Nordic Journal of Literacy Research
Vol. 9 | No. 1 | 2023 | pp. 59–72

Protocol

Adaptvurder: Study Protocol for an Upcoming Adaptive Reading Test

Arild Michel Bakken,* Aslaug Fodstad Gourvennec, Bente Rigmor Walgermo, Oddny Judith Solheim, Njål Foldnes & Per Henning Uppstad
Universitetet i Stavanger, Norge

Abstract
Effective reading instruction requires precise assessment of the learner's current skill level. For young learners, however, assessment often comes at a great cost: Tests take a long time and students are presented with items that are both too easy and too difficult. Recent developments in adaptive testing have the potential for solving both these challenges. In this paper, we take the path of argument-based validity (Kane, 2015) by presenting an *interpretation and use* argument for an upcoming adaptive test. We term this paper a study protocol, in line with the established tradition for protocols for pre-registered empirical trials. The function of the protocol is to communicate openly what often remains tacit knowledge on test development.

Keywords: *reading assessment; adaptive testing; early literacy education; test validity*

Responsible editor: Gustaf B. Skar

Received: January, 2021; Accepted: December, 2022; Published: April, 2023

Introduction

Well-established insights from general pedagogy indicate that a precise assessment of the learner's current proficiency level is key to an effective reading instruction. This is because the support given by the teacher must be adapted to the specific and changing needs of each learner (Black & William, 1998; Vygotsky, 1978), and assessment is required in order to identify these needs.

Daily informal student assessment and local teacher-developed assessment instruments constitute the cornerstone of formative reading assessment. Standardized tests can add important information to assist teacher decisions as these tests often come with high validity and reliability due to thorough development processes. However, standardized reading tests do – in their current state – present some clear drawbacks, one of which is the time required by the test taker in order to achieve a precise estimate of their proficiency. For a young test taker, 90 minutes (for example) is a very

*Correspondence: Arild M. Bakken, e-mail: arild.m.bakken@uis.no

Citation: A. M. Bakken, A. F. Gourvennec, B. R. Walgero, O. J. Solheim, N. Foldnes & P. H. Uppstad. "Adaptvurder: Study protocol for an upcoming adaptive reading test" Nordic Journal of Literacy Research, Vol. 9(1), 2023, pp. 59–72.
http://dx.doi.org/10.23865/njlr.v9.2609

long time to stay concentrated on a test, and any test of such a length will run the risk of creating a less-than-optimal experience for the test taker, which in turn could threaten validity by measuring the test taker's test stamina instead of their reading comprehension. Another significant drawback is that, although they are aimed at providing grounds for an instruction adapted to the specific needs of each learner, linear assessment instruments themselves are not adapted to different proficiency levels. They tend instead to be best suited for learners around the middle of the targeted proficiency scale. This could result in both a less-than-optimal experience for high performing test takers, faced with many items that are far too easy for them, and for low performing test takers, who are asked to solve items that are far too difficult for them. And what is more, the test will also not give optimal information on the proficiency level of these test takers (Magis et al., 2017, pp. 1–2).

A solution to these problems is to take advantage of recent technological advances that facilitate the development of computerized adaptive tests. Such tests adapt progressively to the proficiency level of each test taker, presenting more difficult items to high performing readers and easier items to lower performing readers, such that all students would have, for example, a 60% chance of solving each item. Hence, computerized adaptive tests could make the test experience better for all test takers and produce more precise information – in less time (Magis et al., 2017). A potential drawback, however, is that these tests would actually become more difficult for the strongest readers, giving them a lower success rate than they are used to, potentially decreasing their motivation. On the other hand, it can be argued that an increased sense of challenge could lead to more learning for the strongest readers.

