frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 25 June 2024
DOI 10.3389/feduc.2024.1331777

OPEN ACCESS

EDITED BY
Tove Stjern Frønes,
University of Oslo, Norway

REVIEWED BY
Gustaf Öqvist Seimyr,
Karolinska Institutet (KI), Sweden
Issarapa Chunsuwan,
Thammasat University, Thailand

*CORRESPONDENCE
Bente Rigmor Walgermo
✉ bente.r.walgermo@uis.no

RECEIVED 01 November 2023
ACCEPTED 03 June 2024
PUBLISHED 25 June 2024

CITATION
Walgermo BR, Foldnes N, Uppstad PH, Bakken AM and Lundetræ K (2024) Development and deployment of an adaptive national elementary reading screening test. *Front. Educ.* 9:1331777. doi: 10.3389/feduc.2024.1331777

Development and deployment of an adaptive national elementary reading screening test

Bente Rigmor Walgermo*, Njål Foldnes, Per Henning Uppstad, Arild Michel Bakken and Kjersti Lundetræ

National Reading Centre, University of Stavanger, Stavanger, Norway

Increasingly over the past decade, there has been a demand of more thorough documentation of the quality of reading assessments. Yet, available documentation of high-quality measures are often restricted to general framework descriptions providing psychometric information as a token of test quality. In a modern view of validity, understanding what is being measured and how scores are calculated is a prerequisite for good interpretation and use of test scores. The present article aims to document the research and development process of a national adaptive screening test for reading difficulties, in which the envisioned interpretation and use of test scores is guiding the work. Given the mandatory nature of this test the sample consist of 49,828 third grade students aged 8. Significant outcomes from this design process involve detailed information on: (a) choice of sub-tests and item formats, (b) selection of high quality items, (c) choice and design of adaptive model, and finally, a statement on the challenges that are still to be met for such a test to function optimally. The present paper is among the first to, in an open and detailed manner, describe the development process as well as qualities and challenges of an adaptive reading screening test for students of this young age.

KEYWORDS

adaptive testing, screening for reading difficulties, multistage, routing, screening, early readers, meaningful tasks

1 Introduction

Reading is a fundamental skill that underpins central facets of academic achievement and societal participation. Accurate and comprehensive assessment of an individual's reading abilities is imperative for effective educational interventions. In this, mandatory screening tests play an important role in supporting teachers' identification of students that need intensive reading intervention. And as the use of standardized tests in education has increased worldwide, a demand for more thorough documentation of the quality of reading assessments has been put forward (Evers et al., 2013; Arnesen et al., 2019). Arnesen et al. (2019) suggest that the absence of documentation is due to a lack of expectation from society and end users that assessment quality should be explicitly stated: "*As an example, for mandatory national tests and assessments, this information exists only as internal documents or technical reports. Consequently, important information about the assessments' quality is hidden from the end-users for some instruments, while it is communicated for others*" (p. 485). While this quest still awaits fulfilment, the assessment field is rapidly developing.

Traditional reading assessments often utilize fixed, linear tests, providing a static evaluation of an individual's reading capabilities. However, these one-size-fits-all approaches may not capture the nuanced and diverse specter of reading,

leading to imprecise measurements and potentially inappropriate instructional practices. Recognizing this limitation, there is a growing interest for developing adaptive reading tests that tailor assessment experiences based on the test-taker's responses, allowing for a more personalized evaluation of the individual test taker. Adaptive assessments can measure more precisely, if they present students with items that are optimally informative for the estimated skill level of that particular student. This enhanced precision may be converted into a shorter test time, which would make for a better test experience, especially for the struggling readers. Test experience could also be enhanced qualitatively by adaptive testing, as the items the student is asked to solve are more adapted to their skill level. This pertains in particular to struggling readers, who would get easier items than in linear tests and thereby experience more mastery. Adaptivity clearly has much to commend it in assessments for struggling readers. However, the increasing provision of adaptive assessments is posing new demands on documentation and transparency. Illustrative for this demand is the debate related to the former mandatory national Danish adaptive tests (Bundsgaard and Puck, 2016). Central to the termination of the Danish tests was uncertainty and discontent related to a general lack of transparency about how the test was made, what the content was, how the algorithms worked, and accordingly how valid the test scores were (Flarup, 2020)—i.e., a typical black box-problem, also relevant in other fields using AI based assessment (Brożek et al., 2023).

To provide meaningful feedback and support to students, both teachers and teacher educators need to understand how assessment decisions are made. If the algorithms used for scoring or grading are opaque, it becomes challenging for educators to explain or justify assessment results to students, parents and other stakeholders (Farrow, 2023). This is in particular problematic for adaptive tests used for educational purposes in classrooms, where the traditional, physical test material is dematerialized into algorithms and digital systems (Jiao et al., 2023), and where the human test administrator is replaced by the computer.

In Denmark, the black box-problem can be said to have rendered Danish policy makers and even researchers incapable of having a constructive, critical debate on the issue (see e.g., Andersen et al., 2019; Flarup, 2020), leading up to a polarized situation of *pro et con*. In Wales, National digital, adaptive assessments are less debated—all children aged 7–14 undertake mandatory formative personalized tests in reading, procedural numeracy and numerical reasoning. The digital Welsh adaptive tests were developed in close dialogue with the Danish test developers, from their hands-on experience, and could therefore be considered second generation adaptive tests. Although the Welsh tests are reported to function well according to teachers' experiences with using the tests in classrooms, no scientific open documentation of their construction and functioning exist. New, extensive documentation is likely needed to inform and advance a constructive debate about digital adaptive assessment in society. A token of validity is however that the test developers of the Welsh adaptive tests, Alphaplus, won the EEA (E-Assessment Award, hosted by the E-assessment Association) for best use of formative assessment for the Welch adaptive tests in 2020.

The purpose of the new Screening of Reading Difficulties addressed in the present study is to support teachers in identifying students who need extra follow-up in reading. The quality of such a test is closely tied to its validity, i.e., "the degree to which evidence and theory supports the interpretation of test scores for proposed uses of test scores" (AERA et al., 2014, p. 1). This definition of validity emphasizes that the concept exceeds the boundaries of the test itself. Stressing the difficulty in obtaining this level of quality, Kane (2015) states that potential scenarios for interpretation and use of test scores should be highlighted systematically even before the development of a test starts. In order to follow this recommendation, prior to the current test development, Walgermo et al. (2021) stated challenges in the actual interpretation and use scenario based on experiences with interpretation and use of past generation linear tests. They highlighted the following challenges: (a) past generation screening tests had unintended negative consequences for classroom practices, e.g., items intended only to raise teachers' awareness of difficulties with reading became subject to rote learning, (b) test scores were often incorrectly interpreted, for example, high scores were taken as a sign of high skill, an interpretation not warranted by the test because of its ceiling effect, (c) the test had a disproportionate duration and difficulty (too short and easy for the students not at risk, too long and difficult for the at-risk students), leading to a negative test taker experience, (d) the fixed time window for taking the test was considered rigid by many teachers, as instructional practices and progress differ largely, and (e) teachers had a high threshold for acting upon test scores, due to a lack of resources or knowledge about how to help struggling readers. In summing up the scenario for interpretation and use, Walgermo et al. (2021) point to the need of using "all means available to gear a new test concept to the original purpose of the test" (p. 8) namely identifying the students most at risk of struggling with reading. The question is then how the test can be constructed in a manner that increases the probability that the results will be interpreted and used in accordance with this purpose.

The present paper aims to communicate in detail the research and development process of a new national adaptive screening of reading difficulties. The test was deployed as mandatory for all Norwegian third graders and carried out at a national level from autumn 2022 (October). As emphasized above, the current initiative offers transparency for all stakeholders with interests in the measurement of students' skills. The scientific contribution resides in detailed step-by-step information to test developers worldwide who intend moving into the field of adaptive testing for young students, but lack opportunities to learn from choices made by other test developers. Accordingly, we first present the rationale for the new test, the construct of reading and which aspects of the construct are measured. Next, we describe the choice of adaptive model and the operational tests. We discuss data from the first deployment of the test, including the responses of 49,828 students. Finally, we discuss the challenges and dilemmas that remain.

2 Background for the new test

The present screening test replaces a past generation reading screening tool with roots back to the early 1990s and later versions

of these (Engen, 1999; Tønnessen and Solheim, 1999). The old tools came as a response to teachers' request for material that could support their identification of struggling readers, and obtained the status of a mandatory national screening test from year 2000 and onward (Ministry of Church and Research, 2000). As mentioned above, the design of the new test builds on a review of challenges concerning the old, where the following aspects are emphasized in the development: (a) to bring the test up to date with current theories of reading (and writing), (b) preference for sub-tests that have documented longitudinal prediction, (c) the acknowledgment of teachers' needs, interpretations and uses of screening tests, (d) an expressed need for a shorter and more precise test, and (e) to support and maintain the students' engagement for reading (Walgermo et al., 2021). Central to the development has been a recent extension of the purpose of assessment in the national assessment regulations stating that the purpose of assessment in the subjects is to promote learning and contribute to the desire to learn during instruction, and also to provide information about competence both during and at the end of instruction. While this change does not mean that all evaluation has to lead to a desire to learn, it emphasizes that the way assessment is conducted will have an impact on the students' motivation, self-beliefs (Walgermo and Uppstad, 2023), and interest for further learning related to the subject. This is particularly important for struggling readers, for whom experiences of mastery—or lack of the same—in the assessment situation is likely to have greater impact on their reader self-concept than would be the case for students who are accustomed to master the reading challenges they face (Bandura, 1997). As mentioned, making the test adaptive can increase the sense of mastery of struggling readers, by ensuring that they are presented with easier items, but also by shortening the test. The new tests will therefore be adaptive tests. The Directorate for Education and Training did not agree to the development of an item-level adaptive test (traditionally known as computerized adaptive testing or CAT), but permitted the researchers to develop a multistage test (MST). An MST consists of preassembled modules of items presented in different stages, and is only adaptive after each stage, not after each item.

