Towards a rationale for screening test in reading in first grade – construct, format and cut off

Bente Rigmor Walgermo, Associate Professor, Norwegian Reading Centre, University of Stavanger
Arild Michel Bakken, Associate Professor, Norwegian Reading Centre, University of Stavanger
Per Henning Uppstad, Professor, Norwegian Reading Centre, University of Stavanger

Keywords: screening-tests, reading, cut-off, concurrent and longitudinal prediction, argument based validity

Abstract

As the development of adequate reading skill is considered fundamental for further academic and life success, teachers and policymakers need high quality screening test to early on identify students in need of extra follow up. Increasingly over the past decade there has been a demand of more thorough documentation of the quality of national assessments (e.g. Evers et al., 2013; Arnesen et al., 2018), often precipitating a focus on psychometric standards. Different from diagnostic tests and tests developed for research purposes on an individual level, group administered tests screening children for reading difficulties tend to base design and cut off on available concurrent data and often focus on more isolated aspects of reading. In this study we argue for a better concurrence of prevailing reading theory and measures of reading. Further, in the litterature there seem to be no consensus for how cut-offs are developed and what is more, there are few qualitative benchmarks to guide cut offs for identifying children in need of support in developing adequate reading skill.

In this study we follow the lines of Kane (2017) who states that there is no such thing as a correct cut-score, however *reasonable* cut scores are applied as the appropriate criterion of quality. The approach of Kane (2017) follows the so called Goldilocks principle which entails that "the standard need to be high enough to achieve the goals of the program, but not so high as to cause serious side effects." (Kane, 2017, p. 28). When putting this much value on reasonableness, Kane moves an argument based approach to validity (Chronbach, 1988, Kane, 2013) at the forefront. This, however, does not exclude other aspects of validity, as validity is considered to be unitary, i.e no aspect of validity is more important than the other.

Introduction

As a part of the national quality assessment system in education, Norwegian schools apply a variety of different test with different purposes and designs, where the tests in use also differ when it comes to degree of documented quality (Arnesen et al., 2018). In this study we provide a rationale for the development of a Grade 1 screening test, as a mean for identifying the poorest readers, the readers who need extra follow up in order to reach curriculum goals in reading (Norwegian ministry of education, 2014). Given that the primary goal of a screening test in reading is to identify the poorest readers, this study will as follows by means for construct validity aim to apply items as close to the given construct at target, in this case reading, and additionally make a statement about how we more specifically calibrate cut off.

Construct and validity

Some children experiences particular difficulties in learning how to read. These students often fall behind in mainstream classroom instruction and tend to remain poorer readers than their peers over the school years (Juel, 1988; Scarbourough, 1995; Torrpa et al; 2014). Early screening tests is a mean for identifying children who will have difficulties in learning how to read have been used for decades (Gredler, 1997; Walgermo et al; 2018) (over 70 years in the American context, 30/40 years in Norwegian). The use of screening test in educational context is heavily influenced by the practice of medical screening, where screening is a mean for detecting symptoms or disorders leading to certifications and application of treatment programs. Screening program in education is by far based on the same assumption, that learning problems can be predicted with more or less the same accuracy as medical disorders. *This assumption presupposes that we in early stages accurate and reliable are able to measure children's potential learning problems.* Nonetheless, accurate prediction of first graders later success is deemed problematic in many ways (Gredler, 1997). *The core of the problem originates from the fact that the purpose of the screening test not only is concerned with to what degree a specific skill is acquired, however the test must also give information of the child's potential to acquire reading skill.*

Considering criterion-related validity, relating to concurrent and predictive validity (Thorndike & Thorndike-Christ, 2014), this study points to the need for rethinking the overall design and philosophy of group administered screening tests in reading by using both concurrent and longitudinal design validating the design of a tool for identifying the weakest readers in Grade 1.

