A captivating snapshot of standardized testing in early childhood
Frans, Niek

DOI:
10.33612/diss.95431744

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
https://doi.org/10.33612/diss.95431744

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Introduction
Chapter 1

Context of this dissertation

Between 1956 and 1985, education of four to six-year-olds in the Netherlands took place in separate nursery schools\(^1\). Following the Primary Education Act in 1985, educational provision for this age group was integrated into primary education to improve the transition between nursery classes and primary school (Den Elt, Van Kuyk, & Meijnen, 1996). This integration led to a struggle between two opposing views of education (Den Elt et al., 1996). On the one hand, there was the development-orientated view that originally dominated the nursery schools and places emphasis on learning through play, child initiated learning and assessment through observation. On the other hand, there was the curriculum- or program-orientated view in primary education that underscored the importance of developmental goals, planned instruction and regular testing (Den Elt et al., 1996).

The National Institute for Educational Measurement (Cito) played a major role in the advancement of an integrated teaching method through the development and introduction of a pupil monitoring system and remedial programs (Den Elt et al., 1996). Their vision on education was based on the guiding rule: ‘development-orientated where possible (if children have the capacity to direct their own learning) and curriculum-orientated where necessary (if the child is not able to direct his/her learning).’ (Den Elt et al., 1996, p. 23). While the model includes many developmental aspects, the use of tests and traits from program-orientated education led to resistance of this approach among preschool and kindergarten teachers. Although assessment in preschool and kindergarten (age 4 to 6) still consisted primarily of teacher observation (Dutch Eurydice Unit, 2007), schools were required to administer at least one nationally norm-referenced assessment for both language and mathematics before first grade. This meant that tests from Cito’s pupil monitoring system (Leerling- en OnderwijsVolgSysteem, LOVS) were widely implemented in early education.

To date, the discussion surrounding test use in early childhood education is outspokenly ongoing. This dissertation was written in a turbulent period that saw many policy changes surrounding the use of early standardized normative tests in the Netherlands. In November 2013, when this dissertation was launched, the House of Representatives in the Netherlands voted on a motion to abolish the mandatory administration of a nationally norm-referenced standardized test in

\(^1\) Since the terminology in the Dutch educational system differs from international terminology, the term ‘nursery school’ is used to define education of 4 to 6-year-olds before its integration in primary education in 1985. We will use the term ‘preschool’ to define education after 1985 between the ages of 4 and 5 [groep 1] and ‘kindergarten’ when referring to education of 5 and 6-year-olds [groep 2]. Finally, the term ‘first grade’ is reserved for the Dutch group 3 (age 6–7), at which point education becomes more formal in the Dutch system.
preschool and kindergarten (Rog, Bisschop, van Dijk, Voor dewind, & Klaver, 2013). The motion was accepted by a majority of the House on the basis that the unstable development in preschool and kindergarten makes reliable testing challenging and favors observation as a means of monitoring development.

In the coalition agreement of October 2017, the newly formed government decided to prohibit the use of tests from the LOVS in preschool and kindergarten altogether. The reasons for this decision were later outlined in a letter to parliament (Van Engels hoven, 2018). By 2021 only observational instruments that are endorsed by the ‘Expert group Testing Primary Education’ are allowed to be used to monitor the development of young children. According to the Minister of education, culture and science, the scholastic format of the LOVS tests does not fit the way in which preschoolers and kindergarteners develop and frustrates both teachers and care coordinators. This can be illustrated by the black papers on early childhood education (ECE) presented by the lobby group WSK (Werkgroep en Steungroep Kleuteronderwijs, 2013), which contain 101 stories from educators who express frustrations about testing and increasingly formalized education. In addition, she states that the normative scores do not do justice to the discontinuous development of young children. Since the minister indirectly refers to work from this dissertation as support for this decision, we will reflect on this decision and its motivation more extensively in the discussion of the dissertation.

