Maydeu-Olivares, 72–90. Los Angeles, CA: Sage Publications, Inc.

Centraal Bureau voor de Statistiek. 2012. *Jaarboek onderwijs in cijfers 2012* [Annual Education in Numerals 2012]. Den Haag. https://www.cbs.nl/-/media/imported/documents/2012/48/2012-f162-pub.pdf.

Colpin, M., S. Gysen, K. Jaspaert, R. Heymans, K. Van den Branden, and M. Verhelst. 2006. *Studie naar de wenselijkheid en haalbaarheid van de invoering van centrale taaltoetsen in Vlaanderen in functie van gelijke onderwijskansen* [Study on the Desirability and Attainability of Introducing Central Language Tests in Flanders for the Purpose of Equal Opportunities in Education]. Leuven: K.U. Leuven.

COTAN. 2011. *Toelichting bij de beoordeling Rekenen voor Kleuters Groep 1 en 2 LOVS Cito* [Supplement on the Assessment of Math for Preschoolers and Kindergarteners in the Studenten and Education Monitoring System by Cito]. Amsterdam: COTAN.

COTAN. 2013. *Toelichting bij de beoordeling Taal voor Kleuters (TvK)* [Supplement on the Assessment of Language for Preschoolers and Kindergarteners in the Studenten and Education Monitoring System by Cito]. Utrecht: COTAN.

Cronbach, L. J. 1971. "Test Validation." In *Educational Measurement*, edited by R. L. Thorndike, 2nd ed., 443–507. Washington, DC: American Council on Education.

Darrah, J., and M. Hodge. 2003. "Stability of Serial Assessments of Motor and Communication Abilities in Typically Developing Infants—Implications for Screening." *Early Human Development* 72 (2): 97–110. doi:10.1016/S0378-3782(03)00027-6.

De Wijs, A., F. Kamphuis, F. Kleintjes, and M. Tomesen. 2010. *Wetenschappelijke verantwoording Spelling voor groep 3 tot en met 6* [Scientific Account of the Spelling Tests for Grades 1 through 4]. Arnhem: Cito.

Dockrell, J. E., and C. R. Marshall. 2015. "Measurement Issues: Assessing Language Skills in Young Children." *Child and Adolescent Mental Health* 20 (2): 116–125. doi:10.1111/camh.12072.

Dollaghan, C. A., and T. F. Campbell. 2009. "How Well do Poor Language Scores at Ages 3 and 4 Predict Poor Language Scores at age 6?" *International Journal of Speech-Language Pathology* 11 (5): 358–365. doi:10.1080/17549500903030824.

Duncan, G. J., C. J. Dowsett, A. Claessens, K. Magnuson, A. C. Huston, P. Klebanov, Linda S. Pagani, et al. 2007. "School Readiness and Later Achievement." *Developmental Psychology* 43 (6): 1428–1446. doi:10.1037/0012-1649.43.6.1428.

DUO. 2013. *Toelichting Gewichtenregeling basisonderwijs* [Supplement Adressing the Educational Burden Regulation in Primary School]. https://www.duo.nl/Images/Toelichting, gewichtenregeling basisonderwijs, 29 april 2013_tcm7-39943.pdf.

Einarsdóttir, J. T., A. Björnsdóttir, and I. Símonardóttir. 2016. "The Predictive Value of Preschool Language Assessments on Academic Achievement: A 10-Year Longitudinal Study of Icelandic Children." *American Journal of Speech-Language Pathology* 25: 67–79.

Fuchs, L. S., D. C. Geary, D. Fuchs, D. L. Compton, and C. L. Hamlett. 2014. "Sources of Individual Differences in Emerging Competence with Numeration Understanding Versus Multidigit Calculation Skill." *Journal of Educational Psychology* 106 (2): 482–498. doi:10.1037/a0034444.

Gabadinho, A., G. Rischard, N. S. Müller, and M. Studer. 2011. "Analyzing and Visualizing State Sequences in R with TraMineR." *Journal of Statistical Software* 40 (1): 1–37.

Geary, D. C. 2006. "Development of Mathematical Understanding." In *Handbook of Child Psychology, Cognition, Perception, and Language*, edited by D. Kuhn, R. Siegler, W. Damon, and R. M. Lerner, 6th ed., 777–810. Hoboken, New Jersey: John Wiley & Sons, Inc.

Gilliam, W. S., and E. Frede. 2012. "Accountability and Program Evaluation in Early Education." In *Handbook of Early Childhood Education*, edited by R. C. Pianta, W. Steven Barnett, L. M. Justice, and S. M. Sheridan, 77–91. New York: The Guilford Press.

Goorhuis, S. M., and A. M. Schaerlaekens. 2000. *Handboek taalontwikkeling, taalpathologie en taaltherapie bij Nederlandssprekende kinderen* [Handbook Language Development, -Pathology and -Therapy for Dutch-Speaking Children]. 2nd ed. Leusden: De Tijdstroom.