Although adaptive testing is a well-established research field, adaptive reading comprehension tests for young learners have not yet been sufficiently explored. A notable challenge with bringing adaptivity to reading comprehension tests is that each item often depends on a text which it has in common with other items. Achieving item-level adaptivity is therefore difficult. In the last decade, some efforts have been made towards developing such tests, e.g. in Denmark and Wales. However, the Danish tests in particular have been subject to intense debate (see Bundsgaard, 2018) around many aspects of the test: its purpose, its construct, its qualities, and its usefulness. For an overview, see the recommendations report from the advisory group (Ministry of Children and Education, 2020). Certainly, a more thorough documentation on the development process would have paved the way for a more nuanced debate and made it easier for developers of other tests to build on the knowledge gained. Similarly, to the best of our knowledge, no scholarly studies have yet been published on the Welsh tests. The building of knowledge on the development of adaptive reading tests in early literacy education is therefore in its infancy.

We here present the protocol for an adaptive reading comprehension test for 3rd grade school children in Norway. The test concept also includes adaptive subtests assessing important processes underlying reading comprehension, i.e. word reading and vocabulary. The aim of the protocol is to state the proposed interpretation and

use of the test scores derived from the upcoming test concept and to describe how the test concept can be built in order to enable such a use. We target validity in a broad sense, which is comprised of the suitability of the items, the test design, and the use of the tests, in relation to both the constructs being tested and the larger societal reasons for engaging in such testing (Kane, 2015; Messick, 1980; Stobart, 2009).

To this end, we first present the purpose and constructs of the test, and then a way towards its design: item formats, statistical model, adaptive design, and report format. We discuss these design choices in relation to the construct and purpose of the test.

Part I: Proposed Interpretation and Use

Since validity can be seen as a result of the fitness of the test both to the construct being tested (Messick, 1980) and to its intended use (Kane, 2015), we first present the purpose of the test and the construct of reading comprehension underlying the test.

Purpose of the Test

The Adaptvurder project springs from an initiative of the municipality of Oslo, Norway, in its aim to develop formative assessment instruments that are adapted to all students. Since the test is purely formative, the stakes will be low, compared with tests that could be used for accountability purposes. No results will be reported other than to the teacher. The 3rd-graders will take the test in February, which means that most students will be 8 years old and they will have received reading instruction for approximately 2,5 years. At this stage, the reading skills of students present much variation (Arnesen et al., 2017).

The Adaptvurder test will give precise information on the reading comprehension skills of all students, from the highest to the lowest performing readers, enabling the teacher to adapt their reading instruction to the needs of each student. This type of "adapted education" is a long-standing and consensual principle in Norwegian education. The Education Act states that "education must be adapted to the abilities and aptitudes of the individual pupil" (Education Act, 1998, § 1–3). In order to achieve an adapted reading instruction in each student's zone of proximal development (Vygotsky, 1978, p. 86), the teacher must be able to assess the current skills of all students in an adequate manner (Eysink & Schildkamp, 2021), and the Adaptvurder reading test will be an important tool in this effort.

Test Constructs

The main construct of the Adaptvurder reading test is based on reading literacy, as defined in the PIRLS framework:

> Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment. (Mullis & Martin, 2019)

More specifically, the test measures a part of reading literacy, namely the purpose of understanding written language (and less the use of written language). The result of this understanding / meaning construction is reading comprehension.

According to an influential conceptualization of reading comprehension, it is a product of two factors, decoding and language comprehension (Gough & Tunmer, 1986). These factors are important prerequisites for reading comprehension and they are the two most important predictors for understanding variance in reading comprehension scores (Lervåg et al., 2018). To provide teachers with information that can help them to interpret individual students' results on the reading comprehension test, Adaptvurder includes separate subtests of word reading and vocabulary, to be reported only in the case that students get low scores on them.

Notwithstanding, as laid out in other publications, we see reading as a unified, interpretive skill, in line with the hermeneutical tradition: "Everything in a text, from the individual letters to its deeper meaning, must be interpreted" (Walgermo et al., 2021, p. 16). A hermeneutical view of reading makes it less relevant to draw a clear line between decoding and comprehension, because "the principle remains the same: seeing the whole and the parts in relation to each other" (Tønnessen & Uppstad, 2015, p. 70).