Although the position of standardized testing in ECE seems to be slowly diminishing in the Netherlands, it is still a relevant topic both in the Netherlands and in many other countries. For example, a study by Bassok, Latham and Rorem (2016) showed that standardized tests are increasingly prominent in ECE in the United States. Other scholars (Meisels, Steele, & Quinn, 1989; Shepard, 1994) likewise observe this increase in formal testing of young children. More recently, Roberts-Holmes and Bradbury (2016) talk about the ‘datafication’ of ECE in England, where teachers feel increasingly pressured to produce objective numerical data for national comparison. While Kilderry (2015) writes on ‘the intensification of performativity in early childhood education’ in Australia, where external performance measures increasingly influence on teaching and curriculum. Although this dissertation provides a portrait of a specific context at a particular point in time, the findings that are presented here are relevant to the broader discussion surrounding early childhood assessment. Since international literature is rife with comparable discussions, the insights presented in this dissertation can be applied to similar early childhood assessment contexts.

**The potential value and challenge of early childhood assessment**

There are several important reasons for wanting to assess academic skills in preschool and kindergarten. For example, a meta-analyses by Duncan et al. (2007) showed that mathematics- and
reading achievements between the age of 5 to 6 are the strongest predictors of later mathematics- and reading achievements. This finding was replicated in a study by Romano, Babchishin, Pagani and Kohen (2010). The connection between early learning deficiencies and later academic abilities has led to an increasing interest in the early identification of these children (e.g. Gersten, Jordan, & Flojo, 2005; Scarborough, 2009). The reasoning is that detecting and resolving potential problems at an early age may prevent or ameliorate later learning difficulties. In addition, some authors (e.g. Guralnick, 2005) argue that early academic difficulties may have a cascading effect on later years, resulting in exaggerated problems in later grades. The generally superior effectiveness of early intervention compared to later remediation may be seen as support for this claim (e.g. Barnett, 2011; Heckman, 2000). Although most research on early intervention has been conducted in the US, where differences both in early childhood quality standards and targeted population make comparisons with European countries difficult, results on early childhood intervention in Europe also show generally promising effects (Burger, 2010). Another less pedagogical reason for the increasing role of early childhood assessment can be found in the growing trend in educational accountability that is slowly trickling down through primary education (Bordignon & Lam, 2004; DeLuca & Hughes, 2014).

In a quest for higher grades, teachers may be prompted to start scholastic teaching and testing at an earlier age.

Although there are obvious advantages to early childhood assessment, young children are notoriously difficult to assess accurately (Shepard et al., 1998). This is ascribed to 1) their rapid and discontinuous development, 2) their ‘testability’ and 3) a high influence of their environmental context. Early childhood development occurs at a rate that outpaces growth rates at all later stages in life (Shepard et al., 1998; Shonkoff & Phillips, 2000). Growth at this stage occurs irregularly and simultaneously in various domains such as physical, motor and linguistic development (Nagle, 2000; Shepard et al., 1998). Because of this rapid development with large intraindividual variability, tests given at one point in time may not give a good representation of a child’s later ability (Nagle, 2000).

In addition to their rapid development, young children generally show behavior that is incompatible with the traditional paper-and-pencil format that is used in many standardized instruments. Young children typically have short attention spans and show high levels of activity and distractibility (Nagle, 2000). In addition, as they usually have little or no experience with formal testing, they often do not see the importance of performing well or persisting on test items (Nagle, 2000; Shepard et al., 1998). Finally, the testing situation may be strange and unfamiliar to children as early childhood curriculum is generally aimed at learning through showing and doing in group settings. This unfamiliar context may prevent children from fully demonstrating their abilities (Shepard et al., 1998).
Substantial differences in the impact of the environmental background of children may create further short-lived differences in performance on early tests (Shepard et al., 1998). For example, home backgrounds in which children are hardly exposed to the native language may inhibit children from showing their full potential on any assessment measure that has a built-in language component. More generally, children’s prior experiences with language, working independently and testing may differ widely (Nagle, 2000). Because early test results are a complex mix of these prior experiences and a child’s ability to learn, measures of past learning are not necessarily indicative of a child’s potential to learn (Shepard et al., 1998). The large interindividual differences that stem from these diversities in context and spurts in development means that tests need to be able to measure a wide range of performance (Colpin et al., 2006). This broad variation in performance also makes it difficult to determine at what point performance may be considered problematic (Goorhuis & Schaerlaekens, 2000; Pellegrino, 2012).