As Norway has two mutually comprehensible written languages (bokmål and nynorsk), two versions of the test were developed in parallel. For the old screening test, the nynorsk version was obtained by a translation from bokmål. In the new test, the final design is based on pilots in separate samples for the two written language forms. Quality requirements state that the items and their distractors have to function optimally in both language norms. The current solution gives us two original tests of high quality, instead of one master test in bokmål and an additional translated test in nynorsk.

The new national screening test contains sub-tests that have documented prediction on later reading skills (Walgermo et al., 2021). Interestingly, it is the sub-tests that are closest to reading *per se* that are shown to have the highest predictive value. As reading researchers, we consider this to be important and of positive value, as a test containing measures of actual reading is likely to guide the teachers' awareness more directly toward the act of reading, instead of a focus toward more atomized, theorized aspects of processes involved in reading (e.g., blending and first phoneme isolation tasks). Several decades ago Madaus and Keillor (1988) stated that "it is testing, not the 'official stated curriculum, that is increasingly determining what is taught, how it is taught, what is learned, and how it is learned". This points to so-called "washback" effects, defined by Messick (1996) in the context of language testing as "the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or hinder language learning". Messick (1996) further stated that washback only is problematic when construct validity is threatened: "in the case of language testing, the assessment should include authentic and direct samples of the communicative behaviours of listening, speaking, reading and writing of the language being learnt. Ideally, the move from learning exercises to test exercises should be seamless. As a consequence, for optimal positive 'washback there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test." Messick (1989) also introduced a validity framework where validity was seen as a judgment of how appropriate inferences and actions based on test scores were. It was this "unified model of validity" that Kane (2013) later re-framed in a more practical sense as *argument-based validity*, proposing the interpretation and use of the test scores by stakeholders as the most important aspect of a test's validity.

3 Choice of sub-tests and item formats

When choosing sub-tests and item formats we have aimed to minimize the young students extraneous cognitive load derived from instructions and task design (Hollender et al., 2010). Efforts were made to ensure that the tasks were as user friendly and intuitive (Lehnert et al., 2022) as possible for the age group within the given digital interface.

The skill of reading is an entity that is best reflected in reading fluency. The performance of a skill can be characterized as a flexible combination of automaticity and awareness, in which more complex parts of the text are read with greater awareness—and effort—than less complex parts. In this perspective reading fluency is best defined as "*thinking one's way through a text without the written medium obstructing one's thought*" (Tønnessen and Uppstad, 2015). In order to read with fluency, the student has to understand the text. Fluency therefore implies understanding, and can serve as an indicator for a more general reading competence (Fuchs et al., 2001). The four subtests included in the present adaptive screening of reading difficulties can be traced as having impact on the extent to which the student is able to read with fluency:

- word reading
- vocabulary (linguistic comprehension)
- word writing (spelling)
- reading comprehension of sentences and short texts

These aspects of reading have been chosen by the fact that they tell us more about later reading skills than other aspects. Walgermo et al. (2021) document that sub-tests in past generation

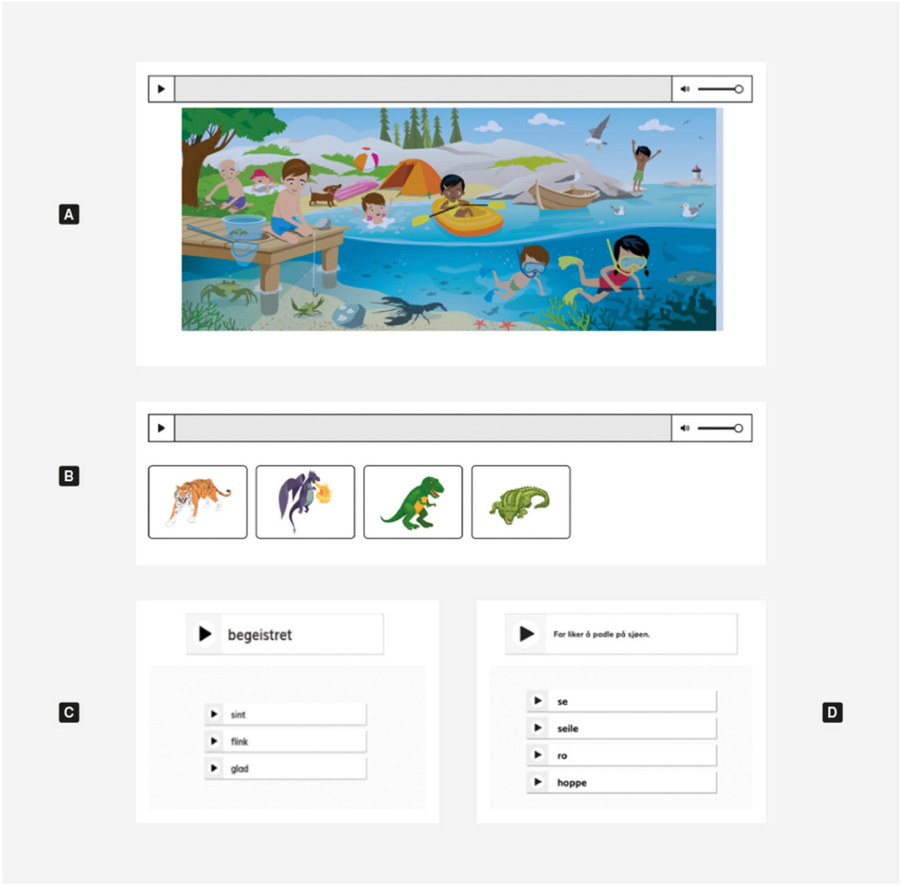


FIGURE 1
Examples of a all vocabulary tasks included in the screening test: Format **(A)** "Click in the picture," Format **(B)** "Find the right picture", Format **(C)** "More isolated synonym tasks," and Format **(D)** "Synonym tasks with more context".

tests measuring reading processes beneath word level (blending, isolated letter knowledge, first phoneme isolation) had small or no predictive value from first grade unto third grade. Rather, it was the sub-tests word reading, writing (spelling) and reading comprehension (sentence reading) that predicted difficulties of reading in third grade in particular. One could therefore say that the focus of sub-skills in the old screening tests either are less important than was assumed when the test was crafted or have gained lower prediction, as early literacy skills have increased due to a change to an earlier onset of reading instruction. In this test, actual reading (word reading, sentence reading, and text reading) constituted most of of the assessment. In addition, we investigated spelling, as it is a well-known predictor of reading difficulties (Graham et al., 2021), and vocabulary, as it predicts later reading comprehension (Quinn et al., 2015). On these empirical grounds, the current screening test in reading contains formats of reading (word reading, sentence reading and text reading), spelling and vocabulary, putting aside sub-tests from the old tests related to analysis of sounds and underlying processes. These changes are also intended to have beneficial effects in guiding teachers' awareness toward actual reading, instead of the focus on underlying processes that demands a much deeper theoretical understanding from the teacher in order to translate into good classroom practice. The vocabulary sub-test is included to measure aspects of the students' linguistic comprehension. In the following, we will describe the chosen formats in more detail.

3.1 Item formats targeting vocabulary (linguistic comprehension)

Our knowledge of words and language develops incrementally (Nagy et al., 2000; Alexander, 2005). Consequently, in order to measure the students' linguistic comprehension at different levels, we have developed and piloted three different item formats ranging from identification of common everyday words/items in a picture to more nuances knowledge of synonyms. All vocabulary formats are presented to the students with sound support. This is important for these tasks to measure language comprehension *per se*, and not reading skill.

3.1.1 Item format 1A: *click in the picture*

This format measures a fundamental understanding of everyday concepts like for instance "boat," "ball," "beach." A voice tells the students to click on a specific item in the picture. Also,

some tasks address prepositions, e.g., "Click on the bird *on* the boat", "the crab *under* the jetty" (see Figure 1A).

3.1.2 Item format 1B: *choose the right picture*

Here the student is presented with four pictures representing words that are somehow related when it comes to meaning, e.g., "rocking horse," "rocking chair," "hammock," and "wheelchair" (see Figure 1B). A voice tells them what item to click on.

3.1.3 Item format 1C and 1D: *find the words that have the same meaning*

Within this format nuances of students' word knowledge is measured using two different synonym tasks: one variant where the target word appears isolated (see Figure 1C), and one second variant where the key word appears in context, in a short sentence (see Figure 1D). A voice tells the students what item to click on. The target item and distractors are also read aloud to the student. While isolated synonym tasks have been questioned in the assessment literature over the past decades (Pearson et al., 2007), mainly due to theories embracing the value of context, recent findings point to their validity in assessment (Walgermo et al., in review).

3.2 Item formats targeting word reading

A prerequisite for fluent reading is stable and automatized decoding skills at word level. In supporting struggling readers, it is considered necessary that the instructor models and explains to the students' different strategies for reading words (Ehri, 2015). Reading words involve recognition of written words based in the alphabetical principle. The notion of word reading is reflected in three different competence aims after 2nd grade in the Norwegian curriculum, i.e., (a) play with rhyme and rhythm and listen to identify the various speech sounds and syllables in words, (b) combine letter sounds into words when reading and writing, and (c) explore and talk about the structure and meaning of words and expressions.

Students' word reading skills develop rapidly over the first years in school. Results from a longitudinal research project including 5,000 Norwegian primary school students (Solheim et al., 2018) show a large diversity in 3rd graders' (8-year-olds) word reading skills. The variation at this point is related to degree of automatized word reading. Consequently, the current subtest measures word reading skills in the span from synthesizing sounds to form words, to rapid and accurate automatic recognition of words. This led us to include two different tasks in the current screening test, tasks both with permanent and timed stimuli (see Figure 2).

