

## Predicting first grade students' writing proficiency

Gustaf B. Skar & Alan Huebner

To cite this article: Gustaf B. Skar & Alan Huebner (2022) Predicting first grade students' writing proficiency, *Assessment in Education: Principles, Policy & Practice*, 29:2, 219-237, DOI: [10.1080/0969594X.2022.2057424](https://doi.org/10.1080/0969594X.2022.2057424)

To link to this article: <https://doi.org/10.1080/0969594X.2022.2057424>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 04 Apr 2022.



Submit your article to this journal [↗](#)



Article views: 395





View related articles [↗](#)



View Crossmark data [↗](#)

## Predicting first grade students' writing proficiency

Gustaf B. Skar <sup>a</sup> and Alan Huebner <sup>b</sup>

<sup>a</sup>Department of Teacher Education, Norwegian University of Science and Technology (NTNU), Trondheim, Norway; <sup>b</sup>Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA

### ABSTRACT

This study aimed to investigate the predictability of writing development and if scores on a writing test in the first weeks of first grade accurately predict students' placements into different proficiency groups. Participants were 832 first grade students in Norway. Writing proficiency was measured twice, at the start and at the end of first grade (time 1 and time 2, respectively). Multilevel linear regression analysis showed that writing proficiency measures at time 1 were significant predictors of writing proficiency at time 2. The results also showed that measures at time 1 could identify students running the risk of not meeting expectations with high precision. However, the results also revealed a substantial proportion of false positives. The results are interpreted and discussed from a formative writing assessment perspective.

### ARTICLE HISTORY

Received 29 March 2021  
Accepted 20 March 2022

### KEYWORDS

writing assessment; writing development; predictability

## Introduction

Writing is a powerful tool for thinking, learning, and communicating. In many school systems, students are expected to learn to write for these purposes from first grade (e.g. Jeffery et al., 2018; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Failing to master writing can impede a student's opportunity for engagement in learning activities in school, and social and societal activities outside school (Gee, 2004; Organisation for Economic Co-operation and Development, 2005). Thus, writing has since the early 2000s been a so-called 'key competence' in Norwegian schools, and through the national curriculum, all teachers are responsible for teaching subject relevant writing (Skar & Aasen, 2021). On 8 June 2018, the Norwegian parliament further stressed the importance of developing writing proficiency by accepting an amendment to the Education Act requiring schools to offer extracurricular writing instruction to students that 'risk falling behind in reading, writing and mathematics [...] so that the expected progression is met' (Opplæringslova [The Education Act], 1998, §1-4, *our translation*). As outlined in Skar et al. (in press) this amendment has been problematic; the regulation was not accompanied by a definition of the 'expected progression', or tools for identifying the students in need of extracurricular writing instruction.

**CONTACT** Gustaf B. Skar  [gustaf.b.sk@ntnu.no](mailto:gustaf.b.sk@ntnu.no)  Department of Teacher Education, NTNU, Trondheim NO-7491, Norway

Although writing is a key competence and there are standards for writing proficiency from second grade, there is no designated official writing test in Norway. Writing is tested as a part of the Language Arts subject exam in tenth grade, but students do not receive a writing score. Consequently, there is a substantial lack of tools for teachers who are responsible for offering extracurricular instructions to students referenced by the Education Act. A recent researcher-initiated project set out to devise indications of what teachers and other experts would regard as expected progression at the end of first, second, and third grade (Skar et al., *in press*). The current investigation should be regarded as a ‘sister project’, as it aimed to investigate the predictability of writing development and if scores on a writing test in the first weeks of first grade accurately predict students’ placements into different proficiency groups – defined by Skar et al. (*in press*) – at end of first grade. If measures of writing can predict future writing performance, teachers could engage in formative assessment by using outcomes of that measure for planning the instruction.

The goal of formative assessment is to increase the probability of successful instruction by adopting a dynamic approach to the curriculum by allowing successive collected information about students’ proficiency guide the subsequent instruction (Black & Wiliam, 2009, p. 9). The validity of formative assessment is contingent upon its consequences; formative assessment is valid if it has desired consequences (Nichols et al., 2009; Stobart, 2012). For writing instruction, formative assessment can be operationalised by using the cycle outlined by Graham et al. (2018). This cycle includes stating the objectives of the writing instruction, instruct students, collect and analyse evidence of writing proficiency, and finally react to the analysis by, for example, altering the instruction to better meet the needs of the students. For example, a teacher can, by observing students’ writing processes and the ensuing texts, focus his/her attention to stages in the students’ processes that currently works sub-optimal.

Meta-analyses have proven the use of formative writing assessment to be successful. Graham et al. (2012) reported an average effect size in elementary education of  $d = 0.42$  for assessment in general, and  $d = 0.80$  for feedback from adults. A subsequent meta-analysis (Graham et al., 2015) found the average effect size for various types of feedback to be  $d = 0.61$  and the effect of adult feedback to be  $d = 0.87$ . Similar findings were also reported by Koster et al. (2015). Koenka et al. (2019) even found a positive effect of providing feedback in the form of grades – as compared to no feedback at all – although written comments were more effective than grades, with an average effect size of  $d = 0.30$  for comments versus grades.

To help teachers who are responsible for providing extracurricular instruction to students running the risk of not developing writing skills in accordance with expectations, there is a need for a tool that can predict future writing performance. Without fulfilling these requirements, an educator would not know what a piece of student writing represents in terms of anticipated development. Previous research has explored both how writing develops and to what extent writing proficiency can be predicted.

### ***Writing development and predictions of writing ability***

Some studies have investigated writing development in elementary school over time to explore developmental patterns. For example, Bae and Lee (2012) measured writing ability in students ( $N = 42$ ) from second to sixth grade in South Korea and found on

average students display gains in all measures (e.g. grammar, coherence, and text length), albeit at different rates. Kim et al. (2015a) investigated the writing proficiency development in US students ( $N = 304$ ) attending first grade. The authors noted a general increase in all measures, and that the pace of development could not be explained by extraneous factors such as eligibility for free lunch. Drijbooms et al. (2017) surveyed writing proficiency development in Dutch students ( $N=93$ ) in grades 4–6 and found that students develop in terms of syntactic complexity, without noticeable development in narrative writing skills. Hsieh (2016), in a cross sectional study of linguistic development of 765 essays written by students in Singapore in grades 3–6, noticed ‘a general upward trend in lexical and syntactic development over the 4 years’. In two cross-sectional studies, Graham (1999), and Graham et al. (2001), investigated handwriting development in US students ( $N=900$ ) in grades 1–9 and US students ( $N=300$ ) in grades 1–3, respectively. Both studies found that the average student displayed an increase in measures of handwriting, but the developmental pace was different between grade levels, with a ‘plateau tendency’ increasing with grades.

