

Identifying texts in the Warning Zone: Empirical foundation of a screening instrument to adapt early writing instruction

Gustaf B. Skar^a, Anne H. Kvistad^b, Marita B. Johansen^c,
Gert Rijlaarsdam^d, and Arne Johannes Aasen^e

Abstract

This article addresses the basis for the development of the screening tool Norwegian Early Writers Signal (NEWS). The aim of the study was to develop a tool for teachers in grades 1–3 to identify student texts in ‘the Warning Zone’, i.e., texts that signal insufficient overall text quality associated with students in need of extra instructional support. Text norms were elicited from a panel of 14 experts in a standard-setting seminar. The standard-setting procedure was a benchmarking-like approach in which panelists chose texts that according to their judgement were in the Warning Zone. Additionally, in an online questionnaire, data on experts’ expectation growth pattern for eight text quality aspects in grades 1–3 were collected. Furthermore, student texts in the Warning Zone were marked and then included in the screening tool to concretize the norms, showing that texts in this zone can take several shapes. The article discusses what steps can be taken to further validate and implement the NEWS tool.

KEYWORDS: WRITING ASSESSMENT, ELEMENTARY SCHOOL, STANDARD SETTING, PROFICIENCY LEVELS

^a Department of Teacher Education, Norwegian University of Science and Technology
email: gustaf.b.skar@ntnu.no; ID: <https://orcid.org/0000-0002-6486-396X>

^b Department of Teacher Education, Norwegian University of Science and Technology
email: anne.h.kvistad@ntnu.no; ID: <https://orcid.org/0000-0003-0916-9294>

^c Department of Teacher Education, Norwegian University of Science and Technology
email: marita.b.johansen@ntnu.no; ID: <https://orcid.org/0000-0002-2574-6920>

^d Department of Teacher Education, Norwegian University of Science and Technology and University of Amsterdam, Research Institute of Child Development and Education
email: g.c.w.rijlaarsdam@uva.nl; ID: <https://orcid.org/0000-0002-2633-7336>

^e Department of Teacher Education, Norwegian University of Science and Technology
email: arne.j.aasen@ntnu.no; ID: <https://orcid.org/0000-0002-9153-2939>

Submitted: 2021-04-14 Accepted: 2022-01-31

Introduction

Developing writing skills while in elementary school is important because one's writing proficiency may have positive or negative implications for succeeding in the school system and in work life. Writing ability has other obvious positive implications, such as promoting communication with others, oneself, and in thinking and learning. Better writers also tend to be better readers (Graham & Harris, 2018).

Research evidence indicates that struggling students should be attended to sooner rather than later (Foorman et al., 1997). Studies have indicated that writing performance to some extent is predictable (Hooper et al., 2010; Juel, 1988; Skar & Huebner, 2022; Wilson, 2018; Wilson et al., 2016). Hooper and colleagues (2010) found that pre-school literacy measures, such as phonological processing, and 'writing concepts' (e.g., writing one's name) showed an average correlation with writing measures from the same students in third, fourth, and fifth grade of $r = .48$. They also found a preservation effect with the relative rankings of students being rather intact across years. Juel (1988) followed 54 children from first to fourth grade and found that writing measures in first grade predicted writing measures in fourth grade with a correlation of $r = .38$. Juel (1988) also found an increase of correlations as a function of grade levels. The measures for second and third grade correlated higher with measures from fourth grade at $r = .53$ and $r = .60$, respectively. Another example is provided by Wilson et al. (2016), who investigated different scoring methods (e.g., automated essay scoring, human text quality scoring, linguistic analysis) to predict at-risk student writers. A 'multivariate prediction model,' including both human scores and automatic scoring, correctly identified 93% of students at risk, although the instrument also wrongly identified 25% of students as at risk, when they were not. Skar and Huebner (2022) also identified a high proportion of students at risk (92%), albeit with a higher rate of false positives (43%). The predictability cited in these studies may be interpreted to mean that writing instruction may preserve differences among students. It points to a 'natural growth pattern' that indicates that over time, instruction contributed the same amount of growth to all students, *if* it contributed. Students that do not meet minimal requirements must be detected very early so that they can receive extra instructional support to accommodate their needs. At the same time, we must acknowledge research also indicates less than perfect predictions. This might partly be due to the fact that a measurement is not perfect – in particular, writing proficiency is a complex construct to measure reliably and validly – but also to different rates of growth and different rates of intervention effects for individual students; however, it may be that while predictability is not perfect, we

have been cautioned that identifying students in the Warning Zone in early grades is a prudent undertaking.