In an historical perspective screening test in Norway like in Scandinavian and English speaking countries have a focus towards the different aspects of the reading process, aspects that a number of studies have proven to predict reading difficulties (e.g. Schneider, Roth & Emnemoser, 2000). These testes are underscored by a view of reading as a process analytical skill (Høien & Lunderg, 1991) where different parts of the test set out to represent different components of the reading process (Engen, 1999; + internasjonale referanser). Consequently, in addition to mapping students reading skill, aspects like letter knowledge and phonemic awareness is assessed with isolated tasks and sum scores within the test. Designing were intended as a signal for Grade 1 teachers as important components in Grade 1 reading instruction. All this were rooted in a theory that reading instruction for struggling readers in particular should have a focus on the parts or components of the reading process (Høien & Lundberg, 1991; 2012). In Norway the mandatory screening test for Grade 1 students have carried the same design for 20 years. Within this timeframe teachers have gathered a lot of experience with the tests and some challenges have emerged that seem indicative to solve for the next generation of reading screening test. These challenges are according to Walgermo et al,. 2018 concerned with: 1) the extent of the tests, 2) inadvertent consequences for classroom practice, 3) incorrect use of test results, 4) different learning pace in formal reading instruction, 5) a high threshold for use of test results.

We argue that the design of screening tests should be guided by both concurrent and longitudinal prediction, and that teachers and the curriculum should constitute the fulcrum for cut-offs when it comes to identifying the poorest readers. This is in line with Kane’s goldilocks principle, that a cut-off should neither be to high or too low, but reasonable. The suggestion for design and cut off based on results from longitudinal prediction and teacher report of number of students in need of intervention, will be discussed with regard to the overall construct, a concept of reading as an interpretation skill (Tønnessen & Uppstad, 2014).

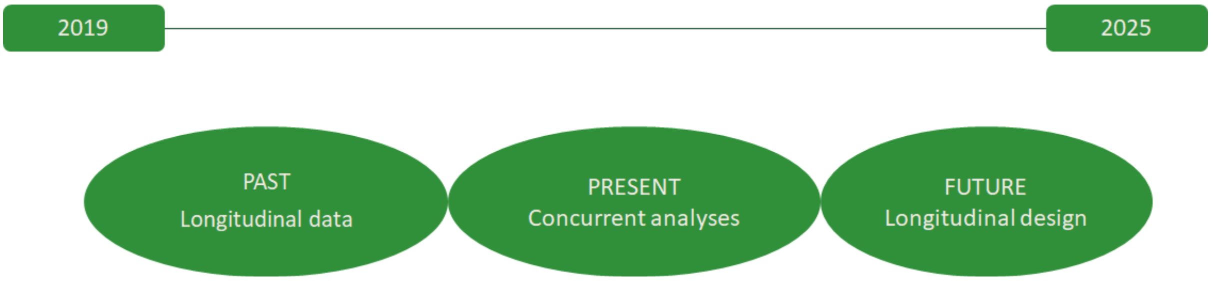

Figure 1
Validation design for the new generation Grade 1 screening tests in reading

The core of our validation efforts have been directed towards both psychometric and theoretical aspects when designing these measures for classroom use as screenings for reading difficulties – at the end of Grade 1. Like figure 1 illustrates the validation design build on three lines of data:

- **logitudinal analyses - from the past** (2014 - 2018)
  From the longitudinal RCT On Track the previous screening tests are used in
- **cocurrent analyses - from present** (2019 -) piloting of the new tests
- **longitudinal analyses - in the future** (2025)

A two-tier model for setting cut offs in screening tests of reading

Central aspects of a "full-skill" logic are, however, often transferred unchanged to tests with a more specific purpose, such as identifying those readers who are likely to fall behind in reading development. The main challenge of this general approach is that it only to a limited extent will be able to discern variation that is due to general reading development from variation stemming from a more specific reading disability. Consequently, the students' reasons for responding to test items are considered alike, and cut-offs are likely to be set at a convenient level of the population as a whole. While the vagarity of cut-offs, however, will prevail, meaning that a cut-off has to be set at a certain convenience level, it can be assumed that the design of test items can be better tuned to identify the unique behavior of those students who the test aims to identify, i.e. the items would be more strongly shaped on the grounds on knowledge of the students we aim to identify by the test, i.e. students running the risk for developing reading difficulties. This represents a change from the established approach, by moving knowledge of errors related to reading difficulties from the background into the foreground.