These studies point to two quite general predictions: given instruction, children are likely to develop as writers, and the nature and pace of that development will presumably differ between grade levels and be related to the certain aspect of writing being investigated.

Other studies investigated predictors of writing proficiency. Campbell et al. (2019) investigated predictors among 62 students in first grade in the US and found name-writing ability, but not demographic variables such as gender or age, at the start of the school year to be a predictor of writing proficiency after six months. Hooper et al. (2010) investigated predictors of narrative writing skills among 65 US students who were followed from kindergarten to grades 3–5 and found the following predictors to be statistically significant: core language abilities, maternal education and prereading skills. Similar findings were reported in a study of 157 US students in grade K–3 (Kim, et al., 2015). Juel (1988) investigated writing predictors among 54 US students in grades 1 and found that writing competence in first grade predicted writing competence in fourth grade ( $r = .38$ ).

Two recent studies (Wilson, 2018; Wilson et al., 2016) investigated the classification accuracy of measures of writing proficiency. In Wilson et al. (2016), the writing scores of 272 US students in sixth grade, awarded after an assessment in the fall, were used to assess to what extent that writing measure could be used to accurately predict students being at risk in the subsequent spring semester. The results indicated that scores on the fall assessments provided acceptable prediction accuracy.

In Wilson (2018), 230 US students’ writing proficiency in third and fourth grade were assessed using a screening tool. A subsequent state assessment was used to identify students meeting or not meeting requirements of the grade levels. The screening tests, based in the tradition of ‘direct assessment of writing (DAW)’, meant that students wrote texts in reply to a prompt specifying a topic and a rhetorical situation. Wilson (2018) reported his models yielded acceptable levels of discrimination for third grade and excellent levels for fourth grade.

Taken together, development and prediction studies indicate that one can presume that writing skills develop over time, and that the development will probably be related to prior achievements. The studies also indicate that development is not likely to be linear.

## The current study

One could argue that promising results of formative assessment, the explorations of writing development and predictability of writing proficiency would call for a Norwegian adaptation of existing formative assessment programmes and writing measures. However, such an approach would not take into account the contextual factors of writing. Theoretical (Berge et al., 2016; Graham, 2018; Russell, 1997) and empirical (Jeffery, 2009; Jeffery et al., 2018; Purves, 1992) findings suggest that different school contexts value different types of writing. Investigations of writing instruction and writing assessment, of which there is little outside the US (Graham & Rijlaarsdam, 2016; Zheng & Yu, 2019), suggest that instruction differs between contexts (Graham et al., 2021). These findings illustrate the importance of conducting studies in various contexts. Replication studies have the ability to extend the knowledge base, and to provide context relevant knowledge about writing development and accuracy of predicting writers' future proficiency. We briefly outline some aspects of the Norwegian school context.

In Norway, students enter school the year they turn six years old. The age difference in a class can be eleven months. A Norwegian school year starts in August and ends in June. Teacher generated grades are awarded from the start of eighth grade, and in tenth grade students sit for an external exam.

Writing is supposed to be taught in all subjects because of its status as a 'key competence', but Language Arts has the main responsibility for writing instruction (Skar & Aasen, 2021).<sup>1</sup> A recent survey among Norwegian teachers in first to third grade (Graham et al., 2021) indicated that teachers typically offered varied writing instruction and devoted approximately 20 minutes per day to writing instruction.

Currently, students sit for national tests in English as a foreign language, mathematics and reading in fifth, eighth and ninth grade. One of the purposes of these tests is to provide teachers with insights into student achievement within these domains, to better plan the subsequent instruction.

## Research questions

We posed two research questions:

- (1) Was students' writing proficiency at the start of first grade a significant predictor of their writing proficiency at the end of first grade, after controlling for numerous school-level and individual-level factors?
- (2) With what accuracy could scores on a writing test at the beginning of first grade accurately predict students' placements into different proficiency groups for text quality?

If writing is predictable, this would imply that the writing measures can be used by teachers in their formative assessment, given that the accuracy was acceptable enough. In other words, our understanding of the possible outcomes relates to both statistical significance and practical significance. Without the latter, the former may be of lesser interest for practitioners. Based on previous research, we expected the writing

proficiency to be predictable, but based on the arguments about context sensitivity above we refrained from formulating a hypothesis about the exact nature of this predictability.

## Method

### Participants

Participants were 832 students from 62 classes in 26 schools. All students attended first grade, and their mean age was 73.6 months ( $SD = 3.3$ ). There were 402 boys, (48% of the total sample), and six hundred forty students (77%) had Norwegian as their first language, 134 students (16%) were bilingual with Norwegian as one of the first languages, while 58 students (7%) had a first language other than Norwegian.

Schools in our sample were located in four municipalities, two urban and two rural, with relatively large and small populations, respectively. There was a mixture of large and small schools with a mean school size of 483.6 students ( $SD = 164.1$ ). The moving average result on the above mentioned national tests in reading, English, and mathematics in the period 2018–2021 for the schools was 50.6 ( $SD = 2.3$ ).<sup>2</sup> On average, 95.4% ( $SD = 5.2\%$ ) of teachers in these schools were certified. The average number of students per special education teacher was 92.5 ( $SD = 34.6$ ) and students in these schools averaged 55.4 instructional hours ( $SD = 8.6$ ), a metric that states the number of instructional hours divided by the number of students.

We argue that the sample was representative of the population. There were 50.2% boys attending first grade in the academic year of 2019–2020. In 2019–2020, 7.9% of students in Norway were entitled to extra-curricular language instruction, which is close to the proportion of L2 speakers in the sample. The size of the schools included in the material was somewhat bigger than the national average size ( $M = 226$ ,  $SD = 166$ ). The municipalities, however, represented the population range of municipalities in Norway with a range from large (population of 673,469, or 13% of the population) to average sized (population of 13,958, or 3% of the population) to smaller municipalities (population of 6,882, or 1% of the population). The municipalities were located at different areas of Norway, and thus represented education in various parts of the country. The proportion of certified teachers in the sample as well as school hours per student were similar to national average, which was 95.% and 61 hours, respectively. Also, the average number of students per special education teacher was similar to the national average, which was 82.4 ( $SD = 98.2$ ). Finally, the average national test score in our sample ( $M = 50.6$ ,  $SD = 2.3$ ) was close to the national average of 50 ( $SD = 10$ ). Thus, we argue that the effects of the measures presented below are generalisable to the Norwegian population.

