General discussion
Research findings

One of the main arguments against formal testing of young children is that test scores in preschool and kindergarten are too unstable to allow inferences about future development (Nagle, 2000). These unstable scores are problematic when identification is seen as one of the main purposes of a test. This is because the argument for identification is not based on current performance, but on the expectation that this performance reflects some unfavorable outcome unless action is taken (Bracken & Walker, 1997; Cronbach, 1971). While the stability of scores plays an important role in early childhood assessment, even highly stable scores that lead to accurate predictions about future performance may not provide teachers with the tools to act appropriately on these predictions to remediate (potential) academic difficulties. In addition, while standardized norm-referenced instruments may provide information that can be used to improve the process of teaching and learning, the explicit judgment that these instruments provide may lead educators to view them primarily as accountability instruments.

This dissertation is written with the ambition to answer three 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?’, and ‘How does the stability of these test scores affect test-based decisions about individual children?’ In the first section of this chapter, we answer the question on the utility of these tests. Next, we answer the two questions related to the stability of these tests scores.

How do teachers’ experience the utility of these tests?

Chapter 2 focuses on how teachers view these instruments as tools for the improvement of teaching and learning. This study used quantitative analyses of a questionnaire to select interesting cases for semi-structured interviews. The questionnaire results showed that educators generally do not view these tests solely as accountability instruments. In fact, they did not seem to make a clear distinction between the instruments’ accountability and improvement purposes. This could indicate that teachers share the view of Taras (2005) that judgment and use are complementary parts of the same assessment process. However, teachers did seem to hold separate ideas about the test’s usefulness and its usefulness to them.

Further interviews with a selection of teachers revealed that although teachers are aware of both the accountability and improvement purpose, they differ substantially in how they experience these purposes. While some teachers view the norm score as a pleasant and welcome confirmation of their own observations, other teachers view these scores as antagonistic to their own observations. Throughout the interviews, teachers invariably spoke in terms of failure if children scored below average. Sometimes this idea was reinforced by the color scheme of the test (‘getting
children out of the red zone’) or by other parties such as the schools’ management team (MT) or parents. Teachers who experienced these tests more positively used the same terms of failure and success but tended to teach classrooms where most children scored above average and/or felt supported by their MT in the interpretation of the test results and subsequent planning of remediation. Although some teachers found the test format unsuitable for young children, others saw the test format primarily as a useful preparation for future testing and formal learning.

While the results of this study show that not all teachers view these types of standardized norm-referenced assessments as negative accountability instruments, it does show that scores interpretations quickly reduce to pass/fail judgments. In line with this finding Faber, Van Geel, and Visscher (2013) found that teachers tend to focus on the normative achievement level that the LOVS tests provide. This is unsurprising given that the manual advises teachers to select children who score in the lowest 20% to 25% for further assessment and intervention. As a result, teachers tend to view these scores as insufficient and adjust their teaching activities to prevent or improve these scores. Such activities include explicitly mentioning the words that are included in the test or offering material in a format that is similar to the one used in the tests. This is often done with good intentions, because teachers think the test measures what children need to know, or because they think it is unfair to test children on unfamiliar formats or content.

As a result of this view and corresponding actions, test scores are likely to become higher than the scores in the original norm group were. Indeed, this inflation of norms could be observed in our quantitative data for most tests and was more pronounced for older tests compared to versions with more recently updated norms. This inflation of norms is a problem that has been noted before by Cito (Keuning et al., 2014). It is considered a problem because the norms no longer describe the distribution that was initially intended and will likely overestimate a child’s performance relative to the population. To prevent this, the norms of some of the older tests have been recalibrated using data from current test administrations (Keuning et al., 2015). As Shepard (1990) notes, periodically updating test norms may provide a solution if the norms are simply outdated. However, if the problem is caused by a curriculum that is focused more narrowly on the tests content, updating the norms might exacerbate the problem by creating a standard that is increasingly unattainable without adjusting the curriculum. Because the preschool and kindergarten tests are already considered to be too difficult by many teachers, the norms for these tests have not been updated since the introduction of the new version (Papenburg, 2015). The test user and test developer seem to have conflicting goals in this respect. Educators prefer to see high scores as a reflection of the quality of their teaching and their students’ performance, while Cito’s objective is that the scores are an accurate reflection of performance relative to the norm group. The focus on these normative scores at both ends may draw attention away from potentially more constructive uses of these instruments.
Chapter 6