3.2.1 Item format 2A: *word reading (decoding) with permanent stimuli*

In the early phases of reading development, students acquire knowledge that enables them to decode words. Word decoding requires knowledge of the letters, to associate written symbols with their corresponding sounds, and to synthesize these sounds into

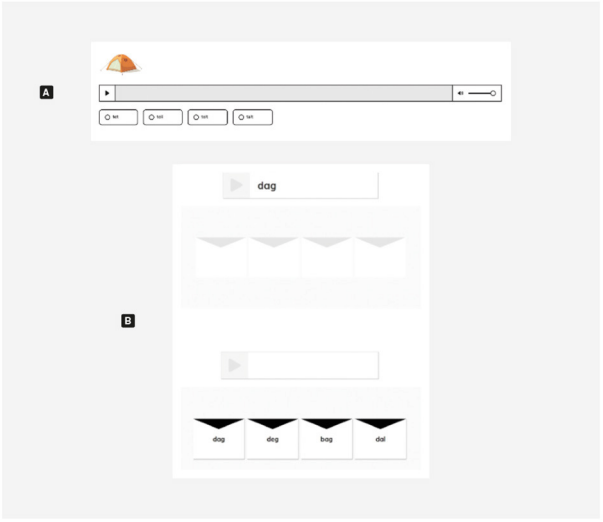


FIGURE 2
Examples of both word reading items included in the reading test. From the top: format **(A)**, item with permanent (not-timed) stimulus: "tent" (telt). The figure further shows, format **(B)**, "day" (dag). The first and second part of a word reading task with timed stimulus. In these tasks the target words are visualized with different time slots: 2 seconds, 1 second, 500 milliseconds or 200 milliseconds. Stimulus words are presented and disappear permanently before four response alternatives are presented. The student is supposed to re-find the target word among the four alternatives.

(meaningful) words. This skill is measured with the format *Word reading (decoding)—with permanent stimuli*. In this item format, a picture is presented to the student during the time it takes to respond. This format is also supported with a sound file that is optional for the students to take advantage of, ensuring that the student will not be uncertain concerning what the target word is (see Figure 2A).

3.2.2 Item format 2B: *word reading (recognition) - with timed stimuli*

By decoding the same words several times in different contexts, the students develop word recognition skills. This involves knowledge of parts of words (e.g., frequently occurring letter combinations) as well as automatic recognition of letter patterns of whole words (Adams, 1994). Word specific knowledge renders the students capable of recognizing words rapidly and accurately—without performing a phonological synthesis. In this way the students' word reading skills get automatized. This automatized word reading skill is measured with the format *Word reading (recognition)—with timed stimuli* from the Adaptvurder-project (Bakken et al., 2023) developed by Rønneberg et al. (in review).

A limitation to both word reading formats is that the they do not render information on whether the students have gained access to the meaning of the words tested. The items therefore only primarily measure whether the students map the written word with the spoken word (see Figure 2B).

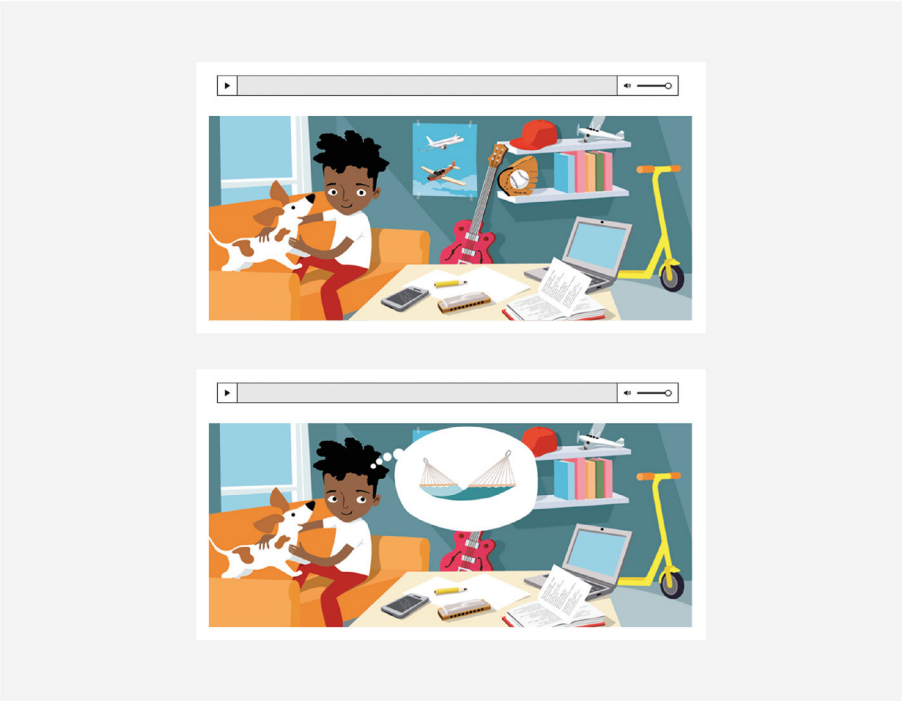


FIGURE 3
In the first picture, the student is introduced to the spelling format: "Hi, this is Olaf who is soon celebrating his birthday. He is very exited about it. Can you help Olaf write a wish list?" The second picture display an actual task: "Hammock" (hengekøye)—write "Hammock".

3.3 Item format targeting word writing (spelling)

Piloting indicates that students find traditional spelling tasks, i.e., the writing of a word out of context, demotivating. Acknowledging that students' motivation for writing is dependent on the writing task (Alves-Wold et al., 2023, 2024) and also in accordance with state-of-the-art? theories of reading and writing, the spelling format in the present test builds on the view that reading and writing are acts of communication (Tønnessen and Uppstad, 2015) and that this element of communication is the motivational driving force in the learning and exercise of these skills. Consequently, in this item format, students are given a clear purpose for their word writing, they are asked to help a boy—Olaf—in writing a wish-list. The students get to see an illustration of a boy in his room, and in a thought bubble Olaf's wish appears as a picture. The student is then asked to write Olaf's wish. These word writing (spelling) items are automatically supported with sound, i.e., the spoken word to be written is given alongside the picture of what is to be written (see Figure 3).

3.4 Item formats targeting reading comprehension

In order to measure accurately—as well as securing mastery for the lowest performing readers - the reading comprehension sub-test involves two formats: a format of sentence reading and a format assessing text comprehension. While reading is an interpretive skill, we here measure the product of interpretation, i.e., comprehension (Tønnessen and Uppstad, 2015; Walgermo et al., 2021).

3.4.1 Item format 4A: *sentence reading*

For the lowest performing readers, the test assigns reading comprehension items that include only one or a few sentences. This task is important in order to measure with precision in the lower end of the performance scale, but also to ensure that the poorest readers will experience a sense of mastery when taking the test. This format is inspired by formats in two Danish reading tests, i.e., SL 60 and SL 40 (Nielsen et al., 1986) (see Figure 4A).

3.4.2 Item format 4B: *reading of authentic texts*

In this format, the emphasis is on using authentic and motivating texts. As for the sentence reading format, the consideration for the lowest performing students (i.e., the target group) is essential. An increased number of short texts has therefore been included, at the cost of longer and more complex texts. The texts included are all piloted in dialogue with students in the target group. While so-called item writers are much used in the field, i.e., writers composing short text items on the basis of a prescribed scheme, all text included in the present screening are either authentic texts chosen because of their

10.3389/feduc.2024.1331777

FIGURE 4
Item examples: Format **(A)**, Sentence reading "Click on the picture that best describes the text". Format **(B)**, Reading of authentic texts: "Read the text and answer the questions".

qualities or high quality literary texts developed specifically for the test by acknowledged children's book authors (see Figure 4B).

3.5 Instructions and learning environment

Prior to the annual deployment of the screening test, digital example tasks are made available for parents and teachers in order to prepare the students for the test. The test was carried out in the students' respective classrooms with their teacher introducing them to the test and modeling the example tasks prior to the test. The test was carried out on the students' personal Chromebook or IPad. All students used headphones while carrying out the test. In addition, short example items are provided within the test itself every time a new task type is introduced to the students.

3.6 The Norwegian school context

In Norway children start formal reading instruction when they enter school in August the year they turn six years. Prior to school start 97% of Norwegian children attend the barnehage where literacy skills are promoted through different playlike activities largely driven by the children's own initiatives. The Norwegian language holds a semi-transparent orthography somewhere in between English and Finnish when it comes to orthographic depth (Seymour et al., 2003).

4 Designing the adaptive screening of reading difficulties

4.1 Pilots—Timeline and procedure

Prior to the full deployment, the screening of reading difficulties were piloted in two rounds. First a linear pilot was carried out for item quality estimations. Second, a pilot with the purpose of testing out the adaptive model was conducted (see Figure 5 for a timeline). Given that the test is mandatory for all Norwegian third graders (from October 2022), the linear pilot was administrated in September 2021. The pilot data were rapidly analyzed and the adaptive model constructed to be piloted in November 2021.