In 2018, by an amendment to the Education Act, the Norwegian government decided to grant all students at risk of ‘falling behind’ in mathematics, reading, and writing the right to extra resources (Opplæringslova [The Education Act], 1998, amended June 8th, 2018). The amendment implies that these students are identified in early grades; however, currently, teachers who wish to monitor students’ writing are referred to a mandatory *reading* test developed by the Directorate for Education and Training¹ that also includes two relevant writing subscales: letter writing and spelling. For grades 2 and 3, only the spelling scale is administered. Scores on the test are used for placement into one of two categories: ‘at or below level of concern’ and ‘above level of concern.’ The cut-off scores have been set such that one-fifth of the student population will fall into the first category, and as such, by default, is classified as ‘running risk of not developing the key competences’²; however, this instrument does not fit the national curricular requirements for the writing curriculum in grades 1–3. This curriculum requires students to write communicative texts, especially descriptive and narrative texts, and provides a definition of writing as a means of thinking and communication.

A major challenge with the amendment to the new law is the absence of national or regional standardized measures of writing proficiency. While there are the abovementioned national and obligatory tests for reading comprehension in grades 1–3 [*kartleggingsprøver*] and obligatory reading tests in grades 5, 8, and 9 [*nasjonale prøver*], teachers are referred to their own writing assessments and their own notions of what might constitute the performance of a student at risk of ‘falling behind.’ The consequence is an absence of a unified way to categorize writing proficiency as ‘falling behind’ or ‘not falling behind’ and the absence of unified methods for identifying students with performances further along the proficiency scale. Therefore, there are most likely significant differences within and between schools in how this classification is performed.

An approach to remedy this problem would be surveying the international research field and the testing market for existing writing tests and implementing them in Norway; however, this approach would fail to consider theoretical and empirical evidence highlighting the contextual nature of writing and fail to consider that what counts as writing proficiency differs between contexts. Several researchers (e.g., Berge et al., 2016; Graham, 2018; Russell, 1997) have provided explanations as to why writing proficiency is far from restricted to cognitive traits and that students’ writing is shaped by and may even contribute to shaping the literacy context the students are part of. Evensen (2002, p. 393), for example, commented

that even macro aspects of texts are shaped by culture; placing the ‘thesis towards the end of an argumentative piece,’ he noted, ‘may strike American readers as unexpected, but in Nordic argumentative conventions, building an argument towards a final conclusion is the default option.’ A similar remark was made by Graham and Rijlaarsdam (2016) in a special issue on writing across the globe. Previously, the first and thus far only international comparison of writing proficiency failed to produce a conclusive answer to the question regarding how students in different contexts perform. The failure was attributed to a lack of commonly shared norms for text quality across contexts (Purves, 1992). Recent comparative investigations of writing curricula in multiple national contexts indicated that more than shared norms may be lacking; studies showed different curricula and writing research traditions in the contexts examined (Jeffery et al., 2018; Jeffery & Parr, 2021). To ensure that teachers can base their screening on a contextually relevant basis, there is a need to develop a writing assessment tool to serve the aim of eliciting context-relevant information regarding writing proficiency.