The aforementioned difficulties in early childhood assessment led various authors to conclude that test results at this age do not provide an estimate that is stable enough to accurately identify children as at-risk for academic problems (Colpin et al., 2006; Dockrell & Marshall, 2015; Shepard et al., 1998). Indeed La Paro and Pianta (2000) concluded from a structured review on the relationship between preschool academic and social assessments and later performance that ‘child-based assessment of skills will not accurately identify “high risk” children’ (p. 476). While emerging academic skills may be related to later academic success, it seems that many instruments are not able to make a reliable distinction between children who will develop academic problems and those who will not. Specifically, many early assessment instruments lead to a sizeable portion of both false positives and false negatives (Law, Boyle, Harris, Harkness, & Nye, 2000; Nelson, Nygren, Walker, & Panoscha, 2006; Scarborough, 2009). Scarborough estimates that about half of the children identified as ‘in need of reading intervention’ may not actually need it. While failing to identify a child in need of intervention may have seriously harmful consequences, ‘false alarms’ may likewise lead to negative educational and psychological consequences such as stigmatization and feelings of incompetence (Abu-Alhija, 2007; Roberts-Holmes & Bradbury, 2016; Scarborough, 2009; Shepard, 1994).

**Assessment and standardized testing**

Throughout this dissertation we will use the term ‘test’ to refer to formal standardized paper-and-pencil instruments (or their digital equivalent) and ‘assessment’ to refer to the broader process that may include tests (as defined above), observations and other means to gain relevant information about children that may be used in a decision-making process (Steenbeek & Van Geert, 2018).
Chapter 1

Although ‘test’ may also be used to describe other forms of assessment, such as systematic observation, the term has acquired a connotation in common parlance with formal, single-answer instruments used to rank individuals (Shepard, 1994). Likewise, while ‘assessment’ can include standardized paper-and-pencil tests it is frequently used to place a substantive focus on the pedagogical role of testing. This division between testing as an accountability instrument to judge the quality of teaching and learning, versus testing as a pedagogical instrument used to improve the educational process can be found throughout assessment literature (Barnes, Fives, & Dacey, 2015; Black, 2002; Brown, 2008; Remesal, 2007). It has been indicated by the terms ‘summative assessment’ and ‘formative assessment’ (e.g. Black, 2002; Newton, 2007) or ‘assessment for improvement’ and ‘assessment for accountability’ (e.g. Brown, 2004).

Often, these two purposes of assessment are seen as antithetical to one another (e.g. Torrance, 1997), with summative assessment representing the negative social aspects of assessment and formative assessment the constructive positive aspects (Taras, 2005). However, several authors have noted that the terms ‘summative’ and ‘formative’ focus on different complementary aspects of the same process (e.g. Newton, 2007; Taras, 2005). While the term summative can only ever refer to an assessment judgment, formative can only be used to refer to a use to which this judgment is put (Newton, 2007). As any use of assessment results inevitably includes some form of judgment, formative assessment can be seen as an extension of summative assessment (Taras, 2005). While the process of summative assessment stops when a judgment is reached, formative assessment continues by using this judgment to shape and improve the educational process (Taras, 2005).

Taras (2005) further stipulates that a judgment cannot be made within a vacuum; some point of comparison such as a norm or goal is necessary. Many forms of standardized assessment very explicitly provide a judgment in relation to a norm or goal, which may explain why they are sometimes referred to as instruments for summative assessment (e.g. Abu-Alhija, 2007; Black, 2003; Lam, 2013; Serafini, 2001). However, this explicit judgment does not preclude their formative use in education (Newton, 2007). Standardization means that the methods used to administer and score an assessment instrument are fixed in a predefined manner that is identical for each member of the target group. One of the advantages is that this procedure can provide a judgment that compares children to some pre-defined norm or criterion. A norm-referenced test places a child’s achievement on a continuum relative to a representative group of children who are generally given the test items before they become publicly available (Bond, 1996). Such tests provide a numerical ranking of the child relative to the standardization sample (Meisels, Wen, & Beachy-Quick, 2010). On the other hand, a criterion-referenced test relates a child’s performance to a set of educational goals that is pre-determined by the test designers (Bond, 1996). As such, these tests provide an indication of a
child’s mastery of a set of independent standards, irrespective of the performance of other children (Meisels et al., 2010).