### Sampling procedure

Students were recruited at the school level as participants as part of a large scale writing intervention study (Skar et al., 2020). After recruitment, half of the schools were randomly selected to an intervention group, while the other half formed the control

group. In the current study, only students from the latter group participated, as estimates from the intervention group probably would be less generalisable to the population, since those students had participated in a writing instruction program.

In this control group 1,333 students from 72 classes in 28 schools were approached, and 1,170 (87.8%) participated after consent from parents and guardians. Of those, 832 (71.1%) students, nested in 62 classes in 26 schools, participated in enough data collection for reliable estimates of writing proficiency to be calculated.

Due to the nested structure of the data, the major analytical strategy for this study was multi-level regression analysis. For regression models, the estimate of statistical power may focus on inference for a coefficient, a variance component, or on the means of particular groups (Snijders, 2005). The current study focuses on the coefficient indicating the effect of the predictability of writing measures, and the estimated statistical power of the study was high (> 99%) To obtain the power stated above, we ran simulations using the *MLPowSim* software (Browne & Golarizadeh, 2021). Given the very high statistical power, we will subsequently discuss issues of practical significance versus statistical significance.

## **Measures**

Students were administered two types of measures related to writing proficiency, on two occasions. In total, there were six tasks, of which four were used to derive measures of text quality, and two were used to derive measures of handwriting fluency. The former measure related to several aspects of writing proficiency, while the latter related to transcription skills.

There were two justifications for using these measures of writing proficiency. First, they have both been used in a previous investigation including 4,950 Norwegian students in grades 1–3, which had provided norming data (Skar et al., 2022). Second, another previous study used the Text Quality Scale associated with the text quality measure to establish cut scores for three proficiency levels, including the level referred to in Norwegian law.

### **Text quality measure**

To obtain a measure of text quality, students were administered two ‘functional’ writing tasks at two time points (Time 1 [T1], Time 2 [T2]), resulting in four texts. The tasks were developed for data collection purposes within the project Functional Writing in Primary School (FUS), detailed in Skar et al. (2020). They were ‘functional’ in the sense of targeting the communicative function of writing. The tasks were similar to those in DIW writing (Wilson, 2018).

The tasks at Time 1, administered at the start of first grade, were called ‘Recess Time’ and ‘Lunch Box’, respectively. Recess Time prompted students to write ‘a letter where you describe what you usually play with during recess time’ to staff at ‘the University in Trondheim’. This letter was to be a reply to a letter from the same researchers asking students across the country to provide information about popular activities during recess time. Teachers first read the letter aloud and then led a discussion about popular recess time activities, how someone might describe something to a non-present reader, and

common features associated with letters. Teachers also presented a visual portraying students playing in a school yard. Students started writing when the teacher was convinced that all students had decided on elements to include.

The Lunch Box task prompted students to describe ‘the lunch box of your dreams to someone who cannot see or smell its content’. This administration also had teachers lead a similar discussion as described above, as well as projecting a visual, this time a blue lunch box containing typical Norwegian fare.

At Time 2 – at the end of first grade – students again were administered the Recess Time task. This time the letter from the researchers was amended with this sentence: ‘it’s been almost a year since you started school – what do you usually play with during recess time now?’ Students were also administered the task ‘Magical Hat’. This task prompted students to picture themselves finding a magical hat, which would grant the wearer any wish. The task further prompted students to recount what happened that day. Teachers engaged students in a pre-writing discussion and provided visual aid, this time in the form of a hat laying on a road surrounded by green fields.

There were two reasons why two tasks were administered at each time. First, research in primary school writing has shown multiple tasks to be effective to increase generalisability of scores (Bouwer et al., 2015; Graham et al., 2016; Kim et al., 2017; Olinghouse et al., 2012). Second, given the use of proper statistical analyses, multiple tasks would allow the researchers to estimate text quality scores for students who were absent for one of the tasks. There were 797 students participating in both tasks at the first measurement point. Fifteen students participated in the recess task only, and 20 students took the lunch box task as the only task. At Time 2, 756 students participated in both tasks. Seventy-six students participated in the recess task only, and 15 students took the magic hat task as the only task.

A pool of trained raters read and rated each student text using eight validated ratings scales, published in Skar et al. (2020). The scores were combined and scaled to a single Text Quality Score using a many-facet Rasch approach (see below). Raters assigned a score of 1–5 on the following rating scales: audience awareness, organisation, content relevance, vocabulary, sentence construction, spelling, legibility and punctuation. Higher scores represented better quality for each aspect of writing assessed. All rating scales had descriptors for all scale steps, and raters were also provided with exemplar text associated with each descriptor.

The scales targeted different aspects of text quality. For example, audience awareness concerned how the text addressed a reader, whereas the rating scale organisation concerned the structure of the text at the macro-, meso-, and micro-level.

The pool of judges consisted of 33 trained raters recruited from the university of the first author. The judges participated in training sessions, focusing on how to understand and use the rating scales. The raters supplied their 25 first ratings to the first author who provided feedback on harshness and consistency. During the rating period, an email list was available for raters to address the first author and fellow judges with questions. When a question was posted on the list, the rating stopped, and raters were given directions according to answers to the question posed.