For the 3rd-graders taking this reading test, their proficiency will reflect a large variation; from the students hardly being able to interpret the very basics of reading, to the students interpreting more complex text with ease. As such, progress in reading for this age-group is also reflected in their capability in moving from reading words, over sentences to texts with increasing complexity. The formats chosen and described below incorporate this insight.

Part II: Description of Choices made to Serve Intended Interpretation, Purpose, and Use

Operationalization of the Constructs

Constructs are in principle unobservable. They are hypothetical, latent variables that are only accessible indirectly, through manifest variables, such as responses to test items (Wilson & Gochyyev, 2013, p. 4). Bringing the test construct into operation, therefore, implies developing the types of items that are best suited to giving access to the latent reading comprehension, by attracting the kind of responses that reveal it.

In this section, we first discuss the item formats that are fit to operationalize reading comprehension. Then, we discuss item formats in word reading and vocabulary.

Operationalization of Reading Comprehension

For reading comprehension, stimuli in the form of texts, words, and sentences are of critical importance, even before the items themselves. For the highest skill levels, the stimuli are texts. As far as possible, we use authentic texts, sometimes with minor changes in order to fit the purpose. The texts are of different length and complexity in order to assess the whole span of skill level. We aim for a broad variation in genre,

text type, and themes, as well as in the disciplinary affiliation of the texts. Although by "written language" we primarily mean alphabetic writing, the texts we have chosen often include non-alphabetic elements such as illustrations, maps or graphs, and the interplay between them and the alphabetic writing is sometimes of interest in the items.

In the PIRLS framework, four processes of comprehension are suggested: (1) focus on and retrieve explicitly stated information, (2) make straightforward inferences, (3) interpret and integrate ideas and information, and (4) evaluate and critique content and textual elements. These processes are useful and frequently used for developing items, and a version is also reflected in the Norwegian Framework for Basic Skills (Directorate for Education and Training, 2017). Targeting these processes, we develop items that are associated with the texts in the test. To each text belongs one or more multiple-choice questions, and these questions are meant to induce the student into engaging in one of the four processes. The distractors represent plausible misrepresentations of the meaning of the text.

In addition to these multiple-choice items, we also use items that are statements about issues present in the text, which the students must decide to be true or false. These are used when it is not possible to make enough plausible multiple-choice items for each text. The true statements concern elements in the text that the students might not perceive, whereas the false statements represent plausible misrepresentations of the meaning of the text. Here too, the students are induced to engage in processes, which real readers would engage in within an authentic reading situation: evaluate whether a preliminary construction of the meaning of the text is correct. Sometimes, this will lead the student to search for information in the text, and at other times the student must make inferences about the text.

To access the lowest levels of reading comprehension, we use a format where the stimulus is a sentence, and the multiple-choice options are images. The sentence describes a situation, and the correct answer is an image representing this situation. The distractors represent plausible misrepresentations of the meaning of the sentence or of words included in the sentence. This format only requires the reader to read the sentence in the stimulus. It could function quite differently than the rest of the reading comprehension items, and any differential functioning will be investigated.

Operationalization of Word Reading

Word reading involves being able to recognize written words. Previous studies have found that most Norwegian students read accurately by the end of Grade 1 (Seymour et al., 2003). However, there are large variations in the degree of fluency (Arnesen et al., 2017). Consequently, the Adaptvurder test will have a subtest that assesses the ability to recognize words quickly and accurately. The item type is a multiple-choice item with single words as stimuli and options. Initially, a stimulus word is presented on the screen. The stimulus word then disappears and is replaced by four words: the stimulus word (key) and three orthographically similar words (distractors). The task is to recognize the stimulus word among the four options. There is no time limit on

providing the answer. The level of difficulty increases along two dimensions: time (duration of the stimulus) and word length.[1] The duration of the stimulus is 2 seconds, 1 second, 500 milliseconds, or 200 milliseconds. Word length differs between 3 and 10 letters. The easiest items present a stimulus word of 3 letters for 2 seconds. The most difficult items present a 10-letter word for 200 milliseconds. Even if this item format is developed to assess word reading, language comprehension is not completely absent, because understanding the meaning of a word will help the students in recognizing the word more quickly (Nation & Cocksey, 2009; Ricketts et al., 2007). These items will be calibrated separately.