There have been few—if any—*official* tools for monitoring students’ writing proficiency in grades 1–3 in Norway. For their research project, Berge et al. (2019) devised their own standards based on interviews with teachers and applied them to 265 texts from the third and sixth grade, respectively. The aim of another research project was to assess writing proficiency based on a nationally representative sample of students in grades 5 and 8. The results were integrated into a pedagogical tool that can be used by teachers, who could administer the same writing tasks within their classes to compare the attainment of their students with the national average (Skar, 2017). A more recent project resulted in developing writing assessments and rating scales for grades 1–3 (Skar, Aasen et al., 2020; Skar, Jølle, et al., 2020), which have been administered to over 4,900 students in grades 1–3 (Skar et al., 2021). This latter project is the most relevant regarding the new law, but neither this nor the other research projects have given indications of students’ performance falling into certain, clearly defined categories. In other words, they cannot be used to identify texts that indicate that a student is ‘falling behind.’

To aid teachers in students’ writing proficiency assessments under the amendment to the law, there is a need for a *signal system* that indicates that a text might represent writing of a student running the risk of falling behind. The aim of this study was to design such a pedagogical tool for Norway for grades 1–3.

Challenges of Defining Writing in the Warning Zone

There are some obstacles when trying to define a 'warning zone', or a zone in which texts indicate a signal insufficient overall text quality associated with students in need of extra instructional support. First, there may be multiple patterns of weakness in and between texts, and second, there may be multiple patterns among students.

Multiple Text Patterns When Comparing Weak Texts

Based on their understanding of the Norwegian curriculum, Skar et al. (2020) identified eight relevant rating scales (or text quality aspects) for assessing writing in grades 1–3. Any text will have strengths and weaknesses on these eight scales of text quality. These characteristics interact and compensate. A text might be weak in spelling but be quite good in relating relationally to the reader (audience awareness, for instance). Two texts, each with different scores on the eight text quality aspects, may be weak texts.

Multiple Patterns Among Students

A student's text is the phenomenon to be screened to identify a student in the Warning Zone. One text cannot indicate the competence of a student because the performance depends partly on the relation between the topic and the student. On an individual level, the relation between different texts is unpredictable (Rijlaarsdam et al., 2013; Verheyden, 2010; Verheyden et al., 2010). The Verheyden studies showed that some students scored much better on some text dimensions in a second text, and on other dimensions, they scored much worse. At the same time, for other students, other text quality aspects for growth were in play.

When we define the Warning Zone, it is a *warning zone for texts, as relevant agents evaluate the text as a whole as worrisome*. That is, *this text has features that make us aware that we must keep an eye on other texts of this student to decide whether s/he needs adapted instruction and practice*. This Warning Zone for texts should be defined by the norms teachers use to determine that a student text has traits associated with a warning zone, meaning there are text quality aspects that are associated with students who *may* run the risk of falling behind.

Whether a text can be identified as a 'text in the Warning Zone' is based on two premises:

1. A student's text does inform the evaluator about the student's writing proficiency to some extent.

2. Teachers and other experts of writing in grades 1–3 can distinguish between text quality features typically associated with struggling students and features typically associated with non-struggling students.

Aim of the Study

The aim of the study was to design a screening instrument that supports teachers in identifying texts of students in the first three grades as belonging to the Warning Zone: a text that indicates insufficient overall text quality that calls for extra instructional support for the student.

The screening instrument builds on curriculum-relevant tasks and a rating instrument – text quality aspects with scale descriptions – that were already available (Skar, Aasen, et al., 2020; Skar, Jølle, et al., 2020). What was needed was to add empirically grounded norms based on the rating instrument that reflect a theory of text and to identify text profiles in the Warning Zone. This addition required insight into teachers' and other experts' norms and their beliefs in patterns of progression. It also required texts that indicate the Warning Zone; texts representing the variations of text profiles that fall in the region under the cut-off score as established by a panel of experts (i.e., teachers and teacher educators).