All texts were blinded, and each text was assessed independently by two raters. There were two separate occasions, Rating Occasion 1 and Rating Occasion 2, that corresponded to Time 1 and Time 2, respectively. To provide an empirical link between the raters and



students, all raters scored 50 student texts from a previous rating session. This empirical link was used to scale the raw scores with a many-facet Rasch model (MFRM; Linacre, 2018):

$$\log \frac{P_{nmij(k)}}{P_{nmij(k-1)}} = B_n - A_m - E_i - C_j - F_x$$

Here,  $P_{nmij(k)}$  represents the probability of student  $n$ , on task  $m$ , rated on rating scale  $i$ , by rater  $j$ , receiving a score of  $k$ , and  $P_{nmij(k-1)}$  represents the probability of the same student under the same conditions receiving a score of  $k - 1$ .  $B_n$  is the ability for person  $n$ ,  $A_m$  is the difficulty of task  $m$ ,  $E_i$  is the difficulty of rating scale  $i$ , and  $C_j$  is the severity of rater  $j$ . Finally,  $F_x$  represents the point on the logit scale where category  $k$  and  $k-1$  are equally probable.<sup>3</sup>

If the data fits the MFRM, one can use the scaled score as an estimate of the outcome of interest – in this case Text Quality. The computer software FACETS converts the scaled score back to a scale of the same length as the original scale. This converted score is called a ‘fair score’ in the FACETS lingo, and it represents a ‘fair’ measure of text quality, which means that it is adjusted for differing task and rating scale difficulties and differing rater severity.<sup>4</sup>

The data fit the model well; specifically, the ‘reliability of separation’-measure, which is a MFRM analogue to Cronbach’s  $\alpha$ , was .95 for Rating Occasion 1 and .96 for Rating Occasion 2, which indicates a reliability fit for high stakes assessment in the measure of students’ writing quality. The standardised residuals also indicated a reasonable fit for the ratings at Rating Occasion 1, as 2.90% were in the range of  $|2-3|$  and 2.0% exceeding  $|3|$ , and a good fit for the ratings at Rating Occasion 2 as 4.10% were in the range of  $|2-3|$  and 0.80% exceeding  $|3|$ . Standardised residuals above 2 should preferably not exceed 5%, and standardised residuals above 3 should preferably not exceed 1% (Eckes, 2011).

### Copying task

To measure the handwriting fluency of students, teachers administered a copying task (Graham et al., 1997). Students were instructed to copy a short text as quickly and accurately as possible within 1.5 minutes. The paragraph was taken from the Group Diagnostic Reading and Aptitude and Achievement Tests (Monroe & Sherman, 1966).

To standardise the task ensure that students understood the text they were going to copy, students were first shown an instructional video explaining how to complete the task. The teachers were instructed to then read the paragraph aloud and to start the task when s/he assessed all children to have understood the instruction. Students were instructed to start and stop copying the text at the teacher’s command.

The handwriting fluency measure for each student was obtained through a process where trained coders would tally all correctly written letters. The coders did not count incorrectly written letters, or letters that were skipped or correctly written letters that did not match the text. To estimate inter-rater reliability, 10% of all tasks were re-coded by an independent co-coder. The reliability was satisfactory ( $\kappa = .812$ , ICC = .99). The score for handwriting was divided by 1.5 to provide an estimate of handwriting fluency per minute.

### **Classifications of writing proficiency**

To classify performance regarding text quality we used cut scores that were established through a standard setting procedure and documented in a separate investigation (Skar et al., *in press*). This investigation based the standard setting procedure on measures of text quality that were closely related to measures in this study; students in Skar et al. (*in press*) were administered the recess time writing task.

In Skar et al. (*in press*), three levels of proficiency were defined: ‘a warning zone’, ‘minimum level’, and ‘aspirational level’.<sup>5</sup> The warning zone denoted proficiency that would yield extracurricular writing education at the end of first grade. The term itself referred to that the performance would warn educators about performance not meeting expectations. For the sake of brevity we will refer to at risk of performing in the warning zone as ‘at risk’. The cut score the at risk level was set to 2.0 on the Text Quality Scale. The minimum level referred to minimally acceptable proficiency at the end of second grade, and the cut score was set to 2.3. The aspirational level referred to the level of proficiency that panellists in the standard setting procedure would wish students to attain at the end of first grade. The cut score for this level was set to 3.7 on the Text Quality Scale.

In the classification process, all students’ performances with Text Quality Scores of  $\leq 2.0$  were classified as ‘at risk’. Performances in the range  $>2.0 \leq 2.3$  were classified as ‘below minimal’, and performances  $\geq 3.7$  were classified as ‘aspirational’. Performances between the minimum level and aspirational level were classified as ‘above minimum, below aspirational’.

As can be noted, the gap between the ‘at-risk level’ and the ‘minimum level’ amounted to 0.3 units on the Text Quality Scale. Although this difference may seem small, it is worth noting that a previous study based on comparable measures of text quality found the average second grade student to score 0.43 units more than the average first grade student (Skar et al., *in press*). The difference between 2.0 and 2.3 can thus be understood as substantial.

### **Covariates**

We included the following covariates for the school level in the analysis: National Test Result, School Size, Proportion Certified Teachers, Students per special education teacher, and School Hours per Student. Data for these variables, which were described in the participants section above, were collected from the database ‘Skoleporten’ [The School Gate], hosted by the governmental agency the Norwegian Directorate for Education and Training (<https://skoleporten.udir.no/>). Covariates for student levels were gender, language background, late administration, and time. Data for the first two variables were collected by teachers who indicated the students’ gender, and if the student learned Norwegian or another language first, or both. Late administration, which was a dummy variable, concerned students that for some reason were administered the task later than the other students. The covariate was included to control for possible effects of differences in time of administration

**Table 1.** Student-level variables (i.e. outcome variables), broken down by gender and language background.

Outcome	Gender	Language	N	Time 1		Time 2	
				M	SD	M	SD
HF	Boy	BL	73	6.02	3.14	12.58	6.69
	Boy	L1	299	5.19	2.60	12.15	6.25
	Boy	L2	30	4.82	2.08	10.51	5.46
	Girl	BL	61	6.15	2.67	14.72	6.48
	Girl	L1	341	6.53	3.71	14.36	6.85
	Girl	L2	28	4.00	3.27	10.83	6.37
	Total		832	5.83	3.20	13.18	6.64
TQ	Boy	BL	73	1.21	0.36	2.26	0.68
	Boy	L1	299	1.28	0.39	2.41	0.55
	Boy	L2	30	1.08	0.18	2.20	0.63
	Girl	BL	61	1.20	0.31	2.52	0.60
	Girl	L1	341	1.36	0.39	2.68	0.53
	Girl	L2	28	1.11	0.21	2.29	0.61
	Total		832	1.29	0.38	2.51	0.58

HF: Handwriting fluency, TQ: Text quality, T1: Time 1, T2: Time 2. L1: Norwegian as first language, L2: Norwegian as second language, BL: Norwegian and another language as first languages.