Operationalisation of Vocabulary

Measuring language comprehension is central to the test's purpose, because it will allow the test to identify readers whose main problem with reading comprehension is insufficient oral language comprehension of Norwegian (e.g. due to being a second language learner or having developmental language disorder).

For measuring language comprehension, the test will have a subtest using various formats that access the student's vocabulary. Vocabulary is, of course, a narrowed-down part of language comprehension skill, and cannot represent the whole skill. However, it is the easiest part of language comprehension to test (and the most frequently tested) and will be considered for the purposes of this test to be a sufficient spot-check of the skill.

To assess the most basic levels of language comprehension, we use a format where the students see a rich thematic picture with many objects. The students hear a word and they are asked to click on the corresponding zone of the image. Next, we measure vocabulary in context: The student hears a sentence and is asked to choose a synonym for one of the words from a list of four options. Finally, a more demanding task is a format that measures vocabulary in isolation: The student hears a word and is asked to choose a synonym from a list of four options. The difficulty in vocabulary thus increases along an axis going from naming to finding synonyms, from concrete words, mostly nouns, to more abstract words, including also verbs and adjectives. In all these formats sound support is essential in order to prevent contamination from the decoding skill. All language comprehension items will be calibrated separately.

Assembling the Item Pool

Building an adaptive test requires a large item pool because all students should receive a sufficient number of items that are adapted to their skill level. The adaptive test also requires a careful statistical examination of the items in order to know

[1] Item difficulty will also be affected by word frequency, but this aspect is not systematically incorporated in item development.

how they would function in the test. For example, for the test to be able to provide students with items that are adapted to their current skill level, it is crucial to know beforehand how difficult the items are.

We base the statistical examination and validation of items on item response theory (IRT). Among the most popular IRT models is the two-parameter logistic (2PL) model, where each item is assigned a difficulty and a discrimination parameter. The 2PL model is required in the national tests in Norway (Directorate for Education and Training, 2017b). A simpler and more robust model is the one-parameter Rasch model. However, we find that this model is too constrained for our purposes. Although it provides a difficulty parameter, no information on discriminatory power is incorporated in the Rasch model. We opted to also model variation in discrimination among items, using the two-parameter model. Among two equally difficult items, we can therefore assess their discrimination in order to choose among them. The item with highest discrimination will be more informative, given equal difficulty levels. An adaptive algorithm will profit by choosing a high-discrimination item and thereby achieving a more precise skill estimation. IRT models with more than three parameters require larger sample sizes. Within our resource constraints, this would imply using a smaller set of items. Overall, we therefore decided that using the two-parameter model would reasonably balance the need for item parameter precision and a sizeable item pool. In addition, such a choice would allow us to align with the national tests. In the present project we estimate IRT models using the R (R Core Team, 2019) package mirt (Chalmers, 2012).

We will conduct two pilot studies in order to evaluate the psychometric adequacy of the proposed items and to calibrate the items for use in an adaptive model. The sampling frame for the pilot studies will be 3rd-graders in big city Norwegian schools, the primary intended user group of the final test. To allow us to calibrate many items, the pilot studies will be organized in different booklets. Some items in each booklet will be present also in other booklets in order to create a link.

Items found to be low on discrimination in the first pilot, that is, that do not adequately discriminate levels of low proficiency from levels of high proficiency, will be deemed inadequate. We will use as cut-off a traditional heuristic of $a = 0{,}85$ with the logistic metric used by mirt (corresponding to $a = 0{,}5$ with a normal ogive metric). Items with a values below this cut-off will be scrutinized by an expert panel, and either modified for entry in the second pilot study or dismissed altogether from Adaptvurder. After the second pilot study, items will again be either dismissed or entered into the final item pool.