We formulated four design principles for the instrument, which is called the *Norwegian Early Writers Signal* (NEWS):

1. *Practicality.* Practitioners must be able to relate texts written in their classes to the NEWS to identify whether the written text shows correspondences to the text that the NEWS presents as examples of Warning Zone texts. Therefore, the NEWS must provide teachers with a reference scale consisting of texts that represent variations of weak texts as results of writing prompts that are common in grades 1–3. Therefore, as a task category, we chose a descriptive narration in a functional context that includes various weak texts in the reference scale, as they represent a national sample from Skar et al. (2021).
2. *Variability in Norms.* The selected texts in the NEWS must represent the variation in the normative belief systems within the community of teachers and must represent possible shifts in these systems across grades. Therefore, we invited a group of relevant experts (i.e., teachers and teacher educators) to select texts that would represent the Warning Zone and to provide insight into how they weigh the eight text characteristics across the grades. The definition of 'in need of extra instructional support' should

be based on a normative system derived from their individual practice. Implication: the normative system may allow teachers to expect different growth rates for different text quality aspects, and the relevance of text quality aspects differs between grades.

3. *Content Relevance.* Texts in the reference scale of the NEWS must include qualities that represent the collective understanding of qualities that must be taught in the first three grades. The definition of writing should fit the national Norwegian curriculum for grades 1–3, which focuses on communicative and functional writing (Skar, Aasen, et al., 2020). Implication: multiple text qualities must be covered from letter knowledge to sentence construction to audience awareness (Skar, Jølle, et al., 2020). Therefore, we annotated the reference scale texts on eight text quality aspects, as they are included in the national curriculum documents.
4. *Validity.* Texts in the NEWS must represent the empirical variability of texts in the Warning Zone for grades 1–3. Therefore, we will use a sample from the Norwegian Reference Data Set, a national representative set of texts, scored by a sample 24 raters, two per text, with a reliability of .94. The instrument must take into account that within a text, different qualities can reach different standards. Implication: different benchmarks with different distributions of qualities (text profiles) may illustrate the Warning Zone.

Research Questions

We formulated three questions, of which the answers would each contribute to the NEWS screening instrument.

- RQ1: According to relevant experts in the field, what are the cut-off scores in a nationally representative set of texts for the early writing grades (grades 1–3) that indicate a need for extra instruction? Answers to this question may indicate in which region of the score distribution respondents experience the Warning Zone as well as whether the ‘normative jump’ (see below) per grade varies.
- RQ2: What is the expected growth pattern for eight text quality aspects in these grades according to experts in the field? This question may reveal whether the eight text quality aspects have the same or different weights, to which extent there is a difference in weight per grade and whether the growth pattern expected from grades 1 to 3 varies per text quality aspect.
- RQ3: What is the variation of text profiles of texts from the national sample that fall into the Warning Zone with respect to the

eight text qualities? We expect that texts that fall in this Warning Zone will vary in the way they realize the eight text quality aspects.

Methodology

To answer the research questions, we implemented three research phases (i.e., three sub-studies): identifying the Warning Zone norms of experts, identifying the norm of the growth pattern of the eight text quality aspects, and identifying text profiles that represent the Warning Zone.

We elicited the norms of experts on two occasions (sub-study 1 and 2): first in a standard setting seminar to mark the Warning Zone and then through an online questionnaire. From the data, we analyzed the variability within this group per grade and the variability in the growth curves for grades 1 to 2 and 2 to 3 for the eight text quality aspects under study. For sub-study 3, we selected texts from the nationally representative text base (Skar et al., 2021) that indicate that an early writer may need extra instructional support in grades 1, 2, or 3. The research team coded these selected texts for all eight text quality aspects, observing the variation in text profiles within the Warning Zone. These texts and our annotations were added as benchmark texts for the screening instrument.

Participants in Sub-Studies 1 and 2

The study included 14 carefully selected participants. Nine participants were teachers from eight different schools in one municipality in Norway. The remaining five participants represented L1 teacher education at four Norwegian universities. All but one teacher panelist (i.e., $n = 13$) chose to participate in Sub-Study 2.

The size of the panel severely limited any aims of national representativeness; however, regarding standard setting, relevance is a more suitable criterion (c.f., Cizek & Bunch, 2007). This group represented relevant background and experience; seven of the panelists had taught for over 16 years, and 11 of 14 panelists had taught for over 11 years. All panelists had experience with early literacy development, and they covered, as a group, the most pertinent areas of writing in first to third grades: teaching, writing development, and special education needs.