**Table 2.** School and student level covariates.

School-level	Short Name	N	M (SD)
National test result	nation_test	832	50.63 (2.29)
School size	school_size	832	438.55 (164.11)
Proportion certified teachers	prop_certif	832	95.39 (5.21)
Students per special education teacher	stud_speced	832	92.46 (34.63)
School hours per student	hours	832	55.42 (8.62)
Student-level			
Gender dummy variable (with girl coded as 1)	gender	832	0.52
Speakers with Norwegian as L1	L1	640	0.77
Speakers with Norwegian as L2	L2	58	0.07
Speakers with Norwegian and additional language as L1	BL	134	0.16

Outcome variables and covariates are summarised below. [Table 1](#) summarises the names and descriptive statistics of student level measures and [Table 2](#) the covariates used in the analyses.

There were moderate correlations between the measures of writing proficiency, spanning from  $r = .24$  ( $p < .001$ ) between handwriting fluency at Time 1 (HFT1) and text quality at Time 2 (TQT2), to  $r = .42$  ( $p < .001$ ) between handwriting fluency at Time 2 (HFT2) and TQT2. The handwriting fluency measures had a correlation of  $r = .37$  ( $p < .001$ ) between Time 1 and Time 2, and the text quality measures displayed a relationship of  $r = .40$  ( $p < .001$ ) between Time 1 and Time 2. Please refer to [Table 3](#) for all correlations.

**Table 3.** Correlations between outcome measures.

	HFT1	TQT1	HFT2	TQT2
HFT1	–			
TQT1	0.310***	–		
HFT2	0.368***	0.312***	–	
TQT2	0.243***	0.394***	0.418***	–

\*\*\*  $p < 0.001$ . Note. HF: Handwriting fluency, TQ: Text quality, T1: Time 1, T2: Time 2.

## Data collection procedures

All tasks were administered by students' own teachers. This was consistent with the tradition in Norway, where external assessments are not introduced before tenth grade, and formal grades are not set until eighth grade. The first author worked with teachers in first grade to design a test administration procedure that would fit the participating students. All teachers who participated were given instructions provided by the first author. Moreover, there were instructional videos available for teachers to consult if they had questions. There was also a telephone hotline teachers could use to call the project group, as well as an email address teachers could use to pose questions. Some teachers had questions not pertaining to the study, such as how to get refunded for post stamps after sending the student texts to the university.

Students were given 45 minutes to complete each 'functional' writing task. The copying tasks were restricted to 90 seconds. Data for the first measurement point was collected during weeks 35–36 in 2019 (i.e. August 26–September 6), which meant that data collection started one week after students entered first grade. Data for the second measurement point was collected in weeks 22–25 in 2020 (i.e. May 25–June 16), which was close to the last day of the second semester of the students' first school year.<sup>6</sup> The data collections were counterbalanced, so that no task exclusively was first, middle or last.

## Analytical strategy

Multilevel linear regression models were used as the main method for analysing the data, as these models are well suited for handling the clustered nature of the data. The R package *lme4* (Bates et al., 2015) was used to fit the models, and the package *lmerTest* (Kuznetsova et al., 2017) was used to obtain the *p*-values. Random intercepts were included for classes and schools. The continuous covariates described in the previous section were standardised so that regression effects can be interpreted in terms of standard deviations. For each of the two outcome variables, three models were fit. The first model (1) was a null model with no predictors, so that the correlation structure could be examined. Then, two more models were fit: (2) a model with all predictors except for the time 1 measurement of the covariate, and (3) a model with all predictors. For example, for the outcome HFT2, model (2) consists of all predictors except for HFT1, and model (3) consists of all predictors. The difference in values between models (3) and (2) quantifies the unique contribution of HFT1 in explaining the variation in HFT2. Furthermore, examination of the regression slope in model (3) for HFT1 will determine whether it was a significant predictor of HFT2 (i.e. research question 1), and similar for TQT1 and TQT2.

The second research question, addressing the accuracy of predicting TQT2, was addressed using model (3) for TQT2. Predicted text quality scores and prediction intervals for each student were obtained from the model using the R package *merTools* (Knowles & Frederick, 2020). The prediction intervals were created using a 68% confidence level, as this is stated by Harvill (1991) as the most common level for creating

intervals around test scores. Then, the intervals were used to assess the accuracy of two different types of classifications: (1) whether the students were at risk vs. not at risk, and (2) whether the students were at risk or below minimal vs. above minimal or aspirational. For (1), students were classified as being at risk if the lower bound of their prediction interval was  $\leq 2.0$ . For (2), students were classified as being at risk or below minimal if the lower bound of their prediction interval was  $\leq 2.3$ . The classification accuracy as assessed by using ‘confusion matrices’, a method used in a similar manner by Wilson (2018). These matrices are tables which summarise the true positive rates (sensitivity) and the true negative rates (specificity) of the classifications, which will be explained below in the context of the current study.

## Results

The effect sizes for the random and fixed effects for models for both outcomes are displayed in Table 4. The random effects can be summarised by the intraclass correlation coefficient (ICC). The ICCs for class and school were both higher for the text quality outcome than for handwriting fluency. This indicates that the text quality scores are (slightly) more highly correlated than writing fluency scores for students in a given class and school. The  $R^2$  values quantify the amount of variation in the outcome explained by the predictors, and the effect size  $f^2 = \frac{R^2}{1-R^2}$  is given in parenthesis. For the HWT2 outcome, the full model explains approximately 21.4% of the variation in HWT2, and HWT1 alone explains 9.4% of the variation in HWT2, when controlling for TQT1, and other covariates. The results for the TQT2 outcome are similar. The full model explains about 17.3% of the variation in TQT2, and the TQT1 predictor explains approximately 9.2% of the variation in TQT2, after controlling for HFT1 and other covariates.

Table 5 displays the parameter estimates and corresponding  $p$ -values for the full models for each outcome variable. There are similarities between the results; gender, HFT1, and TQT1 are highly significant predictors of both outcomes. In particular, holding all other variables fixed, we expect girls to score 1.324 units higher than boys on HFT2 and 0.21 units higher on TQT2. Also, holding all other variables fixed, for every one-standard deviation increase in HFT1 we expect a 2.217 unit increase in HFT2 and a 0.097 increase in TQT2, on average. Similarly, holding all other variables fixed, for every one-standard deviation increase in TQT1, we expect a 1.785 unit increase in HFT2 and a 0.157 unit increase in TQT2, on average.

**Table 4.** Effect sizes for the three models fit to each of the three outcome variables.

Quantity	Outcome	
	HFT2	TQT2
ICC (class)	0.078	0.191
ICC (school)	0.028	0.115
$R^2$ , model without T1 ( $f^2$ )	0.120 (0.136)	0.081 (0.088)
$R^2$ , model with T1 ( $f^2$ )	0.214 (0.272)	0.173 (0.209)
$R^2$ , difference ( $f^2$ )	0.094 (0.104)	0.092 (0.101)

HF: Handwriting fluency, TQ: Text quality, T1: Time 1, T2: Time 2.