4.1.1 The first linear pilot, with the purpose of obtaining IRT quality specifications

Initially a large number of items for the four sub-tests were developed by reading researchers. Then all items were distributed into 16 different digital booklets (eight for each of the two different Norwegian written language norms, Nynorsk and Bokmål). Each booklet consisted of a total of 138 items, resulting in a total of over 2200 piloted items in spelling, vocabulary, word reading and reading comprehension. According to recommendations for sound IRT-calibration, we had an overlap in items between the booklets of 25%. Each third-grade students that took part in this pilot carried out two different booklets within a two-week period during September 2021.These booklets were linear, meaning that all participating students were presented with identical items in

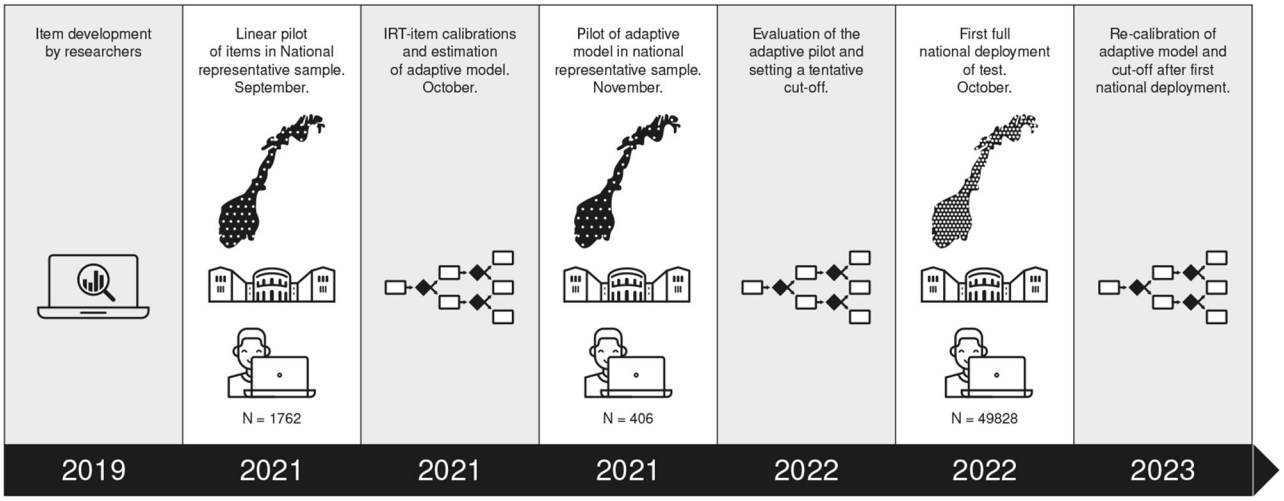


FIGURE 5
Timeline showing phases in developing the adaptive screening of reading difficulties within national representative samples, including the first full deployment.

identical order. The tests were carried out on students' personal computers (Chromebook or iPad).

Item parameters were obtained using a pilot sample of a total of 1,762 third graders, see Figure 5. The number of students that carried out each booklet varied from 214 to 514. The sample of students stemmed from small and large schools from urban and rural areas from every municipality in the country.

Item characteristics were estimated using two-parameter item response theory (IRT) (De Ayala, 2022), where each item is associated with a difficulty and a discrimination parameter. This model is among the most widely used IRT models. The purpose of the screening test was to identify the 20% poorest performing students, a task that do not require a very high level of measurement precision in itself. The two-parameter model was therefore considered convenient for the purpose. The difficulty parameter is located on the same scale as the student ability estimate. This scale comprises both negative and positive values and student ability scores are assumed to be standard normally distributed. A difficulty parameter of, e.g., 0 is interpreted as a probability of 0.5 of getting the item right for a student with ability 0. The second item parameter measures the discriminatory power of the item, i.e., the extent to which the item can discriminate between low and high ability students. The higher the discrimination, the more informative the item is considered to be. IRT calibration was done in R (R Core Team, 2020) with the use of the package mirt (Chalmers, 2012).

Plots of the item parameters for each subtest are provided in Figure 6, with associated summary statistics given in Table 1.

4.1.2 Piloting of the adaptive model

In November 2021 the adaptive MST-model was piloted. Four hundred and six students undertook the test.

4.2 Specifying the MST

A simple six-module Multistage testing (MST) design was adopted, as depicted in Figure 7, where each module consists of four to six items with roughly the same degree of difficulty within each module. For each of the six modules (M_1- M_6) a difficulty level is defined. Each module is constructed from items whose difficulty is close to the nominal difficulty level prescribed for the module. That is, for each module we choose items in our item pool whose difficulties come close to the prespecified difficulty level, while at the same time maximizing item discrimination. Under such conditions the sum score, i.e., the number of correct responses in the module, is a sufficiently precise measure for the students' skill level. We therefore use sum score as our skill measure rather than more advanced IRT measures such as MLE or EAP, as these are highly correlated with the sum score under the current design. Another reason for using sum scores was a more pragmatic one. Our license with the system provider for delivering the MST did not include IRT estimation.

In sum, all participating students (N = 49,828) carried out a total of 68 items, distributed on all four sub-tests. The average time for carrying out the 68 tasks in the test was 20 min. No students worked with the tasks for longer than 35 min before finishing the test. Additionally, all students watched four instruction videos lasting for a total of 3 min.

All students start the test by completing the same initial module M_1 ("Easy items") comprised of items with fairly low difficulty. Then, based on the obtained sum score s_1 students are routed in stage 2 to either module M_2 ("Easier items") if $s_1 < t_1$, or to module M_3 ("More difficult items") if $s_1 \geq t_1$. The same routing logic is applied at modules M_2 and M_3, leading students to one of three modules $M_4 - M_6$ in the final stage. The generic MST design was the same for all four sub-tests, so in the following we focus our presentation on the word reading sub-test. We next describe item selection and threshold calibration. Detailed information of item

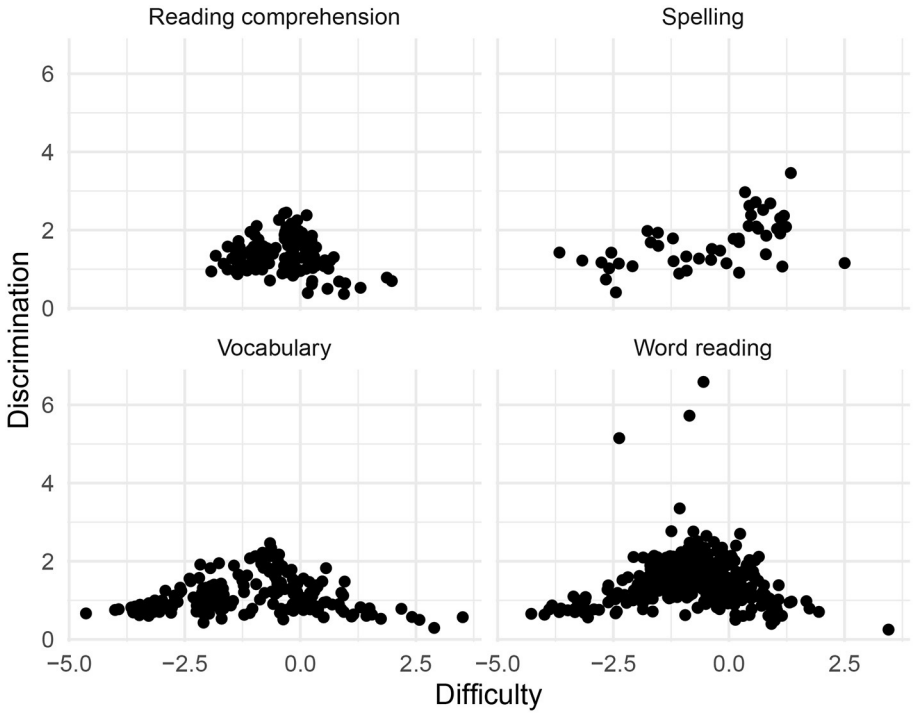


FIGURE 6
Item calibration: IRT parameters for each calibrated item in four subtests.

TABLE 1 **Item calibration: size and IRT parameter summaries for the item pool in four subtests.**

Subtest	Number of items	Mean difficulty	Mean discrimination
Reading comprehension	122	-0.38	1.37
Spelling	46	−0.43	1.69
Vocabulary	177	−1.08	1.13
Word reading	425	−0.84	1.54

selection and threshold calibration for the sub-tests of Vocabulary, Word reading and Reading comprehension can be obtained by contacting the authors. Due to the fact that spelling tasks are more effortful and time consuming to perform—in particular for the struggling students—we had to include fewer spelling items in this sub-test compared to the other subtest.

4.2.1 Word reading MST: designing the first version

Each of the six modules corresponds to a prespecified difficulty level. Referring to Figure 7, we see that the start module consists of relatively easy items. The ideal difficulty level for these items was chosen to be −1. Also, the increment between modules was chosen to be 0.4, which means that items in M_2 have difficulties close to −1.4, and items in M_3 have difficulties close to −0.6, see Figure 8. Furthermore, in stage three the ideal item difficulties were −1.8, −1, and −0.2 for modules M_4, M_5, and M_6, respectively. For each module, we identified from the item pool six items with difficulty close to the specified difficulty level, with as large discrimination value as possible. The resulting item difficulties are plotted in Figure 6. As we can see, the item difficulties in a given module are approximately centered around the ideal difficulties. In Table 2 show mean and standard deviations of item difficulties and discriminations across the six modules for word reading.

The next step in MST specification was to calibrate the threshold values $t_1 - t_3$. These thresholds determine which stage 3 module a test taker is routed to. To simplify the calibration of thresholds and help better understand the MST we also included in our analysis thresholds $t_4 - t_6$ for the stage 3 modules. These threshold were used to introduce a more fine-grained grouping of test takers into four levels. The four levels are labeled risk, low, medium and high. We remark that most students ended up in the medium and high groups, so these label names are only used for convenience to express the ordering of levels. The risk group consisted of individuals who scored below the threshold ($s_4 < t_4$) in M_4. To be classified in the low group, a student could take three possible paths: $M_1 \rightarrow M_2 \rightarrow M_4 : s_4 \geq t_4$, $M_1 \rightarrow M_2 \rightarrow M_5 : s_5 < t_5$, and $M_1 \rightarrow M_3 \rightarrow M_5 : s_5 \geq t_5$. The medium group consists of test takers that scored at least t_5 on M_5 and those that scored below t_6 on M_6. The remaining test takers are labeled as high performers, and these are students who scored at least the threshold value at all stages. Although the grouping of test takers into four performative groups was not part of our mandate, we found it useful when

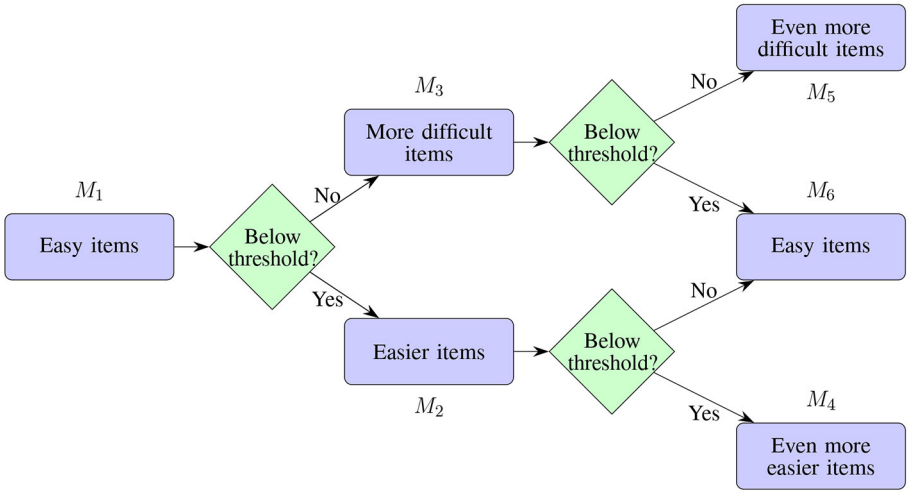


FIGURE 7
Generic design for MST. The number of items in each module, and the thresholds values, may vary for each subtest.