The teacher panelists were recruited by the aid of an administrative officer of the municipality, who sent a letter to the principals of all elementary schools in the municipality not currently participating in writing intervention projects with the university of first, second and fifth author ($N = 23$). These schools were spread out in the municipality. In the letter, the principals were asked to nominate one or two teachers to a panel with the

objective of setting standards for writing proficiency. The criteria for nomination were the following: ‘several years’ experience teaching in grades 1–3’ and ‘a demonstrated interest in teaching writing.’ Ten principals nominated 14 teachers to participate in the panel. While all 14 teachers agreed to participate, circumstances related to the COVID-19 pandemic prevented five from participating. The teacher participants (‘TPs’) served in grades one, two, or three. Three of them were educated as kindergarten and primary school teachers, while the remaining six teachers had teacher certificates based on a five-year teacher college education. Two teachers specialized in special education. Five teachers had over 16 years’ teaching experience, three teachers had experience in the range of 11–15 years, and one had 6–10 years of experience. All teacher participants were women, which does not reflect the workforce, although most teachers (74.9% in 2020) in Norwegian in primary and upper secondary school are women (please refer to online table 12282 at <https://www.ssb.no/en/statbank/table/12282>). The mean age of the TPs was 44.8 years ($SD = 9.0$).

The recruitment of the teacher education participants (‘EPs’) was based on the authors’ collective knowledge of competent EPs across Norway. Five researchers were approached, and all agreed to participate. All worked at teacher education institutions. One EP was an associate professor in early childhood literacy, and the remaining four EPs were assistant professors specializing in early childhood literacy. One of the latter had been involved in developing the national curriculum for the Language Arts subject, two of the latter had led numerous school and professional development projects related to early childhood literacy, and the remaining two were key researchers in a large-scale writing intervention including children in grades 1–2. Among these participants, three had a master’s degree, one had a doctorate, and one had a five-year teacher college degree. All EPs also had teaching experience in L1 instruction in primary and secondary education: one researcher had 1–5 years of experience, one had 6–10 years of experience, one had 11–15 years, and two had been teaching for over 16 years. One researcher participant was male. The mean age of the EPs was 43.6 years ($SD = 6.4$).

All panelists received compensation for their participation in the standard-setting seminar. TPs were awarded a gift set containing two books (one on writing instruction and a children’s book), a notebook, and a pen. EPs were awarded a voucher, which could be used in an online bookstore.

Sub-Study 1: Identifying Norms

There is a plethora of methods for eliciting norms. In test contexts, these are usually presented as methods for deriving ‘cut-off scores,’ or numerical

boundaries separating, for example, categories like ‘basic’ and ‘advanced’ (for a comprehensive review of methods, readers are referred to Cizek & Bunch, 2007). For this investigation, we used a ‘benchmark approach’ based on the work of Harsch and Kanistra (2020). This approach is suitable for establishing cut-off scores, or boundaries for writing assessments based on student texts from a single administration, which was fitting for the present investigation (see materials below). The benchmarking approach asks a judge to associate a single student text with predefined proficiency descriptions, such as ‘in need of extra support’ or ‘not in need of extra support’. The original benchmarking approach consists of two phases. First, judges associated benchmark texts with proficiency levels individually and independently. This phase results in individual scores. Second, there is a consensus phase in which panelists come to an agreement regarding which benchmark text best represents a given level.

In this investigation we used a variant of the benchmark approach. To answer the first research question teachers and teacher educators read completely masked texts and marked which texts represented the level of writing that signaled ‘in need of extra instruction’ per grade. To answer the second research question, we elicited norms about the growth pattern from grade 1–3 for eight text quality aspects—validated rating scale descriptors—via a questionnaire.

Materials

Panelists were provided with a selection of texts and rating scales (text quality aspects).