**Table 5.** Parameter estimates and p-values for full models for HFT2 and TQT2 outcome variables. Continuous covariates were standardised.

Outcome	Handwriting Fluency		Text Quality	
	Estimate	P-val	Estimate	P-val
Intercept	13.288	<0.001	2.471	<0.001
nation_test	0.442	0.548	0.005	0.949
school_size	-0.001	0.607	0.000	0.334
prop_certif	-0.498	0.321	0.017	0.756
stud_speded	-0.409	0.358	-0.003	0.960
hours	-0.152	0.853	-0.008	0.933
Girl	1.324	0.001	0.210	<0.001
L1	0.005	0.993	0.103	0.044
L2	-1.091	0.266	-0.096	0.252
HFT1	2.217	<0.001	0.097	<0.001
TQT1	1.785	<0.001	0.157	<0.001

For Gender, the ref. level was boy. For Language, the ref. level was bilingual.

**Table 6.** Confusion matrix for classifying students as at risk (TQT2  $\leq 2.0$ ).

Predicted Status	Actual Status		Total
	At Risk	Not at Risk	
At risk	132	300	432
Not at risk	12	388	400
Total	144	688	832

As explained above, the full model reported in Table 5 for the TQT2 outcome was used to explore how accurately students can be classified as at risk. Table 6 displays the confusion matrix that summarises the accuracy of the classifications.

The diagonal entries of the confusion matrix denote the true positives (at risk students predicted to be at risk) and true negatives (students not at risk predicted to be not at risk). Thus, 132 of 144 students who were actually at risk were predicted to be at risk, yielding a true positive rate (or, sensitivity) of  $132/144 = 0.92$ . Also, 388 students were predicted to be not at risk out of 688 students actually not at risk, yielding a true negative rate (or, specificity) of  $388/688 = 0.56$ .

Table 7 shows a second confusion matrix that was produced to examine the accuracy of the model for classifying students as at risk/below minimal. We see that the true positive rate was very high ( $241/245 = 0.98$ ), while the true negative rate was low ( $157/587 = 0.27$ ).

**Table 7.** Confusion matrix for classifying students as at risk (TQT2  $\leq 2.3$ ,  $\geq 2.3$ ).

Predicted Status	Text Quality Actual Status		Total
	At Risk/Below Minimal	Above Minimal Aspirational	
At Risk/Below Minimal	241	430	671
Above Minimal Aspirational	4	157	161
Total	245	587	832

## Discussion

We posed two research questions in this investigation. The first research question regarded statistically significant predictors of writing proficiency. The results showed that handwriting proficiency measures at Time 1 (HFT1) were statistically significant predictors of handwriting proficiency measures at Time 2 (HFT2), explaining 9.4% of the variation for HFT2, which equalled an effect size of  $f^2 = 0.104$ . Text Quality at Time 1 (TQT1) explained 9.2% of the variation in Text Quality at Time 2 (TQT2) ( $f^2 = 0.101$ ). Effect sizes in this range are usually considered to be small to moderate (Cohen, 1992; Lorah, 2018). In assessing the magnitude of the effect sizes, however, one should note that the measures were derived from the very first weeks of schooling, where the variance was considerably smaller than at Time 2. This indicates that even small differences among students who were otherwise tightly clustered together was a predictor of future scores. Moreover, the analysis controlled for a large set of student and school level variables, that otherwise might have confounded the effect of time. The writing proficiency of the students in this sample was significantly predictable, even after controlling for gender, age, and within and between school variance, which were all by themselves significant predictors of outcome.

Previous studies, mainly conducted in the US, have also demonstrated predictability in elementary school writing (Campbell et al., 2019; Hooper et al., 2010; Juel, 1988; Kim et al., 2015a). This investigation adds to those studies by confirming predictability also in the Norwegian context, and by including a sample size much larger than most previous studies. It will have theoretical implications, should the pattern of predictability repeat itself in several contexts; this could potentially imply that a student's baseline continues to exert influence over his or her development 'independent' of instruction. We have seen that even at the group level, the slope of writing development is far from a straight line (Bae & Lee, 2012; Drijbooms et al., 2017; Graham, 1999; Graham et al., 2001; Hsieh, 2016; Kim., 2015; Montanari et al., 2016), but previous research (Kim et al., 2015a) has also shown that demographic variables did not explain the relative pace of development. In Kim et al. (2015a) the rank ordering of students did not change enough for demographic variables to be able to explain the pace of development, even though students – on group level – developed as writers.

To our knowledge, no study has investigated whether an intervention could fulfil the supplementary aims of increasing all students' writing proficiency, and tightening the distribution of outcome scores so much that predicting individual scores would be difficult or impossible. Because previous investigations had made it obvious that all students are capable of developing, research that targets the closing of gaps would be interesting as it might help educators better to understand how pedagogical actions could be tailored to individual needs.

The second research question examined with what accuracy text quality measures at Time 1 would predict a student's classification into proficiency groups. The results showed that 92% of the student performances classified as 'at risk' at Time 2 was accurately predicted by measures at Time 1. For student performances classified as 'below minimum', the accuracy was 98%. However, the results also indicated that a substantial low specificity; 632 (80.6%) of students were predicted to be in the below minimum category, 430 (64%) wrongfully so. We note that this pattern of high sensitivity and low specificity is consistent

with the results of the confusion matrices reported by Wilson (2018). While they used a logistic regression approach, the current study used linear regression. The reason for this is because it is because the response variable text quality is continuous, and it is generally recommended that continuous variables should not be collapsed, or discretized (Harrell, 2001). Interestingly, however, despite the different modelling approaches, the end classification results of Wilson (2018) and the current study appear to be concordant.

The pedagogical implications of high sensitivity and low specificity are far from straightforward. An educator mostly concerned with the sensitivity rate would be able to use the measures from Time 1 to assure extracurricular activities to students at risk of not developing writing skills in accordance with expectations, possibly with the outcome that this would lead to a more desired development pace for those students. However, a large proportion of students would – at least as long as the pedagogical decisions are based on this first measure only – receive a suboptimal treatment, that in worst consequence would hinder their development. A worst-case scenario, then, would be that all students end up tightly clustered around a mean score, at the cost of a substantial number of students having their development suppressed.