For determining sample sizes for these two pilot studies, we will consider the recommendation of the European Federation of Psychologists' Association (EFPA), stipulating that an adequate sample size for each item under the 2PL IRT model is 400 (Evers et al., 2013). In addition, we will set up Monte Carlo IRT sample size simulations with the R package SimDesign (Chalmers, 2017), using the setup described by Mair (2018, pp. 126–130).

Since the test concept includes a core test and two subtests of different constructs, three item banks will be calibrated apart: reading comprehension items, word reading items, and vocabulary items. In addition, the dimensionality of each of these item banks will be assessed. Particularly in the reading comprehension and vocabulary item banks, there are several item formats that could represent different constructs.

Choosing an Adaptive Design

After having obtained an item pool with adequate content balance and psychometric validity, the next task is to decide upon the specific design of the Adaptvurder test. This is a complex task that involves striking a balance between many considerations, such as requirements for short test-taking time, test precision, content balancing, test experience, and reporting.

The test concept that we plan to design and implement will have adaptive features. This means that at certain stages, the test assesses the performance of the test taker and estimates their proficiency. This estimate is used to dynamically choose the next test element for the user to solve.

Within this framework many adaptive designs are available. Broadly, we may distinguish between fully adaptive tests, here referred to as computerized adaptive tests (CATs), and so-called multi-stage tests (MSTs) that are adaptive only in the transition from one stage to the next stage (e.g. Magis et al., 2017, pp. 2–3). In a CAT, the basic test element is the item. Items are delivered dynamically on a one-to-one basis, so that the proficiency estimate is updated after each item. In the MST the basic test element is a module, by which we mean a fixed linear test that is content-specific. For instance, a reading comprehension test consisting of three texts associated with ten fixed items would constitute a module. In the MST, proficiency is only updated after a module is completed. The MST may, therefore, be considered as an intermediate between a standard linear test and a fully adaptive CAT.

A major decision that we need to make is whether to adopt CAT or MST for Adaptvurder. The latter is conceptually simpler, and might be easier to implement, at the expense of either measurement precision or test length. The MST might also protect against a possible threat to construct validity in an adaptive test, in that it gives a greater control of each student's path through the test (Yan et al., 2014, p. 6). In a CAT, some aspects of the construct could be underrepresented as a result, for example, of the algorithm's privileging item information over content balancing. It is also a possibility to choose one kind of design for the constructs of word reading and vocabulary and another design for the construct of reading comprehension (e.g. the CAT for the two former and the MST for the latter).

Additional considerations involve the order of content presentation, e.g. whether language comprehension should be assessed first, followed by word reading, before moving to the reading comprehension assessment. Decisions must also be made on a more technical level regarding which item selection method should be used (e.g. the

maximum Fisher information criterion or the Kullback-Leibler divergency criterion; see Magis et al., 2017, p. 44), which stopping criterion to employ (e.g. time, number of items, precision of estimate obtained), how to control item exposure (e.g. using the randomesque method or the Sympson and Hetter method; see Magis et al., 2017, p. 50), which statistical estimator to use in proficiency estimation, and the associated standard error. All these decisions involve balancing practical considerations with what is technically feasible and statistically advisable. We will carefully analyze these issues in order to arrive at a final design that is technically robust, statistically precise, that exerts a low toll on test takers, and that delivers valid information to teachers and other stakeholders.

Simulation as a tool in assessment design

Certain aspects of a proposed design for Adaptvurder will be calculated through simulation. To choose among different designs it is, therefore, useful to compare simulation output for the designs. This may aid in identifying the best design.

In simulation, the proposed test is exposed to thousands of simulated test takers. This will emulate how the test will perform in the population. Across the entire proficiency scale, we obtain information on average test length, and on how precise the final proficiency estimate will be. Suppose that design A and design B are simulated, and we find that the former involves shorter test length than the latter, but is equally precise in proficiency estimation. Then, all else being equal, design A is preferable to design B. We will use such simulations to choose an adaptive design for the test.