**What is the stability of early test scores from the Cito LOVS?**

Chapters 3, 4 and 5 explore the stability of the test scores from the Cito LOVS. Since children who score in the lowest 20% or 25% are generally labeled as at-risk (Koerhuis, 2010; Lansink, 2009), we defined stability in Chapter 3 as the consistency of percentile ranking between measurement occasions. By exploring the achievement levels of 431 children on language and mathematics tests between preschool and second grade we found that only a small portion of children who scored in the at-risk category – only 11% and 17% for language and mathematics respectively – did so consistently. A large number of children who later scored in an at-risk category – 47% and 35% for language and mathematics respectively – achieved far higher scores in preschool and kindergarten. Scores in the highest category were found to be more stable, with an estimated 61% of children who score consistently in this category on consecutive measurement occasions, compared to around 30% for lower categories. Finally, this chapter looked at the between-test correlations and found that the preschool and kindergarten tests correlated far lower with each other on average ($r = .3$) compared to post-kindergarten tests ($r = .6$). In addition, average correlations between preschool/kindergarten and later test administration were generally lower still ($r = .2$). Chapter 3 concludes that these low correlations might be indicative of large intra-individual variation over time in the preschool and kindergarten years.

Although our initial operationalization of stability encapsulates one possible type of stability, a more extensive exploration of the literature identified many different uses of this term that are not captured by this definition. In addition, we noticed how correlation coefficients have several limitations that make them unsuitable for evaluating these different characterizations of stability. In Chapter 4 we extended a framework by Tisak and Meredith (1990) that can be used to describe and evaluate several types of stability that were first pitched by Wohlwill (1973). Broadly speaking, stability is defined here as the predictability of later test scores from previously achieved scores. By specifying how scores are connected over time, Wohlwill differentiates specific types of stability. Tisak and Meredith later show how nested structural equation models can express three of these types. In Chapter 4, we discuss how different assumptions about the stability of test scores can lead to different types of interpretations. In addition, we add a fourth definition to the framework of Tisak and Meredith and show how multilevel models may be used to describe two distinct types of stability of the Cito test scores. First, the **linear stability** of the scores, or the assumption that children’s scores develop at a rate similar to each other and retain their ranking over time. Second, the **function stability** of the scores, which is the more lenient assumption that children’s test scores develop according to a single function, but with distinct individual growth rates. Both assumptions were evaluated using the test scores of 1402 children between kindergarten and third grade. The results showed that function stability provides a significantly better description of the test scores in the
sample. However, they also showed that the gain in model fit is small and test scores of a large group of children were adequately described by the simpler assumption of linear stability. Although there was a small group – 10.7% and 12.1% for language and mathematics respectively – that substantially differed in growth rate, identifying these children based on their test scores was only possible with relative certainty after five or more test administrations (comprising 2.5 years). The results of this study suggest that these tests might not be sensitive enough to identify differences in individual growth rates.

In Chapter 5, we extended these findings to the practical interpretations that teachers may attribute to these test scores. Instead of describing the scores over the entire period, we looked at predictions that teachers may make about the next test score, based on the preceding achievement of the child. We used the two types of stability that were tested in Chapter 4, as these correspond to two recommendations in the manual of these tests. We also varied the amount of information included in the prediction, namely all preceding test scores or only the minimal number of test scores required to make a prediction. These predictions correspond to different perceptions of score stability that some of the teachers in Chapter 2 seemed to hold, based on their interview statements and a small task in which they were asked to order sets of scores (Appendix C).

A subsample of the 1402 children in Chapter 4, consisting of 911 children who were tested with the latest version of the kindergarten instrument, was used to evaluate the accuracy of different predictions. The results showed that predictions of subsequent performance do not become more accurate when the growth rate of individual children is taken into account. On the contrary, predictions that consider individual growth are often less accurate than those made under the assumption of a single average growth rate. This is especially true when predictions are based on the growth between the last two test administrations. The last obtained score often provides a more accurate indication of subsequent performance, even for children who show a substantially different growth rate within the studied period. The recommendation to identify children as ‘at-risk’ when they show stagnations in growth will likely lead to a substantial amount of false positive identifications. Over 60% of children in the sample showed at least one score stagnation between first and third grade. Children who show a significantly different growth rate only make up a small proportion of these stagnations,.12 and .11 for language and mathematics respectively. Although predictions are considerably more accurate when assuming an equal and average growth rate for all children, there is still substantial error in these predictions. Even the best predictions deviate by more than 16 respectively 13 percentile points from the observed score 50% of the time for language and mathematics. As expected, the size of these deviations decreased as children become older and more information is available. Excluding the kindergarten scores from these predictions hardly influenced their accuracy. Likewise, although children who score below the 20th percentile in
kindergarten are more likely to score in this category on subsequent occasions, a sizeable proportion of these children – 18.3% and 37.9% for language and mathematics – scored higher on all subsequent test administrations.