An alternative would be to adopt a more dynamic approach, as the one suggested by Graham et al. (2018), where the teacher frequently elicits information of the individual's writing performance for purposes of planning instruction. Such an approach would also be in line with research indicating the limited generalisability of one-off assessments (Bouwer et al., 2015; Kim et al., 2017; Olinghouse et al., 2012). A cautious start, taking at risk warnings seriously might initially not fit all students, but would be sure to identify those that indeed are in need of extracurricular instruction. Subsequent assessments will presumably provide more nuanced information, enabling the educator to better tailor instructional needs to the individual students.

### **Limitations**

This backdrop of this study was the need for context relevant tools for identifying students' strengths and potential weaknesses to better tailor instruction to individual need. While the study – as intended – did provide much needed insights into predictability and accuracy, future research will be needed to transform the tools that were explored in this investigation to pedagogical tools that can be applied in classroom settings. Such research would include teacher participants using the assessment tools to elicit evidence about students' performance and also include a series of instructional steps that would follow, depending on student outcomes. It would be preferable if such research could be designed in a way that would allow researchers to draw causal conclusions, so that teacher community could be provided with answers on how effective the tools were for improving the trajectories of students' writing development.

### **Conclusion**

This study showed that writing proficiency in the end of first grade was predictable using measures of writing proficiency at the start of first grade. It also showed a high ability to predict which students that performed at the at risk-level and below minimum level at end



of first grade. If nothing else, this predictability suggests for teachers to assess their students and act on that assessment to ensure a dynamic and tailored instruction, rather than a static curriculum, which might preserve and strengthen differences between students.

## Notes

1. There may, of course, be a discrepancy between the intended, official curriculum, and the one enacted. We do not know if writing is taught in all subjects.
2. We chose to use this metric, as it is the statistic the Norwegian Directorate for Education and Training uses for computing school-level added value. Initially, we were going to include added value instead, but the Directorate for Education and Training does not disclose data for all schools, which meant that several students in the sample would have been excluded.
3. There were in effect two rating occasions and scores were scaled for each occasion. To make scaled scores from each occasion comparable, we used scaled scores from the fifty texts as 'anchor values' when scaling scores from the second rating occasion.
4. This is made possible since the MFRM is robust against missing data, which means that if the data fits the model well, a student can get an estimated value on combinations of items, tasks and raters s/he did not encounter empirically.
5. Please note that the project reported on in Skar et al. ([in press](#)) generated more cut scores than were eventually published. All cut score statistics are available upon request to the first author.
6. Due to logistic reasons, 77 students were administered one of the two functional writing tasks after the first data collection period, and four students were administered the test a few days after the second data collection. Two students were administered the writing tasks late at both data collection occasions. Also, 53 students were administered the copy task after the first data collection period and two students were administered the test a few days after the second data collection. Rather than excluding these students from the analysis, however, we chose to include 'late administration' as a covariate in a preliminary analysis. This covariate was not significant, and thus all subsequent analyses were performed without this covariate.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the The Research Council of Norway [288795].

## Notes on contributors

*Gustaf B. Skar*, Ph.D., is Professor of Language Arts Education.

*Alan Huebner*, Ph.D., is Teaching Professor of Applied and Computational Mathematics and Statistics.

## ORCID

Gustaf B. Skar  <http://orcid.org/0000-0002-6486-396X>

Alan Huebner  <http://orcid.org/0000-0003-2141-5793>

## References

- Bae, J., & Lee, Y.-S. (2012). Evaluating the development of children's writing Ability in an EFL context. *Language Assessment Quarterly*, 9(4), 348–374. <https://doi.org/10.1080/15434303.2012.721424>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Berge, K. L., Evensen, L. S., & Thygesen, R. (2016). The wheel of writing: A model of the writing domain for the teaching and assessing of writing as a key competency. *The Curriculum Journal*, 27(2), 172–189. <https://doi.org/10.1080/09585176.2015.1129980>
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31. <https://doi.org/10.1007/s11092-008-9068-5>
- Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, 32(1), 83–100. <https://doi.org/10.1177/0265532214542994>
- Browne, W., & Golalizadeh, M. (2021). MLPowSim. University of Bristol. <http://www.bristol.ac.uk/cmm/software/mlpowsim/>
- Campbell, K., Chen, Y.-J., Shenoy, S., & Cunningham, A. E. (2019). Preschool children's early writing: Repeated measures reveal growing but variable trajectories. *Reading and Writing*, 32(4), 939–961. <https://doi.org/10.1007/s11145-018-9893-y>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037//0033-2909.112.1.155>
- Drijbooms, E., Groen, M. A., & Verhoeven, L. (2017). How executive functions predict development in syntactic complexity of narrative writing in the upper elementary grades. *Reading and Writing*, 30(1), 209–231. <https://doi.org/10.1007/s11145-016-9670-8>
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.
- Gee, J. P. (2004). *Situated language and learning: A critique of traditional schooling*. Routledge.
- Graham, S. (1999). Handwriting and spelling instruction for students with learning disabilities: A review. *Learning Disability Quarterly*, 22(2), 78–98. <https://doi.org/10.2307/1511268>
- Graham, S. (2018). A revised Writer(s)-Within-Community model of writing. *Educational Psychologist*, 53(4), 258–279. <https://doi.org/10.1080/00461520.2018.1481406>
- Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, 89(1), 170–182. <https://doi.org/10.1037/0022-0663.89.1.170>
- Graham, S., Bollinger, A., Olson, C. B., D'Aoust, C., MacArthur, C. A., McCutchen, D., & Olinghouse, N. (2018). *Teaching elementary school students to be effective writers: A practice guide (Issues 2012–4058)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. <http://ies.ed.gov/ncee/>
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing. *The Elementary School Journal*, 115(4), 523–547. <https://doi.org/10.1086/681947>
- Graham, S., Hebert, M., Paige Sandbank, M., & Harris, K. R. (2016). Assessing the writing achievement of young struggling writers. *Learning Disability Quarterly*, 39(2), 72–82. <https://doi.org/10.1177/0731948714555019>
- Graham, S., McKeown, D., Kiuahara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, 104(4), 879–896. <https://doi.org/10.1037/a0029185>
- Graham, S., & Rijlaarsdam, G. (2016). Writing education around the globe: Introduction and call for a new global analysis. *Reading and Writing*, 29(5), 781–792. <https://doi.org/10.1007/s11145-016-9640-1>