Graham, J. W. 2009. "Missing Data Analysis: Making it Work in the Real World." *Annual Review of Psychology* 60: 549–576. doi:10.1146/annurev.psych.58.110405.085530.

Graham, J. W. 2012. *Missing Data: Analyses and Design*. New York: Springer. doi:10.1007/978-1-4614-4018-5.

Heckman, J. J. 2000. "Policies to Foster Human Capital." *Research in Economics* 54 (1): 3–56. doi:10.1006/reec.1999.0225.

Janssen, J., N. Verhelst, R. Engelen, and F. Scheltens. 2010. *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8* [Scientific Account of the Student and Education Monotoring System Math Tests for Grades 1 through 6]. Arnhem: Cito.

Kim, J., and H. K. Suen. 2003. "Predicting Children's Academic Achievement From Early Assessment Scores: A Validity Generalization Study." *Early Childhood Research Quarterly* 18 (4): 547–566. doi:10.1016/j.ecresq.2003.09.011.

Koerhuis, I., and J. Keuning. 2011. *Wetenschappelijke verantwoording van de toetsen Rekenen voor kleuters* [Scientific Account of the Preschool Math Tests]. Arnhem: Cito. http://toetswijzer.kennisnet.nl/html/tg/22.pdf.

La Paro, K. M., and R. C. Pianta. 2000. "Predicting Children's Competence in the Early School Years: A Meta-Analytic Review." *Review of Educational Research* 70 (4): 443–484. doi:10.3102/00346543070004443.

Lansink, N., and B. T. Hemker. 2012. *Wetenschappelijke verantwoording van de toetsen Taal voor Kleuters voor groep 1 en 2 uit het Cito Volgsysteem primair onderwijs* [Scientific Account of the Preschool Language Tests in the Cito Monitoring System Primary Education]. Arnhem: Cito. http://www.toetswijzer.nl/html/tg/18.pdf.

Leseman, P. 2004. "De toegevoegde waarde van vroeg testen [The added value of early testing]." *Pedagogiek* 24 (1): 3–11.

OECD. 2008. *Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value-Added of Schools (Vol. 2008)*. Paris: OECD Publishing. doi:10.1787/9789264050259-en.

Romano, E., L. Babchishin, L. S. Pagani, and D. Kohen. 2010. "School Readiness and Later Achievement: Replication and Extension Using a Nationwide Canadian Survey." *Developmental Psychology* 46 (5): 995–1007. doi:10.1037/a0018880.

Rovict, B.V. n.d. *Leerlingen, BRON* [Students, BRON]. http://www.rovict.nl/downloads/FAQ_NNCA.pdf.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Shepard, L., S. L. Kagan, E. Wurtz, W. Graves, P. E. Patton, R. Romer, James B. Hunt et al. 1998. *Principles and Recommendations for Early Childhood Assessments*. Darby: DIANE Publishing.

Shonkoff, J., and D. Phillips. 2000. *From Neurons to Neighborhoods: The Science of Early Childhood Development*. Washington, DC: National Academy of Sciences - National Research Council. http://eric.ed.gov/?id=ED446866.

Siegler, R. S. 2002. "Variability and Infant Development." *Infant Behavior and Development* 25 (4): 550–557. doi:10.1016/S0163-6383(02)00150-9.

Snijders, T., and R. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London: Sage Publications, Inc.

Snow, K. L. 2006. "Measuring School Readiness: Conceptual and Practical Considerations." *Early Education and Development* 17 (1): 7–41. doi:10.1207/s15566935eed1701.

Van Buuren, S., and K. Groothuis-Oudshoorn. 2011. "MICE : Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67.

Van Eerde, H. A. A. 2009. "Rekenen-wiskunde en taal: een didactisch duo [Math and Language: A Didicatical Duo]." *Reken-wiskundeonderwijs: Onderzoek, Ontwikkeling, Praktijk* 28 (3): 19–32.

Verhelst, N. D., C. A. W. Glas, and H. H. F. M. Verstralen. 1995. *One-Parameter Logistic Model OPLM*. Arnhem: Cito.

Verhelst, N. D., H. H. F. M. Verstralen, and T. J. H. M. T. H. J. M. Eggen. 1991. *Finding Starting Values for the Item Parameters and Suitable Discrimination Indices in the One-Parameter Logistic Model*. Arnhem: Cito.

White, I. R., P. Royston, and A. M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–399. doi:10.1002/sim.4067.

Zubrick, S. R., C. L. Taylor, and D. Christensen. 2015. "Patterns and Predictors of Language and Literacy Abilities 4–10 Years in the Longitudinal Study of Australian Children." *PLoS ONE* 10 (9): 1–29. doi:10.1371/journal.pone.0135612.