The results in Chapter 4 and Chapter 5 support the claim that early test scores are relatively unstable compared to those in later years. Although the test scores do provide a global estimate of ability relative to other children in the same educational stage, the scores of individual children show large and seemingly unstructured intra-individual variation around this estimate. This may make interpretations of a single score unreliable. However, it is important to distinguish between the types of interpretations that teachers may want to make with these tests. The results of Chapters 4 and 5 show that interpretations based on individual growth in scores likely lead to conclusions that are less accurate compared to interpretations based on a child’s estimated ability. Under most circumstances, predictions are more accurate when they are based on the assumption that children’s scores develop at the same rate. Similar to other studies on emerging academic skills (Duncan et al., 2007; La Paro & Pianta, 2000) each study found a weak to moderate connection between performance on the preschool and/or kindergarten test and later achievement. In addition, this link was relatively weaker for language domain tests compared to mathematics tests (Duncan et al., 2007). When used in a decision-making process, the lack of stability signifies that extreme caution is required when interpreting these results as an estimate of ability. Like other authors (Law et al., 2000; Nelson et al., 2006; Scarborough, 2009) we found substantial proportions of false negative and false positive identifications. Although, it is important to note that scores tend to be slightly more stable in the extreme score ranges of these tests, even the best predictions have wide error margins.

**Critical reflection**

This dissertation is an example of how readily available data can be collected to efficiently generate large representative samples. A major strength of the quantitative studies is the scale at which data were collected. In cooperation with schools we were able to relatively quickly collect assessment data on a large number of children and, more importantly, over a longer period of time. Longitudinal data of this scale is necessary to draw reliable conclusions about the stability of such test scores. On the other hand, retrospectively collecting data at this scale also provided significant challenges. Although we did our best to include data on intervention effects such as grade repetition and Individualized Educational Plans (IEP), very often this information was imprecise or unavailable. Collecting the data retrospectively meant that some information could not be retrieved. For example, not all schools kept detailed records on the IEP of individual children. This either meant that only general information on the presence of an IEP or no information at all was available. We did look at the effects of grade repetition and IEP, but often excluded these effects as the addition of this
information did not change the conclusions and it was unclear what having an IEP meant for individual children at any given time (i.e. individual or group instruction, the type and goal of intervention). A more detailed evaluation that included specific intervention effects on these tests might have been possible if specific information was available on individual intervention. Although this was a downside, the available data did not indicate a substantial or structural influence of individual intervention. Furthermore, the use of retrospective data meant that children who were referred to special education in this period might not have been included in our sample. This is a limitation in the generalizability of our results to this specific population. On the other hand, it is likely that these children are also omitted from the norm-population used to calibrate these tests. As such, our sample should still adequately reflect the test population. Finally, the use of retrospective data ensured that the study did not influence daily classroom practice, which is beneficial for the participating schools and children. It also ensures that testing activities were not influenced by participant bias.

A second methodological challenge was the occurrence of missing data. This was both an interesting result and a potential source of bias. Especially in Chapter 3, the occurrence of missing data was dependent on the achievement level of children. The schools in this sample chose to test children again at the end of the year only if they scored in the at-risk achievement levels in the middle of the year. Later in the interview study, we found that this is the ongoing policy in some schools. Either schools test these children again to evaluate the effects of intermediate educational activities, or because these children need to ‘pass the test’ since they failed to do so the first time. Multiple imputation was needed to alleviate bias that resulted from testing policies such as these. Fortunately, this bias was less apparent in the larger dataset and the longitudinal nature of the data meant that we could use available test scores to make a reliable estimate of the missing score.