Texts

Examples of student writing were collected from a reference dataset (RDS), which comprised a nationally representative sample of student texts that had been collected recently (Skar et al., 2021). The RDS contained information about the writing proficiency of 4,950 students in grades 1–3, which was measured using students’ responses to an extended writing task in which students wrote a letter to researchers at the university in Trondheim informing the addressees of what the students enjoyed doing during recess time. These texts were blinded and then assessed by two raters per text (24 trained raters in total) on the eight rating scales (Skar, Jølle, et al., 2020). The rating scales, which had been validated to fit the Norwegian curricular context (Skar et al., 2020), were: audience awareness, organization of content, content relevance, vocabulary, language use (sentence construction), spelling, handwriting (legibility), and punctuation. These eight rating scales contained descriptors for five levels each, ranging from a score of 1

to 5. The ratings were fitted to a many-facet Rasch measurement model. The output was a single-scale compound score – a *Text Quality Score* – in the range of 1–5 for each text, with higher numbers indicating higher proficiency. This score represented the student’s performance score, the average score across the rating scales after accounting for rater differences, task difficulties and differences in rating scale difficulty.

From the RDS, we assembled a booklet containing 100 texts by randomly selecting texts that scored near integers. The sampling strategy was employed to ensure sufficient spread in the material. The texts were arranged in descending order so that the first page contained the highest-scoring text and the last page the lowest-scoring page. All texts were completely blinded: panelists had no access to the score of the text, grade, gender, or language background of the student.

Rating Scales

Next to the texts, we provided the panelists with the same eight rating scales for writing in grades 1–3 that had been used to mark texts in the RDS. The ratings scales (text quality aspects) are presented in Appendix A.

Procedures

The standard-setting seminar had the following pattern: first, two training sessions, second, a warning zone text identification session, which was the data collection session, and third, a handwriting warning zone identification session.

The objective of the first training session (30 minutes) was to set common grounds: to familiarize the panelists with the range of text quality and to share which mixtures of text qualities could point to texts in the Warning Zone. The latter were used to highlight to the panelists that text quality is a multi-faceted construct and that when selecting benchmark texts, the panelists should be cautious not to make ‘halo mistakes’ (Eckes, 2011), such as by neglecting adequate or inadequate audience awareness in a text with poor and good spelling, respectively.

The panelists then completed a first round of benchmarking. To prevent the ordering effect, the panelists were randomly assigned to a different sequence of work. Some began to identify texts that indicated a need for extra instructional support from grade 3, some from grade 2, and some from grade 1. The researchers were available for questions and comments. After the first round, which lasted 45 minutes, the panelists were allowed to share and discuss their choices to construct shared knowledge about the eight text quality aspects and the way they were visible in texts. The discussions, which were moderated by the research team, lasted for 75 minutes.

The third session was the data collection session (60 minutes). The panelists were instructed to repeat the procedures from the first round, identifying one text per grade that indicated a need for extra instructional support. The panelists were informed that the benchmarking from this second round was the one that the researchers were going to use. All benchmarks were recorded on a separate sheet, which was collected by the researchers.

Each panelist recorded one text per grade (i.e., three texts per panelist), and overall, the panel identified a total of 42 texts.

To enable an investigation on discriminant validity, in a fourth session, the panelists identified Warning Zone texts in terms of handwriting. For this task, they were given a second booklet, containing the same texts as the previous one, but it was ordered differently, with the text scoring highest on handwriting on the first page. Otherwise, the procedure was identical.

Defining Cut-off Scores for Overall Text Quality per Grade

To further identify texts that would serve as examples in the pedagogical tool, we created a datafile in which we linked chosen texts to Text Quality Scores. As a concrete example, consider the benchmarks chosen for first grade (See Appendix B).