Although a norm- or criterion-referenced judgment can potentially be put to some formative use, Brown and Hattie (2012) point out how instruments must go beyond reporting a single rank score in order to lead to improvements in teaching and learning. Test results should provide teachers with further diagnostic information that can support the identification of specific strengths and gaps in the child’s abilities and taught curriculum. Moreover, the content of the test must be well aligned with key aspects of the curriculum (Brown & Hattie, 2012). According to Taras (2005), the possibility of formative use is a necessary step that justifies and enriches summative assessment. This view is in line with one of the general principles of the Early Childhood Assessment Resource Group (Shepard et al., 1998) that assessment should bring about ‘a clear benefits for children - either in direct services to the child or in improved quality of educational programs’ (p. 5). While the unification of a summative judgment with a formative use is possible, the formative use of standardized norm-referenced instruments is still a controversial topic (e.g. Nagy, 2000; Torrance, 1997).

The Cito preschool and kindergarten tests

Since the National Institute for Educational Measurement has historically played a major role in the development of standardized norm-referenced tests, the majority of schools in the Netherlands use the Cito preschool and kindergarten tests to monitor children’s development in the first two years of education (age 4 to 6). The ‘language for preschoolers and kindergartners’ (Taal voor Kleuters, TvK, Lansink & Hemker, 2012) and the ‘mathematics for preschoolers and kindergartners’ (Rekenen voor Kleuters, RvK, Koerhuis & Keuning, 2011) tests consist of standardized norm-referenced instruments. Both tests contain one instrument that is administered in the middle and end of preschool and one instrument that is administered in the middle and end of kindergarten. These tests are part of a larger pupil monitoring system (Leerling- en OnderwijsVolgSysteem, LOVS), which contains a collection of instruments designed to monitor a child’s development from preschool to the end primary education and is used by over 80% of primary schools in the Netherlands (Gelderblom, Schildkamp, Pieters, & Ehren, 2016; Veldhuis & Van den Heuvel-Panhuizen, 2014).

Content and construct

Both instruments (i.e. TvK and RvK) consist solely of multiple-choice items. The instruction for each item is read aloud by the classroom teacher (or in the case of digital tests by the computer) and the child is asked to answer the question by underlining the correct answer in a booklet. An example of such a question for the RvK tests is shown in Figure 1.1. The TvK language tests are designed to measure children’s receptive language ability (Lansink & Hemker, 2012). In preschool, children are tested on 48 multiple-choice items that measure receptive vocabulary (linking a word or description
Chapter 1

to a person, object, action or situation) and comprehension of spoken language. These items aim to measure a child’s conceptual awareness, defined as recognizing concepts and understanding the meaning of short, spoken texts. The kindergarten instrument contains 60 multiple-choice items that measure a child’s awareness of language in addition to their conceptual awareness. Awareness of language is defined as a child’s understanding of the shape and sound of written and spoken language independent of its meaning.

These items focus on sound and rhyme (recognizing alliteration and tail rhyme), recognition of first and last words, phonemic synthesis (forming a word by combining several phonemes) and knowledge of written texts.

The RvK mathematics tests measure children’s emerging numeracy ability (Koerhuis & Keuning, 2011). The instruments in preschool and kindergarten contain 46 and 48 items respectively that measure performance on three categories: number sense, measurement, and geometry. Number sense items measure a child’s understanding of the number line, number symbols, concepts of quantity and simple arithmetic operations. Measurement items relate to understanding of and working with concepts of length, weight, volume and time. These include notions of long, wide, empty, heavy, earlier, etcetera. Finally, geometry items measure a child’s ability in spatial orientation, construction of basic shapes, and performing operations with shapes and figures.