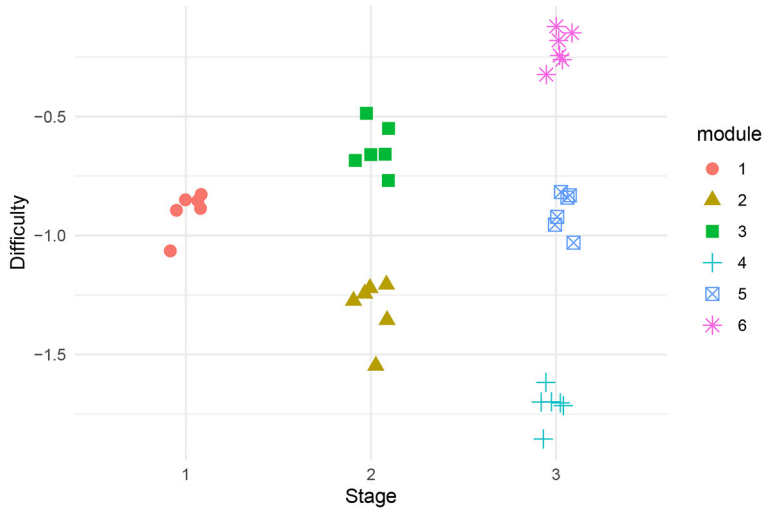


FIGURE 8
The difficulties of items in the word reading MST.

thinking about MSTs. In addition, in future implementations it may be useful to furnish teachers with a similar, more fine-grained assessment beyond the risk vs. no-risk dichotomy.

The calibration was conducted by an exhaustive search among plausible threshold sets. In each iteration, we fixed thresholds t_1 to t_6 and then calculated the proportion of test takers ending up in the four subgroups, looking for settings where the risk and low groups comprised ∼20% of test takers. These calculations relied on IRT-deduced probabilities and on carefully enumerating the possible paths throughout the MST that leads to each category. For instance, to end up in the low group, three trajectories are possible

- $M_1 : < t_1 \rightarrow M_2 : < t_2 \rightarrow M_4 : \geq t_4$
- $M_1 : < t_1 \rightarrow M_2 : \geq t_2 \rightarrow M_5 : < t_5$
- $M_1 : \geq t_1 \rightarrow M_3 : < t_3 \rightarrow M_5 : < t_5$.

The probability of each of these paths was calculated for each value of test-taker true ability θ and summed up. Then the overall proportion of "low" may be calculated by integrating over all θ. To exemplify, let us assume at test taker with ability $\theta = -1$ takes the MST, where all thresholds have been set to 4, except $t_4 = 5$. The probability for each of the paths above may then be calculated using basic probability calculus derived from the IRT parameters, yielding 0.180, 0.395, and 0.152. Hence the total probability that a $\theta = -1$ student ends up in the low group is 0.727. The full probability model for the four groups, i.e., the probability of being classified into each of the four groups as a function of the skill level θ is shown in Figure 9.

The proportion of students ending up in the low group is then obtained by integrating overall test-taker θ values. If this is done, for the given threshold values, the calculations yield that the risk,

low, medium, and high proportions are 8.7, 14, 26.2, and 50.9%, respectively. Given that we want to screen for the lowest 20% percentile, we decided that ending up in the risk or low (8.7 + 14%) groups could indicate screening status.

It is important to note that these theoretical calculations are only approximately correct when benchmarked against a future real-world test administration if (A) The IRT-model is approximately correct, e.g., latent trait normality, unidimensionality and local independence must hold; and (B) the population of future test-takers is identical to the population which was sampled during the initial pilot study. Importantly, this latter condition may not hold if test administration for the pilot occurred at a different time of the academic year compared to the time interval in the final test administration.

From an American context Guthrie (2004) reports that at least 10–15% of students have not established the basic oral reading fluency needed to read texts that are common in the beginning of 3rd grade. This number confirms that a cut off at 20% is reasonable for including students that need extra follow-up in order to develop sufficient reading skills.

4.2.2 Piloting the proposed word reading MST

The proposed MST design was piloted in November 2021, with $N = 406$ test takers. As shown in Figure 10, there was a large discrepancy between the perceived difficulty levels of items at the two test occasions. This is a somewhat surprising finding that may partially be explained by the fact that the item calibration pilot was conducted at an earlier stage (August/September) than the MST (November/December). Hence, students were generally not as proficient in reading when taking the pilot, compared to the students who took the MST three months later. It was uniformly the case that during the MST pilot, the items were easier than during the original IRT pilot. For instance, the first item in the start module was scored correctly by 75% of the test takers in the original IRT pilot, while this item was scored correctly by 87% of the test takers during MST piloting. The discrepancy in item difficulty across the two test occasions resulted in fewer test takers scoring below the thresholds than being expected. Very few students were routed to the the two more easy modules in the last stage, see Figure 11.

Given that the MST pilot results indicated that the test was easier than we initially expected when designing the MST, we decided to shift the final cutoff for screening. That is, we decided to put more trust in the empirical MST pilot results than in our initial theoretical calculations underlying the MST construction. Therefore, we decided to flag test-takers as at-risk if they were not routed to the hardest module at stage 3, or if they were routed to the hardest module, but answered correctly on less than three of the six items in this module. This cutoff was chosen so that ∼20% of the test-takers were deemed at-risk in the MST pilot.

TABLE 2 Mean value for item difficulty and discrimination for modules in word reading MST.

Module	1	2	3	4	5	6
Difficulty	−0.90	−1.31	−0.63	−1.71	−0.90	−0.21
Discrimination	3.07	2.20	3.23	2.05	2.11	2.32

4.2.3 Deployment of word reading MST as a national screening test

The word reading MST was deployed to $n = 49,828$ test takers. The results are illustrated in Figure 12. We see that in this large-sample deployment the test flow is largely as expected from the pilot MST, and that the chosen cutoff for at-risk status performed well. In the figure we also depict the percentages of test-takers for each possible number of correct answers in the final module by horizontal bars. We see, e.g., that almost 31% of test takers were were routed to the hardest module in the last stage and made no errors in that module. The cut-off of correctly answering at least three items in this hardest module resulted in 19.1% of test takers being flagged as at-risk (100−30.6−24.6−16.3−9.4 = 19.1).

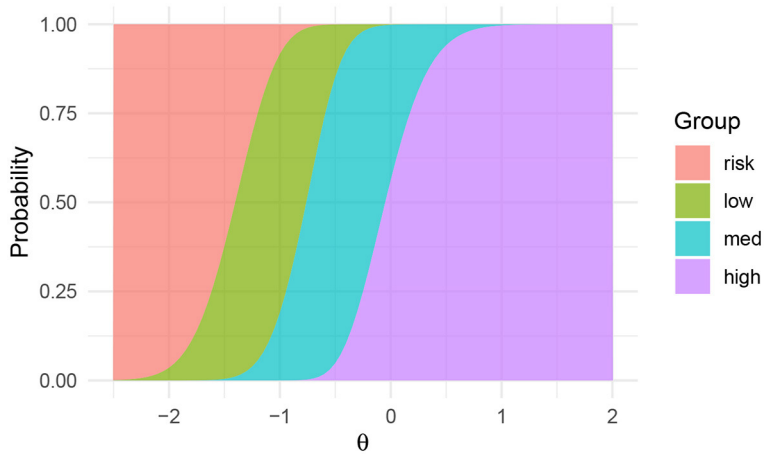


FIGURE 9
Word reading. Probability of being associated with group, as a function of skill level θ.

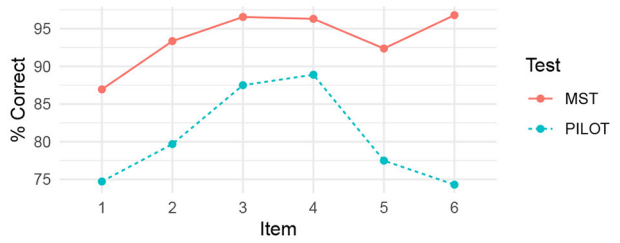


FIGURE 10
Percentage correct for the six items in start module for Word reading MST. Pilot refers to the original item calibration results. MST refers to the MST pilot.

5 Report formats, teachers' guidance material and students' experiences with undertaking the test

5.1 Reporting of results to teachers

How students' performance is reported to teachers plays a crucial role for the interpretation and use of the test results—and is consequently a cornerstone for the test's overall validity (Kane, 2013). After the students have completed the test, teachers get immediate feedback concerning which students' performance are considered to indicate that the student is at risk of reading difficulties (see Figure 13). Norwegian Directorate for Education and training have decided that results exclusively are displayed for students categorized to be at risk and in need of follow up. From the ministry's perspective this is done to align the presentation of results with the purpose of the test, namely aiding teachers in identifying students who lag behind in reading.

The digital display gives the teacher descriptive information regarding number of correct and incorrect responses relative to the four subtests (vocabulary, spelling, word reading, and reading comprehension). Only results for the students who are found to be at risk are displayed. In addition, detailed information of the performance of the at-risk students is only given for the subtests in which the students scored within the follow-up area. For the sake of argument-based validity (Kane, 2013), the presentation of test scores and follow-up material in the screening test aims to provide an easily accessible—and even intuitive—description of how teachers go from score to intervention.