The dual quantitative and qualitative nature can be seen as a strength of this dissertation. Our studies capture important characteristics of the test scores that influence test interpretations, as well as an insight into the experiences of its users. It is important to note that the qualitative study should be seen as a description of a range of possible views about these instruments. A small-scale selective sampling of teachers was chosen to maximize both the detail of information that could be collected as well as the variation between cases. Contrary to our quantitative studies, this method does not focus on generalizing frequencies and magnitudes to a broader population. Instead, it delves deeper into the underlying beliefs and motivations that teachers have about test use. Although these results cannot be used to answer questions about how many teachers experience these tests in a certain way, the approach is more suitable for answering questions about how and why teachers experience them as such.
Chapter 6

Although the larger samples used in the quantitative studies were largely representative of the general population in the Netherlands (see Appendix B), Northern schools were slightly overrepresented. As a result, there was a small underrepresentation of the proportion of children with a foreign heritage (NNCA). This may have attributed to the overall higher scores, but does not explain the magnitude of bias found in the norms since the proportion of NNCA children is generally small to begin with. Although we are confident that the conclusions of these studies can be generalized to the Dutch population to a high extent, it is hard to say how these conclusions generalize to other early childhood tests. Early childhood assessment takes unique forms between and within different countries. While some instruments may be comparable in form or content it is difficult to ascertain the degree of comparability with other instruments. To allow international readers to draw their own conclusions on this point, we have included detailed descriptions and sources of the evaluated instruments and related our findings to international studies. Although many of the practical and methodological issues described in this dissertation play a role in international research, it is important to keep contextual and instrumental differences in mind when extending these findings to other countries. Comparative studies between assessment instruments, such as those presented by Vincent-Lancrin (2010), may also help generalize these findings.

Besides the external validity of the studies, one could raise questions about the construct validity of the results. The developers of the tests make it clear that the preschool and kindergarten tests measure a different construct than any later test, particularly by using two incomparable scoring scales. However, considering that the preschool/kindergarten tests were designed to measure strong predictors of later language ability (Lansink & Hemker, 2012) or emerging numeracy skills that play an important role in the subsequent development of mathematical skills (Koerhuis & Keuning, 2011), it is reasonable to assume that there is considerable overlap between the constructs. In addition, in practice, comparisons between the tests are made frequently as this is made possible using percentile scores. For the mathematics domain, it stands to reason that a test that measures a child’s emerging numeracy ability (COTAN, 2011) should predict a child’s numeracy ability (COTAN, 2010) in later grades. On the other hand, there are a series of tests for the language domain that measure different, distinct constructs nested within the language domain. In the first study, we chose the spelling tests as these showed strong bivariate correlations and were more consistently administered compared to other tests. However, the spelling tests measure a productive language skill that does not necessarily have a strong theoretical relation with the emerging receptive language skills measured in preschool and kindergarten. As such, we selected the reading comprehension test in later studies as it measures a receptive language skill. In addition, several authors (Gough & Tunmer, 1986; Scarborough, 2009) describe the importance of emerging language comprehension and word recognition skills, which are important components of the kindergarten language test
Discussion

(Lansink & Hemker, 2012), for skilled reading at a later age. As such, comparison of these tests makes sense from a practical and a theoretical perspective. Although there are also different versions of the kindergarten and preschool tests – an old version, a new paper version and a new digital version – we treated these tests as measuring the same construct. This decision was made because research by Lansink and Hemker and by Koerhuis and Keuning (2011) indicates that the tests measure the same latent skill to a high degree. We did account for differences in the norms of the old and new versions of these tests in every study, since the old tests showed a higher degree of norm-inflation.

Throughout this dissertation, we have tried to avoid the term validity as much as possible in relation to our evaluation of these instruments. This decision was made to avoid ambiguity about the subject of this study. Newton and Shaw (2016) describe varying views about what the term ‘validity’ means and what it should entail. Although a full review of these different perceptions goes beyond the scope of this dissertation, we thought it important to address this discussion briefly in relation to our results. Traditionally, validity of an instrument was defined in terms of what a test measures. However, over the last century this definition has been broadened to refer not only to the test itself, but also to its interpretation, its use and its (intended and unintended) consequences (Cronbach, 1971; Messick, 1989; Shepard, 1997). To some extent, these discussions about what validity should entail centers around questions of responsibility (Newton & Shaw, 2016; Shepard, 1997). Since validity is an important concept in determining the quality of instruments, discussions surrounding its meaning tend to devolve into discussions about what aspects of assessment test developers and publishers should evaluate when a new instrument is marketed (Newton & Shaw, 2016).