Simulation will also be used, after a final design has been chosen, to fine-tune parameters in the adaptive algorithm. For instance, we can decide on which estimator to use for skill estimation, and which to use for standard error estimation. The stopping criterion may be fine-tuned according to what works best in the simulations, and so forth. By varying algorithm parameters and simulating across the resulting conditions, knowledge is gained on how the final test will operate in terms of many relevant outcome variables. We must then again balance the importance of such outcome variables in order to arrive at specific parameters to be used in the final design for Adaptvurder. Simulations in Adaptvurder for CAT designs will be conducted using the R package catR (Magis & Barrada, 2017), while MST designs will be simulated with the package mstR (Magis et al., 2017).

A third pilot study will be conducted in order to validate the adaptive design of the test. An optimal design will have been selected from the simulations. This optimal design will now be tried out with real items on real students. In this respect, we remark upon a possible threat to the validity of our simulations. Ideally, simulations for adaptive tests should be run on the exact same software platform that will be used when the adaptive design is piloted on real students and later launched on its actual platform for real-world use. Our IRT calibration and our design simulations will be conducted in the widely-used R software environment. One reason for our choice of R as software platform is its transparency as open source software and as part of the

GNU project (Stallman, 1997). The use of R improves reproducibility of analyses and workflows are available through shared scripts. Our plan is to make all our code for IRT calibration and simulation openly available online, allowing for full transparency in our procedures. However, Adaptvurder in its operative implementation will run on commercial software. Therefore, the IRT engine and the adaptive algorithm in Adaptvurder will be programmed in a closed source environment different from the code we used for item calibration and simulations. We plan to inspect the IRT algorithm and CAT implementation closely as our commercial partner develops the engine, in order to make sure that estimators and adaptive algorithms, as implemented in the commercial system, will operate in the same manner as our corresponding R procedures. This will allow our simulations in the design phase to also be representative of Adaptvurder's performance, as implemented on the commercial platform.

Report Format and Guidance Material

The *raison d'être* of formative assessments is that the results from the tests should be used to adapt instruction. Previous research has however identified important challenges when it comes to pedagogical use of test results, both at an international level (Datnow & Hubbard, 2016; Schildkamp & Datnow, 2020; Schildkamp & Kuiper, 2010; Vanlommel et al., 2016) and within the Norwegian context (Gunnulfsen & Roe, 2018; Mausethagen et al., 2018). Some of these are to be found at the administrative level, while others are linked to teachers' test literacy, perceived ability, or motivation to interpret and act on interpretations of the results.

As the key purpose for the development of Adaptvurder is to provide information that supports differentiated instruction for students across all reading skill levels, it is crucial that the test results are communicated in a way that supports pedagogical use of the information.

More precisely, the material should (a) support teachers' and schools' ability to understand and interpret the test results in order to act on them while planning and carrying out adapted reading instruction, (b) motivate teachers and schools to interpret and act on the test results, and (c) support teachers' and schools' ability to enhance knowledge and competence about good reading instruction.

In order to meet these criteria, reports and guiding material will be developed in three phases. Phase 1 involves developing alternative sketches for test reports based on formats from existing adaptive reading tests (the Danish and Welsh tests), getting feedback on the sketches from a reference group of teachers, and on this basis develop two or three alternatives of test report formats. Elements that could enter into such a report are examples of the most difficult items which the students managed to solve (representing their actual development level, in Vygotsky's terms) and the easiest items they did not manage to solve (representing their zone of proximal development). Together with these examples the teacher could see explanations of

what it took or what it would have taken to solve these items. This would give the teacher access to concrete items that would be particularly interesting to look closer at for each student, which in turn can support the teachers' instructional decisions, e.g. in selecting texts well adapted to the student's reading skills, or as a starting point for constructive conversations about the student's reading with the student or parents. Other examples could be reader profiles (what kind of reader is the student) or learner paths (a visual representation of each student's way through the test).

In any case, for the test to be truly formative, the teacher must be able to make sense of the student data in order to identify learning needs (see Eysink & Schildkamp, 2021). It is therefore important that the teacher can peek into the black box of the test and see all the items and the student's responses. This is possible in a low-stakes test like this, where item exposure is not an issue. The teacher could for example look closer at some selected students' way through the test. Teacher guidance material associated with the test could offer concrete pathways to making sense of the students' answers. Opting for an MST (see previous section) would in addition ensure that the whole class have read the same texts in the routing module, even if the test is adaptive. This could form the basis for whole-class conversations about these texts.