5.2 Guidance material

The follow-up material for the current reading screening tests follow principles from the research-based program for struggling readers CORI (Concept-Oriented Reading Instruction) (Guthrie, 2004). The central aim of this material is to guide the teachers in how to tutor the students through their obstacles in reading while enabling them to interpret meaningful texts. In these follow-up sessions, efforts to trigger and maintain students' motivation are crucial both when it comes to students' self-efficacy and reading interest. This practice is in accordance with recent research indicating that these motivational factors develop reciprocally in concert with reading skill (Chapman et al., 2000; Morgan and Fuchs, 2007; Walgermo et al., 2018; Toste et al., 2020).

A key point in all reading interventions for students who struggle with reading or are reluctant readers, is to provide the students with engaging texts that trigger their situational interest. Many such triggers of situational interest will in time develop into more stable individual interest for reading (Renninger and Hidi, 2015). And students who are interested in and have a positive attitude toward reading tend to become more skilled readers (Petscher, 2010). In the current follow-up material acknowledged authors and illustrators have developed new high quality texts that target interest and level of difficulty for struggling 8-year-old readers. Each text is available in three levels of difficulty in order to give the whole range of struggling readers a sense of mastery when working with the texts. Figure 14 displays two of the texts that are included in the present guidance material.

Given that this screening test—for the reasons stated above—does not include a subtest of letter knowledge, teachers are urged to map the letter knowledge of the students who score within the follow-up area on the subtests of Word reading and Spelling. For this the teachers are provided with a specific letter knowledge test that is individually administered on paper and maps the students' knowledge of three dimensions of letter knowledge; letter writing, letter recognition and letter recall. With this additional information on the struggling students' level of letter knowledge, the teacher can give attention to the students' application of difficult letters when reading the texts described above.

5.3 Teachers' and students' experiences

Importantly digital tests and systems build to evaluate children should continuously and thoroughly be evaluated by their users (Markopoulos and Bekker, 2003; Lehnert et al., 2022). Thus, data on teachers' and students' experiences with the test were collected through extensive interviews with teachers in three different parts of the country, as well as self-reported data concerning students' self-efficacy and interest for different tasks and texts within the test situation. Additionally, student data was gathered when the test was piloted, while teacher interviews were gathered both during pilots and after the first mandatory national deployment of the test. Researchers were also present in classroom as observers both under the two pilot stages, as well as during the deployment of the test in November 2022.

Students generally report that they enjoy working with the tasks and that they feel competent during the test. A separate and more specific study is conducted related to the dynamics between students' interest, self-efficacy and skill when carrying out sub-tests within the screening test (see Walgermo et al., in review).

A point for improvement reported by the teachers when it comes to facilitating the interpretation and use of test scores, relates to their need for information regarding students who are close to, but not in, the follow-up area (within the ten lowest percent above the cut-off). The teachers also report that students' actual answers could be made even more easily accessible within the digital feedback interface. An additional point of improvement

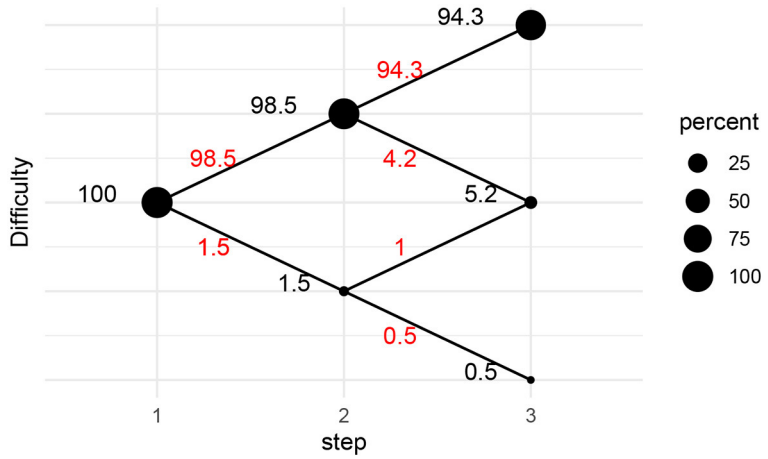


FIGURE 11
Flow percentages for the MST pilot. Percentages in red are percentages of test takers being routed between modules. Percentages in 1black refer to the percentage of test takers for a given module.

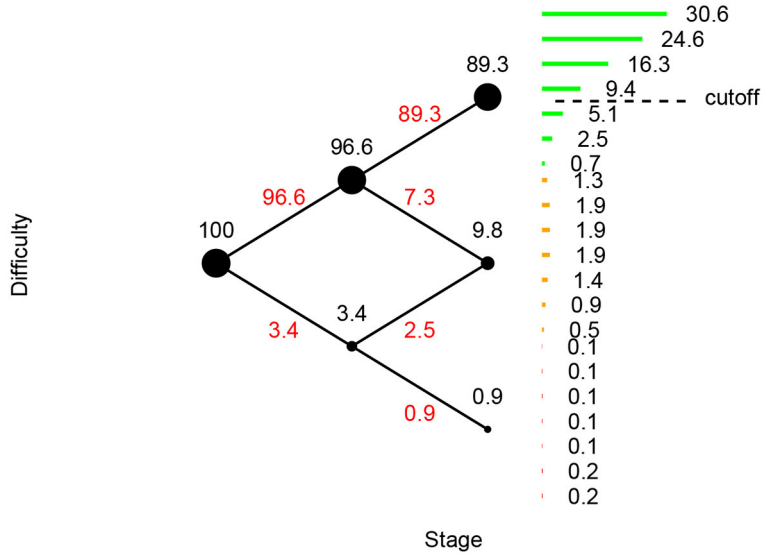


FIGURE 12
Flow percentages for the MST deployment. Percentages in red are percentages of test takers being routed between modules. Horizontal bars on the right-hand side indicate percentages of test-takers for each number of correct items.

relates to to the test length and the need for a test as short a possible, in particular for the lowest performing readers.

6 Discussion

The overall aim of this paper is to document the research-guided development of a new adaptive multistage screening test for reading difficulties for young students. To our knowledge, very few such detailed descriptions are openly available to, e.g., parents, teachers, policy makers, researchers, and test developers. Tests are typically still treated as black boxes. Above, we have presented the rationale of the test, content, choices made and methods applied, while also giving a picture of what the students actually face when taking the test. What has not been shared, is the full overview of the actual items with their psychometric properties. These are considered test specific information that should not be copied onto other assessments. Consequently, this choice aligns with central open science values (Burgelman et al., 2019), stating that scientific data and methods should be shared "as open as possible, and as closed as necessary". In this case, the purpose is to develop a screening that identifies those students who need extra help and support in their reading development. With this aim in mind, we will in the next sections discuss to what extent all means available

FIGURE 13
The figure shows how the teachers are presented with the scores of the students found to be at risk of Reading Difficulties (RD), and are in need of extra follow-up in order to develop adequate reading skills. In accordance with the aim of the screening test, no information is given concerning the performance of students who on the basis of their performance not are found to be at risk of reading difficulties.

has been applied in order "to gear a new test concept to the original purpose of the test" (Walgermo et al., 2021, p. 8). When evaluating the functioning and validity of a test its purpose and use must be the primary point of departure (Kane, 2013, 2015).

6.1 Adaptation model: limitations and possibilities

As mentioned, the choice of the present multistage (MST) adaptation model was made by the Directorate for Education and Training, beyond the influence of the research team. Some challenges arise from this choice of adaptive model: (A) *Test length*. The duration of the test was on average 20 min excluded instructions, meaning that adaptivity's potential for reducing test length seems to be realized by this form of MST. As pointed out by Walgermo et al. (2021), test length represented a challenge for the old tests "The test may be particularly taxing on those students who struggle the most—60 min is a long time to spend working on something that you do not really feel that you have mastered" (p. 6). The adaptive solution in the present screening test is routing the students based on their sum scores within each module. While this model gives good precision concerning at-risk students, the same precision could have been achieved by computer adaptive testing at item level (CAT) (Wainer et al., 2000), with live IRT estimations based on students' performance. Such a CAT model would reduce the number of items required and hence the test time substantially, at no cost in precision (Van der Linden and Glas, 2000). (B) *Uneven test termination time*. The adaptive model also gave an additional challenge compared to the old tests by the fact that the MST made students finish at different times. When all the modules are completed the student has completed the test. In the old tests, students were guided through the material, item by item, by the human test administrator. The new situation therefore calls for a pedagogical action in order to make the transfer from the test over to new tasks for each individual student at different time points during the test. (C) *Limited use of available item qualities*. The present MST model does not exploit all the information available in every item. First, the MST model itself implies that students take sequences of items before their performance is estimated. This might be unfortunate for the target students—the lowest performing readers—as their test taking effort is vulnerable, and a more effective skill estimation therefore is beneficial. Second, the prescribed technical platform was not able to use IRT values for the items in each module for estimating skill level. As a consequence, only the number of correct responses was used to route students to the next module. IRT values were used to create the modules—in which the items had similar difficulty—but not to routing between modules.