The texts selected to represent the Warning Zone in first grade were associated with the following Text Quality Scores taken from the RDS: 2.95, 2.93, 2.9, 2.2, 2.19, 2.18, 2.13, 2.12, 2.12, 1.93, 1.41, 1.16, 1.16, and 1.16. Notably, there was considerable variation among the panelists. While three panelists chose the text that was scored 1.16, one panelist chose a text that was scored 2.95 on the same scale, running from 1–5. To address the variability in the panelists' choices, we defined outliers as $\bar{x} + 2SD$, where \bar{x} was the average score of texts chosen by the panelists and SD was the standard deviation. Continuing the first-grade example, the average was 2.0, and the standard deviation was 0.6, which meant that scores outside the range of 2.0 ± 0.6 were defined as outliers, in this case 2.90, 2.93, and 2.95. The ensuing adjusted mean (the 'trimmed mean') was 1.8 with a standard deviation of 0.5. Please refer to Table 2 for the observed and the adjusted means for all grades. As cut-off scores, the adjusted means were used, so for first grade the cut-off score was 1.8.

Quality of the Cut-off Scores

The variability among the panelists indicates that they differed substantially regarding their holistic judgement of benchmarks representing proficiency levels. This finding merits some attention because it indicates variability even among a group of experts, which in turn makes it plausible

to suspect that one would find a similar variation among teachers at large, which further indicates the need of the tool developed. However, we did investigate the quality of the cut-off scores in two ways: discriminant validity and convergent validity.

Discriminant Validity

In the national sample, the average correlation between handwriting scores and the other text quality scores was $r = .45$ ($SD = 0.06$), indicating that both scores represent (somewhat) different qualities. We may expect that the panelists were able to discriminate between these qualities as well. A non-correspondence would indicate that the panelists were able to separate the two constructs 'text quality' and 'handwriting', whereas correspondence would indicate the contrary.

For third grade, there was no overlap between texts benchmarked for handwriting and text quality. The non-overlap in first and second grades was 79%. Consider first grade, where panelists bookmarked nine unique texts (of 100) for handwriting and 11 (of 100) for text quality. Three of these texts were bookmarked both for text quality and handwriting, thus making six texts unique for handwriting and eight for text quality. The same proportions were observed for second grade. While it could not be ruled out that the proportions for first and second grade were the result of chance, we interpreted the large non-overlap (79%) as an indication that the panel was able to distinguish between the two constructs. Table 1 reports number of texts and the number of overlapping texts.

Table 1. Teachers' choices of texts representing text quality and handwriting

Grade	Different texts chosen		Duplicates	HW + TQ No duplicates	Proportion non-overlap
	HW	TQ			
1st	9	11	3	14	0.79
2nd	10	10	3	14	0.79
3rd	12	13	0	25	1.00

Note: HW = handwriting, TQ = text quality.

Convergent validity

We investigated the increments between cut-off scores. It was decided that it would be satisfactory if cut-off scores for the different proficiency levels were ordered between grades, so that cut-off scores for the at-risk of falling behind level would increase as a function of grade. This was found to be true (Table 2, results section).

Sub-Study 2: Identifying Growth Pattern Beliefs

To identify the beliefs that panelists held about the progression per grade per text quality aspect, the panelists were invited to participate in a follow-up study, filling in a short questionnaire anonymously. They individually indicated per grade, for each of the eight text quality aspects separately, which of the five quality descriptors (Appendix A) would indicate best that the text would fall in the Warning Zone. They had no access to the scores generated in the national assessment study and no access to the student texts they had worked with during the standard-setting seminar. Each panelist ($N=13$) generated 24 scores: 8 (text quality aspect descriptors) \times 3 (grades).

As previously stated, some of the variability was addressed by the investigation and deletion of outliers. In this case, an outlier was defined as outside the 95% confidence interval of the mean score. The outlier analysis resulted in the detection of 18 outlier scores, or 5.7%, of the 312 score set. The outliers were not randomly distributed: they were observed for Organization of content (grades 1 and 2, $n = 7$), Audience awareness (grade 3, $n = 4$), Vocabulary (grade 3, $n = 4$), Punctuation (grade 1, $n = 2$), and Content relevance (grade 3, $n = 1$). Most of the outliers were found in grade 3 ($n = 9$), and fewer were found in grades 1 ($n = 4$) and 2 ($n = 5$). Not all panelists contributed to the proportion of outliers: one panelist generated five outliers, one three, two scored two, five just one, and three scored no outliers. We kept all scores in the analyses; this variation would reflect the variation in practice. Due to the small sample, we set a significance level of $p < .10$, being aware of a small risk of Type II errors.