From our perspective, we believe that the process of evaluating the quality of an instrument should not be limited to the psychometric properties in an initial calibration study. Instead, continuous evaluations should encompass broader consequences that the instrument has on the educational system and on the professionalization of test users. This is important since unforeseen and undesirable consequences may quickly result from the emphasis that is placed on the outcome by one or more stakeholders. Test developers and some of the larger organizations that make use of the results (e.g. the education inspectorate) could share the responsibility for this process. As the developer of these tests, Cito has access to large amounts of longitudinal test data and as such should take a leading role in the evaluation process. With regards to the term ‘validity’, we tend to agree with Newton and Shaw (2016) that this term may have discontinued its usefulness in describing specific aspects of test evaluation. Whilst stating that a decision, result and/or instrument is ‘valid’ can convey a general positive meaning, authors should include more specific concepts that make it clear what aspects of the assessment process were evaluated.

In our study, we used the concept of stability as an important assumption in test interpretation. Although we learned a lot about the meaning of the concept of stability, a question
Chapter 6

that remains difficult to answer is its relation to predictive validity. In the definition of stability that was used in this study, Wohllwill (1973) equates stability with the predictability of later behavior from earlier behavior. This is akin to the definition of predictive validity: the extent to which a test score predicts a criterion obtained sometime after the test is given (Cronbach & Meehl, 1955). Predictive validity is typically studied by correlating test performance with the selected criterion. We chose stability as it takes into account both changes in score magnitude as well as the relation between scores over time. Moreover, stability can be used to specify not only that a score predicts future performance but also how scores are related over time. Finally, stability can be used to describe both inter- and intraindividual performance whereas predictive validity is typically used to describe test performance for the total sample. In our view, stability as defined in Chapter 4 provides a clearer concept in relation to test use for identification.

Implications and recommendations

Many changes in the early childhood assessment system in the Netherlands have already taken place since work on this dissertation started. Most importantly, the minister of education, culture and science, Ingrid van Engelshoven, has made a decision to prohibit the use of tests from the LOVS in preschool and kindergarten by 2021. According to Van Engelshoven (2018), the comparison of individual performance to national averages does not do justice to the discontinuous development of kindergartners and preschoolers. Indeed, our results showed that although the tests provide some meaningful information about individual children’s future scores, the weak stability of these scores presents considerable limitations on the test’s utility for educational decision-making. This means that educators should not base decisions on one or two test administrations. Moreover, decisions should not be based on the perceived growth or decline between two test administrations. It is far more likely that this change is the result of a single peak or plunge in performance than that it represents a structural rise or decline in development. This is especially true for scores before first grade, but should be considered for later test scores as well.

Any studies on the subject of stability should use a clear definition of the type of stability that is being studied and use analytical techniques that support this definition. In Chapter 4, we found that this is often not the case and adapted an analytical framework that might be used to differentiate between different types of stability. Although research on existing data has many benefits, a prospective study that follows the educational assessment process in real time is unmistakably desirable. While this is considerably more costly in terms of time invested by teachers and researchers, such a study could better encapsulate intervention effects and evaluate these as part of the nomological network of the construct being measured (Shepard, 1997).
Besides the discontinuous development of young children, a second argument made by Van Engelshoven (2018) is that educators object to the scholastic format and normative nature of these tests. Our study found mixed results on this issue. While there were indeed educators who had objections to the test format and the use of normative scores, others saw these instruments as a preparation for formal education and experienced these norms as a pleasant confirmation of their own observations. Of greater concern is how teachers invariably perceived these norms as a pass/fail criterion. These beliefs show how the normative judgment of such tests quickly attracts the focus of attention, which draws attention away from other potentially more diagnostic purposes. Although these instruments were designed as a low-stakes assessment instrument to monitor the development of individual children, some teachers feel pressured to perform by these normative judgments. These ideas may be reinforced by Cito’s labeling of all children in the lowest achievement levels as ‘at-risk’ and marking these scores with a red color in the scoring systems. However, there is little theoretical basis for using such scores as an inherent and objective criterion for determining which children are at-risk. The psychometric method used to create these tests places 20% to 25% of children in these categories by design. Although these tests can be used to provide a global estimate of a child’s rank in the population, tying this rank to a criterion that determines risk status is an arbitrary decision. This is because risk status, as defined by these tests, is not determined by a certain minimal performance, but solely by performance relative to others.