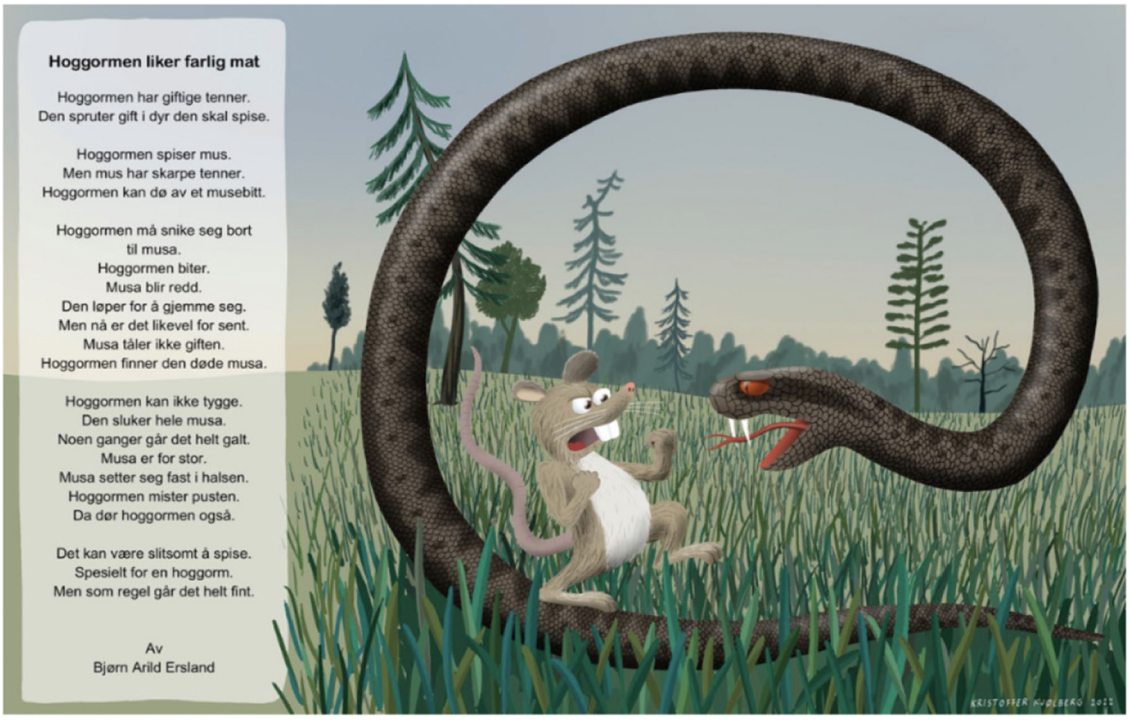
6.2 Cut-off and report format: limitations and possibilities

A cut-off close to 20% corresponds well to teachers' estimation (Berk, 1976; Livingston, 1995) of how many students are lagging behind in reading in their classes. Following the so-called *contrasting group method* (Berk, 1976) Walgermo et al. (in preparation) document that teachers with certainty identify 14,3% of the students to lag behind, and 28% when asked to add those students whose reading risk status they consider uncertain. Still, the cut-off faces challenges when it comes to prediction value. The chosen cut-off is considered sufficient and convenient for the screening purpose, i.e., to secure that students receive necessary instructional support, but is not sufficient for a diagnostic purpose. The cut-off question is also central to the question of how to construct a comprehensive report format. A clear cut-off augments the risk of false positives/false negatives, and unduly boosts the authority of the screening test. An actionable report format is therefore crucial to the interpretation and use of the test, i.e., its validity (Kane, 2013). As seen in Figure 12 the report format of this version of the test faces some black-box challenges. It lacks the opportunity for the teacher to see items in the test and how the student scored on individual items. In the evaluation of the Danish tests, this feature became a stumbling stone for teachers: they missed the opportunity of seeing the student's performance throughout the test. In the current test, a functional report format is likely to succeed, as the subtests are understandable, they concern reading words and text, and writing words, i.e., there are no subtests that provide knowledge unfamiliar to teachers.

A future development of the current screening of reading difficulties could include a follow-up area within the test that was

10.3389/feduc.2024.1331777

A



B

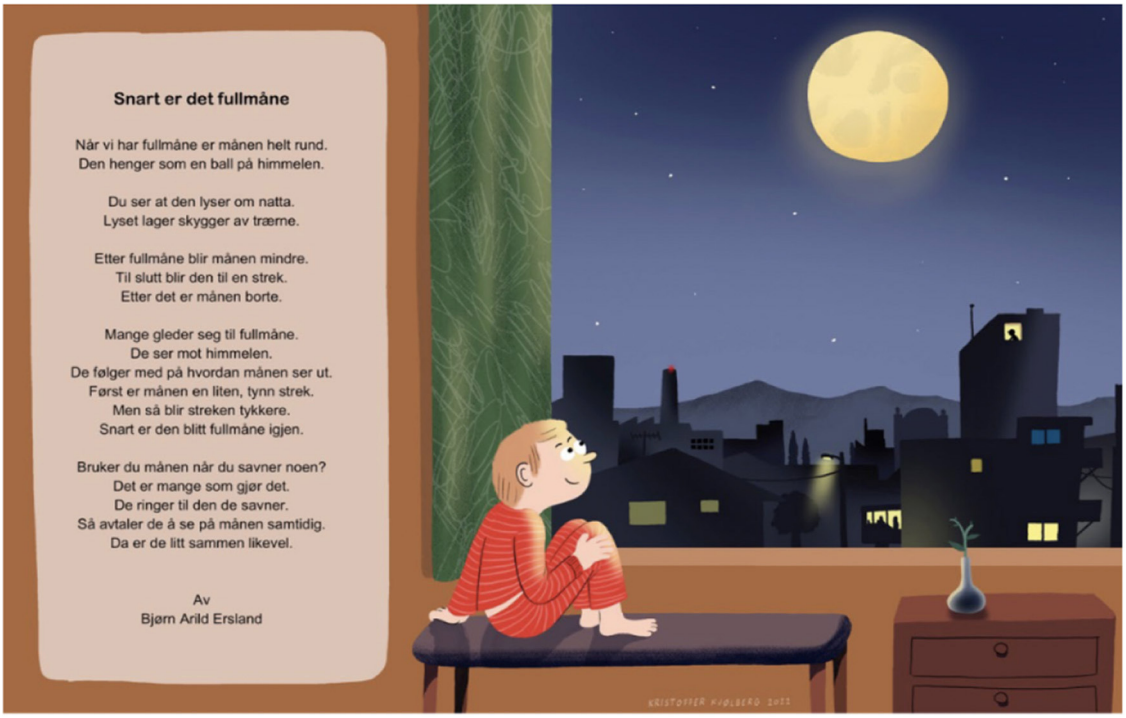


FIGURE 14
The figure shows two high quality texts developed by acclaimed illustrators and authors specifically for the guidance material. **(A)** Displaying a non-fiction test while **(B)** displays a fiction text. Each text is available to the teachers in three difficulty levels to ensure that meaningful texts with the right level of complexity will be accessible for every student.

set in order to be sure to include all true positives, accompanied with an easy manageable procedure for routing false positives out of the follow-up program.

The threshold for the different sub-test is yearly monitored and adjusted when needed by the test developers in collaboration with the Directorate for Education and Training.

6.3 The use of test results: limitations and possibilities

The current screening test is mandatory for all Norwegian students at the onset of 3rd grade. A recurrent issue in the history of Norwegian screening tests in reading is teachers' wish for having a test that goes beyond the purpose of the screening, namely a test that is normally distributed, providing detailed information on students at every skill level. Underpinning these teachers' wishes is a logic that says that a test that is taken by all students, should also give information on the same students. Despite this general opinion of teachers, however, the test is designed to only give information on the target students, by showing clear ceiling effects for the non-target group of students. Unfortunately, however, screening scores have inadvertently been widely misused by teachers to be interpreted as good information on all students. To counteract this misuse of the test, the report format of the current test only gives information on the students that are identified (see Figure 12). As such, this element of the current test represents a convenient regulation of misuse, while the fulfilment of the teachers' wishes could have been within reach, if the test purpose had been adapted.

7 Concluding remarks

The present paper has shed light upon the process of developing a new adaptive screening test in reading. Our aim of full transparency includes sharing with readers what we consider to be the strengths of the test, as well as its limitations and areas for improvement—which will inform future test development. In short, the present adaptive test has several strengths in that it is grounded in new perspectives on reading and writing development. Also, the adaptive design is solid and runs well. The most pressing area for improvement is comparing this MST design to a full adaptive design (CAT). A full adaptive design may improve the test when it comes to students' experience because it could reduce the test length substantially, while keeping its level of accuracy in identifying students at risk of reading difficulties. Pertinent empirical questions in this is the potential for better test taker experience for struggling readers, plausibly driven by shorter test time and a better and faster adaptation to the individual students' skill level.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The Norwegian Ministry of Education has not allowed the datasets for public distribution yet. Requests to access these datasets should be directed to njal.foldnes@uis.no.

Author contributions

BW: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. NF: Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. PU: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. AB: Conceptualization, Validation, Writing – original draft, Writing – review & editing. KL: Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The development of the present adaptive reading screening test was funded by the Norwegian Directorate for Education and Training and the Norwegian National Reading Center.

Acknowledgments

Thank you to Kirsti Thisland for steady administrative work with the new reading screening test throughout the development phase. Also, thanks to Anne Marta Vinsrygg Vadstein and Liv Kristin Børlykke Øverneng for representing the Nynorsk perspective and being responsible for all nynorsk translations.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adams, M. J. (1994). *Beginning to Read: Thinking and Learning About Print*. Cambridge, MA: MIT Press, 433–473.

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Alexander, P. A. (2005). The path to competence: a lifespan developmental perspective on reading. *J. Liter. Res.* 37, 413–436. doi: 10.1207/s15548430jlr3704_1

Alves-Wold, A., Walgermo, B. R., and Foldnes, N. (2024). Assessing writing and spelling interest and self-beliefs: does the type of pictorial support affect first and third graders' responses? *Assess. Writing* 60:100833. doi: 10.1016/j.asw.2024.100833

Alves-Wold, A., Walgermo, B. R., McTigue, E., and Uppstad, P. H. (2023). Assessing writing motivation: a systematic review of k-5 students' self-reports. *Educ. Psychol. Rev.* 35:24. doi: 10.1007/s10648-023-09732-6

Andersen, S. C., Bleses, D., Damm, A. P., Gensowski, M., Gørtz, M., Gregersen, M. K., et al. (2019). *31 forskere: Drop kritikken af de nationale tests*. Hentet Juni 2020 fra Politiken. Available online at: https://politiken.dk/debat/kroniken/art7168552/Drop-kritikken-af-de-nationale-tests

Arnesen, A., Braeken, J., Ogden, T., and Melby-Lervåg, M. (2019). Assessing children's social functioning and reading proficiency: a systematic review of the quality of educational assessment instruments used in norwegian elementary schools. *Scand. J. Educ. Res*. 63, 465–490. doi: 10.1080/00313831.2017.1420685

Bakken, A. M., Gourvennec, A. F., Walgermo, B. R., Solheim, O. J., Foldnes, N., and Uppstad, P. H. (2023). Adaptvurder: study protocol for an upcoming adaptive reading test. *Nordic J. Liter. Res*. 9, 59–72. doi: 10.23865/njlr.v9.2906

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. W. H. Freeman and Company. New York, NY: W. H. Freeman.