Traditionally, a requirement for reading screening test development has been that items should be in a specific difficulty range considering the whole population, i.e. the item should be mastered by between 70 and 90 % of the students. However, by the use of IRT, the scrutinization of very simple (>90%) items has become feasible and sophisticated.

In the exploration that follows, we will apply a heuristic model of how to approach the issue of both item development and cut off for this specific purpose (see figure 1). In this approach, we first maximally build items around knowledge of the poorest performing readers. Next, when piloting these items we will be able to identify an inner tier of low performers (figure 1), committing errors which a very high percentage of students typically would not do (90-

99%). These items are built on knowledge on the kind of errors students with high risk of developing reading difficulties would do. The establishment of these items forming a separate tier is motivated by their relative high discrimination in IRT analyses. As the percentage indicates, however, we only identify up to 10 % of the students by this first tier items.

The second tier is different from the first, as these items - still being easy - in addition also tap general developmental aspects of reading. This is reflected through piloting, showing a lower discrimination and greater variation in skill level. From the perspective of longitudinal prediction, this means that items in this category will contribute to the identification of false positives in a way that items in the inner tier don't. In line with this reasoning adds an assumption that items of the inner tier will have better predictive value for reading difficulties than items of the outer tier. To put it short, frameworks expelling items above the 90% percentile may have come to expel the core identifiers for the groups of students that the test aims to identify. This assumption is underpinned by the fact than only a small percentage of students (below 5%) are considered to have severe deficits when it comes to reading.

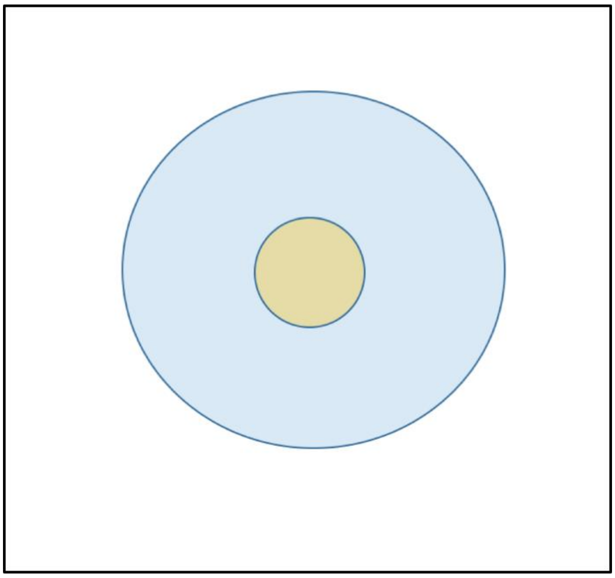

Figure 1. A two-tier model for setting cut off in a screening test for reading difficulties

The inner circle denotes items that directly target students with reading difficulties by presenting items of the unique errors these students are likely to make. Items of the inner circle are characterized by a very high ease (90% +). The outer circle represents items that to a larger extent tap the low performance of students relative to the full population. The two tiers are assumed to play different roles when considering validity (criterion based validity versus concurrent validity)

To some extent, we inadvertently arrived at this point, nonetheless, it can be seen as an consequence of not paying enough attention to the development of more domain and content-specific principles for item construction. To give an example: for an item that is mastered by 60 % of the students, the poorest student will fail - as will also a quite large percentage of students (<40%). For an item that is mastered by 90% of the students, and where the item is carefully built on knowledge of the characteristics of struggling readers, the test is likely to identify at risk students with higher reliability and validity.