Scores and test calibration

The items in the LOVS tests are calibrated on large representative samples using Item Response Theory (IRT) models (Verhelst, Verstralen, & Eggen, 1991). One of the advantages of these
models is that items and persons can be compared on a single scale that estimates the difficulty of the items and the child’s position on the latent trait. By using so called anchor items (items that are calibrated at different grade levels) in the calibration of the item bank, it is possible to compare scores on different tests over time. The estimate of the child’s latent trait position is known as the ‘ability score’ [vaardigheidsscore] and can be used to compare the child’s estimated ability level to the population and to his/her previously obtained scores (as these are measured on the same scale). The language and mathematics tests have their own specific latent trait scale. Starting in first grade, language ability is measured with separate specific instruments that each measure a distinct language skill. While ability scores are comparable within a (sub)domain, comparisons across domains cannot be made with these scores.

Likewise, the preschool and kindergarten tests are calibrated on their own latent trait scales, independent from the scales of language and mathematics tests that are used from first grade onward. Although these independent scales give the impression that these tests deal with distinct and unrelated constructs, one of the primary arguments for the content of the preschool(kindergarten) tests is that these skills are strong predictors of later language ability (Lansink & Hemker, 2012) or play an important role in the subsequent development of mathematical skills (Koerhuis & Keuning, 2011). As such, although direct comparisons between the different scales cannot be made, a positive relation between these test scores should be expected.

<table>
<thead>
<tr>
<th></th>
<th>I 20% highest scoring pupils</th>
<th>II 60% - 80%</th>
<th>III 40%-60%</th>
<th>IV 20% - 40%</th>
<th>V 20% lowest scoring pupils</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>25% highest scoring pupils</td>
<td>B 50% - 75%</td>
<td>C 25%-50%</td>
<td>D 10% - 25%</td>
<td>E 10% lowest scoring</td>
</tr>
</tbody>
</table>

*Figure 1.2: Ability levels of Cito LOVS tests according to the new (top) and old (bottom) classification.*

By comparing the ability score to the norm distribution, teachers are also provided with a percentile score that expresses the percentage of children in the norm population who obtained a lower score than was achieved by a particular child. This is usually reported in an ability level expressed as a letter between A and E (old classification) and/or a Roman numerical between I and V (new classification, since 2013), as shown in Figure 1.2. Using these scores, a teacher can compare the performance of classrooms and individual children to national norms and performance on previous tests. Finally, to give teachers an indication of specific item types that a child or classroom struggles with; scores on the different item categories described in the previous paragraph can be compared to the expected score given the child’s overall performance.
Chapter 1

The quality of all instruments from the LOVS was extensively studied when the test norms were established. Results of these studies are publically available and describe the process and idea behind the development of these tests as well as psychometric qualities such as the reliability of test scores and correlation with scores from other instruments (Koerhuis & Keuning, 2011; Lansink & Hemker, 2012). In addition, the instruments were evaluated by the Dutch Committee of Test Affairs (Commissie Testaangelegenheden Nederland, COTAN), an independent committee that judges the quality of the test material and the reliability and validity of test scores. The preschool and kindergarten tests were found to have reliable scores and good overall quality (COTAN, 2011, 2013). Construct validity of the instruments was judged ‘satisfactory’ since correlations with older versions of the instruments were high, but correlations with other instruments were not reported. Finally, the criterion validity of these instruments was not explored as the developers report that these tests are not intended for predictive use.