Berk, R. A. (1976). Determination of optional cutting scores in criterion-referenced measurement. *J. Exp. Educ*. 45, 4–9. doi: 10.1080/00220973.1976.11011567

Brożek, B., Furman, M., Jakubiec, M., and Kucharzyk, B. (2023). The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artif. Intell. Law* 32, 1–14. doi: 10.1007/s10506-023-09356-9

Bundsgaard, J., and Puck, M. R. (2016). N*ationale test: danske lærere og skolelederes brug, holdninger og viden*. DPU, Aarhus Universitet.

Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., et al. (2019). Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Front. Big Data* 2:43. doi: 10.3389/fdata.2019.00043

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw*. 48, 1–29. doi: 10.18637/jss.v048.i06

Chapman, J. W., Tunmer, W. E., and Prochnow, J. E. (2000). Early reading-related skills and performance, reading self-concept, and the development of academic self-concept: a longitudinal study. *J. Educ. Psychol*. 92:703. doi: 10.1037/0022-0663.92.4.703

De Ayala, R. J. (2022). *The Theory and Practice of Item Response Theory*. New York, NY: Guilford Publications.

Ehri, L. C. (2015). *How Children Learn to Read Words, 293-310. Oxford Library of Psychology*. New York, NY: Oxford University Press.

Engen, L. (1999). *Kartlegging av leseferdighet på småskoletrinnet og vurdering av faktorer som kan være av betydning for optimal leseutvikling: En beskrivelse av den faglige prosessen med å utvikle nasjonale kartleggingsprøver for småskoletrinnet, og en vurdering av forholdet mellom fonologiske delferdigheter, ordlesings- og tekstlesingsferdigheter blant elever i 1. og 2. klasse*. Bergen: Institutt for samfunnspsykologi, Psykologisk fakultet, Universitetet i Bergen.

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., et al. (2013). Assessing the quality of tests: REVISION of the efpa review model. *Psicothema* 25, 283–291. doi: 10.7334/psicothema2013.97

Farrow, R. (2023). The possibilities and limits of XAI in education: a socio-technical perspective. *Learn. Med. Technol*. 48, 266–279. doi: 10.1080/17439884.2023.2185630

Flarup, L. (2020). *Evalueringen af de Nationale Test. tværgående evalueringsrapport*. VIVE. København: VIVE - Det Nationale Forsknings- og Analysecenter for Velfærd.

Fuchs, L. S., Fuchs, D., Hosp, M. K., and Jenkins, J. R. (2001). "Oral reading fluency as an indicator of reading competence: a theoretical, empirical, and historical analysis," in *The Role of Fluency in Reading Competence, Assessment, and Instruction* (New York, NY: Routledge), 239–256.

Graham, S., Aitken, A. A., Hebert, M., Camping, A., Santangelo, T., Harris, K. R., et al. (2021). Do children with reading difficulties experience writing difficulties? A meta-analysis. *J. Educ. Psychol*. 113:1481. doi: 10.1037/edu0000643

Guthrie, J. T. (2004). "Differentiating instruction for struggling readers within the cori classroom," in *Motivating Reading Comprehension* (New York, NY: Routledge), 173–193.

Hollender, N., Hofmann, C., Deneke, M., and Schmitz, B. (2010). Integrating cognitive load theory and concepts of human-computer interaction. *Comput. Hum. Behav*. 26, 1278–1288. doi: 10.1016/j.chb.2010.05.031

Jiao, H., He, Q., and Yao, L. (2023). Machine learning and deep learning in assessment. *Psychol. Test. Assess. Model*. 64, 178–189.

Kane, M. (2013). The argument-based approach to validation. *Sch. Psych. Rev*. 42, 448–457. doi: 10.1080/02796015.2013.12087465

Kane, M. (2015). "Validation strategies: delineating and validating proposed interpretations and uses of test scores," in *Handbook of Test Development*, eds. S. Lane, M. Raymond, and T. Haladyna (Routledge), 80–96.

Lehnert, F. K., Niess, J., Lallemand, C., Markopoulos, P., Fischbach, A., and Koenig, V. (2022). Child-computer interaction: From a systematic review towards an integrated understanding of interaction design methods for children. *Int. J. Child Comp. Interact*. 32:100398. doi: 10.1016/j.ijcci.2021.100398

Livingston, S. A. (1995). "Standards for reporting the educational achievement of groups," in *Proceedings of the Joint Committee on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES)*, Vol. 2 (Washington, DC: National Assessment Governing Board and National Center for Education Statistics), 39–51.

Madaus, G. F., and Keillor, G. (1988). The influence of testing on the curriculum. *Teach. Coll. Rec*. 89, 83–121. doi: 10.1177/016146818808900505

Markopoulos, P., and Bekker, M. (2003). On the assessment of usability testing methods for children. *Interact. Comput*. 15, 227–243. doi: 10.1016/S0953-5438(03)00009-2

Messick, S. (1989). "Validity," in *Educational Measurement, 3rd Edn*., ed. R. L. Linn (New York, NY: Macmillan Publishing Co, Inc; American Council on Education), 13–104.

Messick, S. (1996). Validity and washback in language testing. *Lang. Test*. 13, 241–256. doi: 10.1177/026553229601300302

Ministry of Church and Research (2000). *Second and Seventh Grade Reading Assessment*. (*f-037-00)* [*Circular*]. Oslo: Government Document.

Morgan, P. L., and Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Except. Child*. 73, 165–183. doi: 10.1177/001440290707300203

Nagy, W., Scott, J., and Kamil, M. (2000). "Vocabulary processes," in *Handbook of Reading Research*, Vol. 3, eds. M. L. Kamil, P. B. Mosenthal, P. D. Pearson, and R. Barr (Lawrence Erlbaum Associates), 269–284.

Nielsen, J. C., Kreiner, S., Poulson, A., and Søegård, A. (1986). *SL-håndbog. Sætnings-læseprøverne SL60 & SL40*. Copenhagen: Dansk psykologisk Forlag.

Pearson, P. D., Hiebert, E. H., and Kamil, M. L. (2007). Vocabulary assessment: what we know and what we need to learn. *Read. Res. Q*. 42, 282–296. doi: 10.1598/RRQ.42.2.4

Petscher, Y. (2010). A meta-analysis of the relationship between student attitudes towards reading and achievement in reading. *J. Res. Read*. 33, 335–355. doi: 10.1111/j.1467-9817.2009.01418.x

Quinn, J. M., Wagner, R. K., Petscher, Y., and Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: a latent change score modeling study. *Child Dev*. 86, 159–175. doi: 10.1111/cdev.12292

RCore Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Renninger, K. A., and Hidi, S. (2015). *The Power of Interest for Motivation and Engagement*. New York, NY: Routledge. doi: 10.4324/9781315771045

Rønneberg, V., Jensen, M., Foldnes, N., and Solheim, O. J. (in review). Development of a digital item format for an adaptive word reading test in a semi-shallow orthography.

Seymour, P. H., Aro, M., and Erskine, J. (2003). Foundation literacy acquisition in european orthographies. *Br. J. Psychol*. 94, 143–174. doi: 10.1348/000712603321661859

Solheim, O. J., Frijters, J. C., Lundetræ, K., and Uppstad, P. H. (2018). Effectiveness of an early reading intervention in a semi-transparent orthography: a group randomised control. *Learn. Instruct*. 58, 65–79. doi: 10.1016/j.learninstruc.2018.05.004

Tønnessen, F. E., and Solheim, R. G. (1999). *Kartlegging av leseferdighet og lesevaner på 9. klassetrinn*. Kirke: utdannings-og forskningsdepartementet.

Tønnessen, F. E., and Uppstad, P. H. (2015). *Can We Read Letters? Reflections on Fundamental Issues in Reading and Dyslexia Research*. Rotterdam: Sense Publishers. doi: 10.1007/978-94-6209-956-2

Toste, J. R., Didion, L., Peng, P., Filderman, M. J., and McClelland, A. M. (2020). A meta-analytic review of the relations between motivation and reading achievement for k-12 students. *Rev. Educ. Res*. 90, 420–456. doi: 10.3102/0034654320919352

Van der Linden, W. J., and Glas, C. A. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer Academic Publishers.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000). *Computerized Adaptive Testing: A Primer*. Milton, GA: Routledge.

Walgermo, B. R., Foldnes, N., Bakken, A. M., Gourvennec, A. F., Rangnes, H., and Uppstad, P. H. (in review). Equitable digital vocabulary assessment: what item formats do we need to build a fair vocabulary test?

Walgermo, B. R., Foldnes, N., Uppstad, P. H., and Solheim, O. J. (2018). Developmental dynamics of early reading skill, literacy interest and readers' self-concept within the first year of formal schooling. *Read. Writ*. 31, 1379–1399. doi: 10.1007/s11145-018-9843-8

Walgermo, B. R., and Uppstad, P. H. (2023). "Enhancing students' identities as readers and writers through assessment," in *Becoming Readers and Writers: Literate Identities Across Childhood and Adolescence*, eds. C. J. Wagner, K. K. Frankel, and C. M. Leighton (New York, NY: Routledge).

Walgermo, B. R., Uppstad, P. H., and Lundetræ, K. (in preparation). Den utfordrende, men nødvendige vurderingen av lesing i første klasse.

Walgermo, B. R., Uppstad, P. H., Lundetræ, K., Tønnessen, F. E., and Solheim, O. J. (2021). Screening tests of reading: time for a rethink. *Acta Didactica Norden* 15:8136. doi: 10.5617/adno.8136