In phase 2, these alternative formats will be tested out in connection with pilot 3 (the first adaptive pilot). Teachers will receive test reports for their students in one of the alternative formats. The perceived usefulness of the reports, and teachers' motivation to act on the results based on the reports, will be investigated through focus group interviews and surveys, comparing the different formats. Phase 3 is comprised of improvement of the most compelling format from phase 2, alongside the development of additional guiding material for pedagogical use of the reading test results. The guidance material will also include more general professional development material on reading instruction and on using assessment data for differentiating instruction, using examples from the test. The quality of the report format and guidance material, when it comes to their ability to support adapted reading instruction, will be subject to investigation through survey and focus group interviews that will be carried out both just after the final test is operational in February 2023 and towards the end of the school year.

Conclusion

In this protocol, we have outlined a proposed interpretation and use of a formative adaptive reading test for 3rd grade. We have also outlined a pathway towards a test concept that would enable such a use. Together, these two moves form what Kane calls an interpretation and use argument (2015).

As stated by Kane, it is "appropriate for test developers [during the development stage] to make the case for the proposed interpretation and use of the scores, but at some point, it is necessary to shift to a more critical evaluation of the claims being

made" (p. 71). At the end of the project, we will return to the interpretation and use argument here laid out and evaluate it critically in a validity argument.

Author Biographies

Arild Michel Bakken is Associate Professor of Literacy Studies at the Norwegian Reading Centre, University of Stavanger. His research interests include reading assessment, literature and early literacy instruction.

Aslaug Fodstad Gourvennec is Associate Professor in Literacy Studies at the Norwegian Reading Centre, University of Stavanger. She specializes in the fields of literature didactics and the L1 subject in secondary school.

Bente Rigmor Walgermo is Associate Professor in Special Needs Education at the Norwegian Reading Centre, University of Stavanger. Her main field of research concerns the relationships among interest, motivational beliefs and reading skill.

Oddny Judith Solheim is Professor in Special Needs Education at the Department of Education and Sport Sciences, University of Stavanger. She specializes in the fields of engaging literacy instruction, reading and writing difficulties and assessment of reading.

Njål Foldnes is Professor in Statistics at the Norwegian Reading Centre, University of Stavanger. He conducts research on statistical methodology for the social sciences and on innovative teaching approaches in higher education.

Per Henning Uppstad is Professor in Special Needs Education at the Norwegian Reading Centre, University of Stavanger, and Professor II at Volda University College. His research interests include technology enhanced learning, reading and writing, educational assessment and early literacy intervention.

Acknowledgements

This work was supported by The Research Council of Norway under Grant 285207.

References

Arnesen, A., Braeken, J., Baker, S., Meek-Hansen, W., Ogden, T., & Melby-Lervåg, M. (2017). Growth in oral reading fluency in a semitransparent orthography: Concurrent and predictive relations with reading proficiency in Norwegian, grades 2–5. *Reading Research Quarterly, 52*(2), 177–201. https://doi.org/10.1002/rrq.159

Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*(1). https://doi.org/10.1177%2F003172171009200119

Bundsgaard, J. (2018). Pædagogisk brug af test [Pedagogical use of tests]. *Sakprosa, 10*(2). https://doi.org/10.5617/sakprosa.6007

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2017). *SimDesign: Structure for organizing Monte Carlo simulation designs. R package version 1.6.* Phil Chalmers.

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change, 17*(1), 7–28.

Directorate for Education and Training. (2017a). *Framework for basic skills.* https://www.udir.no/in-english/Framework-for-Basic-Skills/

Directorate for Education and Training. (2017b). *Rammeverk for nasjonale prøver* [Framework for the national tests]. https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover2/

Education Act. (1998). *Act relating to primary and secondary education and training* (LOV-1998- 07-17-61). Lovdata. https://lovdata.no/NLE/lov/1998-07-17-61

Evers, A., Hagemeister, C., & Hostmaelingen, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests (Tech. Rep.Version 4.2. 6)*. European Federation of Psychology Associations.