While these instruments are primarily designed to describe and monitor a child’s academic proficiency (Koerhuis & Keuning, 2011; Lansink & Hemker, 2012), this is not an argument for the absence of predictive use. On the contrary, like in many tests the monitoring of academic proficiency is conducted to guide educational decisions which have an inherent predictive nature (Shepard, 1997). In addition to determining the language and/or mathematics proficiency of individual children or groups of children (Koerhuis & Keuning, 2011; Lansink & Hemker, 2012), the test manual describes these tests as identification instruments that determine whether or not the language and/or mathematics development of children requires additional attention in general or in specific areas (Koerhuis, 2010; Lansink, 2009). Consequently, intervention effects and future outcomes become an inherent part of the tests validity framework (Shepard, 1997). This idea has been brought to attention by Messick (1989) as ‘consequential validity’ but has been a conventional part of test validity for decades (Shepard, 1997). In a chapter on validity, Cronbach (1971) explained a decision as a choice between several courses of action and stated that ‘The justification for any such decision is based on the prediction that the outcome will be more satisfactory under one course of action than another’ (p. 448, emphasis added). According to Cronbach, the role of tests in this process is to reduce the number of incorrect predictions. When a test score is used as a basis for remediation efforts, the prediction is made that this test score signifies a problem in a child’s future development and intervention is necessary to change this predicted outcome (Bracken & Walker, 1997). As such, claiming that tests do not have a predictive purpose is only true if no intended use other than describing performance is specified (Shepard, 1997).

While adequate prediction of problems in the child’s academic development is a requirement when an instrument is used for early identification and remediation, it is not sufficient. Another key issue is whether the instrument provides diagnostic information that can used to make curricular
adjustments (Brown & Hattie, 2012; Messick, 1989). Although these instruments provide suggestions for remediation in the form of scores on specific item categories, it is unclear how teachers use this information. As phrased by Tiekstra, Bergwerff and Minnaert (2017), the issue whether these instruments are able to bridge the gap between diagnosis and intervention is unresolved as of yet.

Aim and outline of the dissertation

Although detecting difficulties in children’s language and/or mathematics development at an early age can lead to the prevention or early remediation of later academic problems, children are notoriously difficult to assess accurately at an early age. The instability of early development – and by extension of early test scores – is one of the main arguments used by several scholars and the Dutch Minister of education, culture and science to plea against the use of tests like those developed by Cito in ECE. While Cito claims that their standardized preschool and kindergarten tests can be used to identify which children need additional attention in their language- and/or mathematics education, there is no report on the relation between these scores and later achievement. Furthermore, it is unclear if teachers consider the information that these tests provide useful in guiding their remediation efforts. To evaluate the utility of these instruments in identification and remediation efforts this dissertation answers the following research questions: How do teachers experience the utility of the Cito preschool and kindergarten tests in their daily educational activities? What is the stability of early test scores from the Cito LOVS? How does the stability of these test scores affect test-supported decisions about individual children? The dissertation is built up in two sections: Chapters 2 and 3 provide a first exploration of the way teachers use and view these tests and the stability of these test scores. The findings in these chapters are used in Chapters 4 and 5 to further explore the stability of these scores in relation to test-supported decision-making.

Chapter 2 examines teachers’ conceptions of the Cito preschool and kindergarten test as an instrument that can be used to improve teaching and learning. A questionnaire was distributed among 97 early childhood educators to explore their conceptions of the purposes of these tests. In addition, in-depth interviews with a small selection of teachers with varying conceptions were used to investigate factors that may influence teachers’ conceptions of these tests.

Chapter 3 presents the findings of a pilot study on the predictive validity and stability of the Cito preschool tests. It looks at the intra-individual variability of test scores from 431 children as expressed by transitions between achievement levels. Furthermore, this chapter is a first examination of the predictive validity of the TvK and RvK tests for later language and mathematics abilities.

Chapter 4 delves deeper into the meaning of the term ‘stability’ and how this concept plays a role in early childhood assessment. Building on the definitions formulated by Wohlwill (1973) and
Chapter 1

work by Tisak and Meredith (1990), the chapter presents a framework that can be used to evaluate stability of test scores. We use this framework to evaluate the stability of the mathematics and language scores of 1402 children between kindergarten and third grade.

Chapter 5 presents a practical evaluation of the test manual recommendations from the perspective of the teacher. Score expectations that take into account individual growth rates are compared to expectations that only consider the child’s achievement level. Predictions of ability scores under both expectations are compared for the language and mathematics scores of 911 children between kindergarten and third grade.

In the final chapter, we present an overview of the major findings and conclusions of these studies. In addition, we address the limitations of this dissertation as well as recommendations for further research and practice. In particular, we reflect on the policy-driven decision made by the Minister of education, culture and science to abolish use of these tests by 2021.