- Graham, S., Skar, G. B., & Falk, D. Y. (2021). Teaching writing in the primary grades in Norway: A national survey. *Reading and Writing*, 34(2), 529–563. <https://doi.org/10.1007/s11145-020-10080-y>
- Graham, S., Weintraub, N., & Berninger, V. (2001). Which manuscript letters do primary grade children write legibly? *Journal of Educational Psychology*, 93(3), 488–497. <https://doi.org/10.1037/0022-0663.93.3.488>
- Harrell, E. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer.
- Harvill, L. M. (1991). Standard Error of Measurement. *Educational Measurement: Issues and Practice*, 10(2), 33–41. <https://doi.org/10.1111/j.1745-3992.1991.tb00195.x>
- Hooper, S. R., Roberts, J. E., Nelson, L., Zeisel, S., & Kasambira Fannin, D. (2010). Preschool predictors of narrative writing skills in elementary school children. *School Psychology Quarterly*, 25(1), 1–12. <http://dx.doi.org/10.1037/a0018329>
- Hsieh, Y. (2016). An exploratory study on Singaporean primary school students' development in Chinese writing. *The Asia-Pacific Education Researcher*, 25(4), 541–548. <https://doi.org/10.1007/s40299-016-0279-0>
- Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing*, 14(1), 3–24. <https://doi.org/10.1016/j.asw.2008.12.002>
- Jeffery, J. V., Elf, N., Skar, G. B., & Wilcox, K. C. (2018). Writing development and education standards in cross-national perspective. *Writing & Pedagogy*, 10(3), 333–370. <https://doi.org/10.1558/wap.34587>
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, 80(4), 437–447. <https://doi.org/10.1037/0022-0663.80.4.437>
- Kim, Y. G., Al Otaiba, S., & Wanzek, J. (2015). Kindergarten predictors of third grade writing. *Learning and Individual Differences*, 37, 27–37. <https://doi.org/10.1016/j.lindif.2014.11.009>
- Kim, Y. G., Puranik, C. S., & Al Otaiba, S. (2015a). Developmental trajectories of writing skills in first grade. *The Elementary School Journal*, 115(4), 594–613. <https://doi.org/10.1086/681971>
- Kim, Y. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and Writing*, 30(6), 1287–1310. <https://doi.org/10.1007/s11145-017-9724-6>
- Knowles, J. E., & Frederick, C. (2020). *merTools: Tools for analyzing mixed effect regression models* [R package version 0.5.2]. <https://cran.r-project.org/package=merTools>
- Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2019). A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology* 41(7), 922–947. <https://doi.org/10.1080/01443410.2019.1659939>
- Koster, M., Tribushinina, E., de Jong, P. F., & van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research. *Journal of Writing Research*, 7(2), 249–274. <https://doi.org/10.17239/jowr-2015.07.02.2>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Linacre, J. M. (2018). *Facets® (version 3.80.4). computer software*. Winsteps.com.
- Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, 6(1), 8. <https://doi.org/10.1186/s40536-018-0061-2>
- Monroe, M., & Sherman, E. E. (1966). *Group diagnostic reading aptitude and achievement test*. C. H. Nevins.
- Montanari, S., Simón-Cerejido, G., & Hartel, A. (2016). The development of writing skills in an Italian-English two-way immersion program: evidence from first through fifth grade. *International Multilingual Research Journal*, 10(1), 44–58. <https://doi.org/10.1080/19313152.2016.1118670>

- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. <http://www.corestandards.org/>
- Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice*, 28(3), 14–23. <https://doi.org/10.1111/j.1745-3992.2009.00150.x>
- Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: recent insights into theory, methodology and practices* (pp. 55–82). Brill. <https://doi.org/10.1163/9789004248489>
- Opplæringslova [The Education Act]. (1998). Act relating to Primary and Secondary Education and Training (the Education Act). *LOV-1998-07-17-61*. <https://lovdata.no/dokument/NL/lov/1998-07-17-61>
- Organisation for Economic Co-operation and Development. (2005). *The definition and selection of key competencies - Executive Summary*. <http://www.oecd.org/pisa/35070367.pdf>
- Purves, A. C. (1992). Conclusion. In A. C. Purves (Ed.), *The IEA study of written composition II: Education and performance in Fourteen Countries* (Vol. 2, pp. 199–203). Pergamon.
- Russell, D. R. (1997). Rethinking Genre in School and Society: An Activity Theory Analysis. *Written Communication*, 14(4), 504–554. <https://doi.org/10.1177/0741088397014004004>
- Skar, G. B., & Aasen, A. J. (2021). School writing in Norway: Fifteen years with writing as key competence. In J. V. Jeffery & J. M. Parr (Eds.), *International perspectives on writing curricula and development. A Cross-Case comparison* (pp. 192–216). Routledge.
- Skar, G. B., Aasen, A. J., & Jølle, L. (2020). Functional writing in the primary years: protocol for a mixed-methods writing intervention study. *Nordic Journal of Literacy Research*, 6(1), 201–216. <https://doi.org/10.23865/njlr.v6.2040>
- Skar, G. B., Jølle, L., & Aasen, A. J. (2020). Establishing scales to assess writing proficiency development in young learners. *Acta Didactica Norge*, 14(1), 1–30. <https://doi.org/10.5617/adno.7909>
- Skar, G. B., Kvistad, A. H., Johansen, M. B., Rijlaarsdam, G., & Aasen, A. J. (in press). *Identifying texts in the warning zone: Empirical foundation of a screening instrument to adapt early writing instruction. Writing & Pedagogy*.
- Skar, G. B., Lei, P.-W., Graham, S., Aasen, A. J., Johansen, M. B., & Kvistad, A. H. (2022). Handwriting fluency and the quality of primary grade students' writing. *Reading and Writing*, 35(2), 509–538. <https://doi.org/10.1007/s11145-021-10185-y>
- Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1570–1573). Wiley-Blackwell.
- Stobart, G. (2012). Validity in formative assessment. In J. Gardner (Ed.), *Assessment and Learning* (2nd ed., pp. 233–242). SAGE.
- Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology*, 68, 19–37. <https://doi.org/10.1016/j.jsp.2017.12.005>
- Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, 27, 11–23. <https://doi.org/10.1016/j.asw.2015.06.003>
- Zheng, Y., & Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from assessing writing (2000–2018). *Assessing Writing*, 42, 100421. <https://doi.org/10.1016/j.asw.2019.100421>