Eysink, T. H., & Schildkamp, K. (2021). A conceptual framework for Assessment-Informed Differentiation (AID) in the classroom. *Educational Research, 63*(3), 261–278. https://doi.org/10.1080/00131881.2021.1942118

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6–10. https://doi.org/10.1177%2F074193258600700104

Gunnulfsen, A. E., & Roe, A. (2018). Investigating teachers' and school principals' enactments of national testing policies. *Journal of Educational Administration, 56*(3), 332–349. https://doi.org/10.1108/JEA-04-2017-0035

Kane, M. (2015). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 80–96). Routledge.

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018). Unpicking the developmental relationship between oral language skills and reading comprehension: It's simple, but complex. *Child Development, 89*(5), 1821–1838. https://doi.org/10.1111/cdev.12861

Mair, P. (2018). *Modern psychometrics with R*. Springer International Publishing.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, 76*(1), 1–19. https://doi.org/10.18637/jss.v076.c01

Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR.* Springer.

Mausethagen, S., Prøitz, T., & Skedsmo, G. (2018). Teachers' use of knowledge sources in "result meetings": Thin data and thick data use. *Teachers and Teaching, 24*(1), 37–49. https://doi.org/10.1080/13540602.2017.1379986

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012. https://doi.org/10.1002/j.2333-8504.1979.tb01178.x

Ministry of Children and Education. (2020). *Rådgivningsgruppen for evaluering af de nationale tests: anbefalinger, januar, 2020* [Advisory group for the evaluation of the national tests: recommendations, January, 2020]. https://www.uvm.dk/publikationer/2020/200206-raadgivningsgruppen-for-evaluering-af-de-nationale-test

Mullis, I. V., & Martin, M. O. (2019). *PIRLS 2021 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement.

Nation, K., & Cocksey, J. (2009). The relationship between knowing a word and reading it aloud in children's word reading development. *Journal of Experimental Child Psychology, 103*(3), 296–308. https://doi.org/10.1016/j.jecp.2009.03.004

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software manual]. https://www.r-project.org/

Ricketts, J., Nation, K., & Bishop, D. V. M. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading, 11*, 235–257. https://doi.org/10.1080/10888430701344306

Schildkamp, K., & Datnow, A. (2020). When data teams struggle: Learning from less successful data use efforts. *Leadership and Policy in Schools*, 1–20. https://doi.org/10.1080/15700763.2020.1734630

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education, 26*(3), 482–496. https://doi.org/10.1016/j.tate.2009.06.007

Seymour, P. H., Aro, M., & Erskine, J. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*(2), 143–174. https://doi.org/10.1348/000712603321661859

Stallman, R. (1997). *Linux and the GNU Project*. Free Software Foundation. http://www.gnu.org/gnu/linux-and-gnu.html

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*(2), 161–179. https://doi.org/10.1080/00131880902891305

Tønnessen, F. E., & Uppstad, P. H. (2015). *Can we read letters? Reflections on fundamental issues in reading and dyslexia research*. Brill.

Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2016). Data use by teachers: The impact of motivation, decision-making style, supportive relationships and reflective capacity. *Educational Studies, 42*(1), 36–53. https://doi.org/10.1080/03055698.2016.1148582

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Walgermo, B. R., Uppstad, P. H., Lundetræ, K., Tønnessen, F. E., & Solheim, O. J. (2021). Screening tests of reading: Time for a rethink. *Acta Didactica Norden*, *15*(1). https://doi.org/10.5617/adno.8136

Wilson, M., & Gochyyev, P. (2013). Psychometrics. In T. Theo (Ed.), *Handbook of quantitative methods for educational research* (pp. 1–30). Brill Sense.

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 3–20). Chapman and Hall. https://doi.org/10.1201/b16858