Moreover, this system motivates teachers in their attempt to keep children out of the ‘red zone’ either by reacting to these scores or by taking action to prevent scores in these categories in the first place. As a result, the norm scores of these tests are subject to inflation over time. Periodically updating these norms does not provide a definitive solution and may increase dissatisfaction with and focus on test performance. As such, such ‘rank indicators’ may diminish valuable curricular practices that are not directly related to test performance. Although it may seem like an admirable aim to increase the performance of low-scoring children, it is statistically impossible to keep all children out of the lowest scoring 20%. If the norms are continuously corrected, an expected 20% of children will always fall within the ‘red zone’. If increasing the scores of individual children is seen as an important goal, this will inevitably result in a shift of the entire score distribution. Although standardization can provide structure to an assessment process, it does not provide objectivity since decisions about what to test and when to act are inherently subjective. It runs the risk of confusing what is easily measurable with what is important (Roberts-Holmes & Bradbury, 2016). We would plea against the use of simplistic comparative judgments as an evaluative measure, as they quickly turn the educational process into a rat race that increasingly focuses on narrowly defined criteria. Furthermore, test developers should consider how such judgments influence the broader educational process and examine ways that would prevent undesirable
consequences. As noted by Shepard (1990), test developers should realize that test-curriculum alignment is a reciprocal process. After test content is chosen to fit the curricular goals, the curriculum frequently undergoes adjustments in response to the test.

Instead of tests, observations are put forth as the preferred method of assessment Van Engelshoven (2018). Indeed, Cito has already started development of a new observational instrument ‘Kleuters in Beeld’ which will become available in the 2019-2020 school year. Although it should be clear that our research did not evaluate structured or unstructured observations as an alternative means of assessment, we would like to address several potential advantages and disadvantages of this method of assessment. First, observations related to learning goals are generally less invasive in the curriculum. Since children can be observed in their daily curricular activities, this creates a more natural assessment setting and reduces the chance that children’s performance is impeded by unfamiliarity with the assessment format. As such, this may lessen the gap between the assessment outcome and the curriculum. Second, since observations are focused on short daily activities they might be conducted more frequently. This reduces the impact of temporary lapses in performance or chance successes. However, the same jumps in development may make it necessary to observe similar activities repeatedly to get an accurate picture of the child’s limitations and strengths. These considerations are in line with what some authors would term ‘authentic assessment’ (Bagnato, Mclean, Macy, & Neisworth, 2011; Meisels et al., 2010). This form of criterion-referenced testing emphasizes assessment through frequent observation of problem solving in naturally occurring settings. A potential limitation in this respect is that repeated structured observation of all children may be time consuming and difficult to accomplish. Reliable observations may require more extensive training of educational professionals compared to standardized test administrations, which may be costly but also benefits the professional development of teachers (Smoorenburg, 2013). Selective testing of children by the teacher could reduce this time consumption. However, this method may increase the number of false negatives if the teacher fails to recognize potential or emerging (academic) problems.

In addition, by relating the proposed observations to population norms as the government desires (Rijksoverheid, 2018), these observations run the same risk of narrowing current curricular activities. Although it was already announced that the norm system would be different from the one used by the Cito tests and would make use of broader age-related norms, developers should think deeply about the goal for these norms. When doing so, they should realize that monitoring without a clear purpose cannot be defined as a goal. If the main goal is to identify children who are at-risk in their language and/or mathematics development, then norms should be focused on evidence-based criteria that adequately predict which children experience problems in their future school career. Furthermore, because current tests are designed to measure the entire range of performance, they
are less sensitive in the extreme tails of performance. A test that is designed for identification should have especially reliable scores in these extremities. Longitudinal studies should subsequently explore the link between early test performance and later difficulties in language and/or mathematics. It should also be clear that remediation can indeed reduce this risk status. According to Scarborough (2009), ‘there are indications that preschool training that successfully ameliorates early speech/language impairments is not effective in reducing such children’s risk for later reading problems, as it ought to be if those language weaknesses are a causal impediment to learning to read’ (p. 109). If successful remediation does not affect a child’s future problems, then identification of these problems may become meaningless. As of yet the evidence that supports the use of traditional tests in early childhood is unconvincing and incomplete. However, the use of other forms of assessment such as observation should be subject to the same scrutiny.

With regards to the tests that are the main subject of this dissertation, our results show that individual scores are not stable enough to adequately support the recommended score interpretations employed in the educational decision-making process. Perhaps more importantly, the different expectations that the test user and test developer have for the norm scores interact in a negative way that risks an increased focus on the test content and norm scores. It is nearly impossible to maintain representative norms on a test that is used repeatedly, while simultaneously encouraging test users to avoid low norm scores. This will eventually elicit practices that at best do not contribute to the educational process, and at worst may be considered harmful. Although – like the test scores themselves – these results present a snapshot of a situation that is context specific and subject to change, the important issues presented in this dissertation are likely to persist in any context where a simple norm-indicator is used to determine (potential) academic difficulties in children in preschool and kindergarten.