Sub-Study 3: Identifying Text Profiles in the Warning Zone

We used the norms elicited through the standard setting seminar (Sub-Study 1) to identify texts from the reference data set (RDS) in the range of , where was the adjusted mean score of bookmarked texts in grade i , and SE was the standard error associated with that mean. As an example, for first grade, the adjusted mean score was 1.8, and the $SE = 0.14$. Please refer to Table 2 for all statistics.

For each grade, we selected three texts to represent texts that had features associated with the writing of students in need of extra attention. These were the first texts in the RDS that had a matching score to . For first grade, the three texts chosen had a *Text Quality Score* of: 1.5, 1.8, and 2.1. All texts were different texts than the ones selected by the panelists.

In total, nine texts were selected. These texts were marked by the first, second, and third author, applying the five-point scales for each of the eight

text aspects. After individual scoring, a conference followed in which raters compared scores and resolved any differences. The ensuing result was plotted in a matrix adjacent to the student text to ensure a visual pedagogical aid.

Results

Sub-Study 1: Which Text Profiles Represent Texts that Indicate a Need for Extra Instructional Support in Grades 1–3 According to Relevant Experts?

The standard setting seminar resulted in three cut-off scores, indicating students in need of extra attention: 1.8, 2.8, and 3.1 for 1st, 2nd, and 3rd grades, respectively (Table 2).

As is evident, the difference between first and second grade was wider (one scale step) than between second and third grade. The relatively larger leap between first and second grade indicates that the panelists perceived the Warning Zone as quite different for first and second grades compared to second and third grades, where the Warning Zone was more overlapping.

Table 2. Warning Zone norms: Cut-off scores based on panelists' text selection

Grade	With Outliers		Without Outliers	
	Mean	SE	Mean	SE
1 st	2.0	0.17	1.8	0.14
2 nd	2.8	0.10	2.8	0.11
3 rd	3.2	0.18	3.1	0.19

Sub-Study 2: What is the Expected Growth Pattern for the Eight Text Qualities in these Grades According to Experts in the Field?

The online questionnaire was used to investigate whether teachers weighted the eight aspects differently: what did they rate as more important? Moreover, did the internal system of text quality vary across grades? Did the panelists rate different text qualities as more important in one or the other grade?

The levels chosen varied (see Appendix A for the levels), as did the variation. For first grade, one of the panelists chose level 1 for all eight rating scales, other panelists combined levels 1 and 2, and one panelist chose levels 2 and 3. This variation continued for grades 2 and 3. For grade 2: Two panelists combined levels 1 and 2 across the rating scales, with the most weight on level 2; three panelists chose rating scale scores on levels 2 and

3, with different in weights on level 2 or 3; and two panelists had scores for some aspects on level 4. Grade 3 showed even more variation across the levels, with level 5 next to scores on levels 3 and 4. Please refer to Table 3 for the descriptive statistics. The variation showed that very few texts within each of the three Warning Zones would be unidimensional. This means that not one or two text quality aspects or qualities must be considered but all.

Table 3. Warning Zone norms: Panelists' mean scores and standard errors per grade and text aspect

Text Quality	Grade 1		Grade 2		Grade 3	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Audience Awareness	1.54	0.22	2.38	0.18	3.08	0.18
Organization of Content	1.23	0.12	2.00	0.20	2.54	0.18
Content Relevance	1.38	0.14	2.08	0.24	2.69	0.24
Vocabulary	1.77	0.20	2.54	0.22	3.00	0.23
Sentence Construction	1.38	0.18	2.31	0.17	2.62	0.18
Punctuation	1.23	0.12	1.92	0.21	2.85	0.19
Spelling	1.77	0.20	2.62	0.18	3.23	0.23
Handwriting (Legibility)	1.62	0.21	2.54	0.22	3.31	0.31

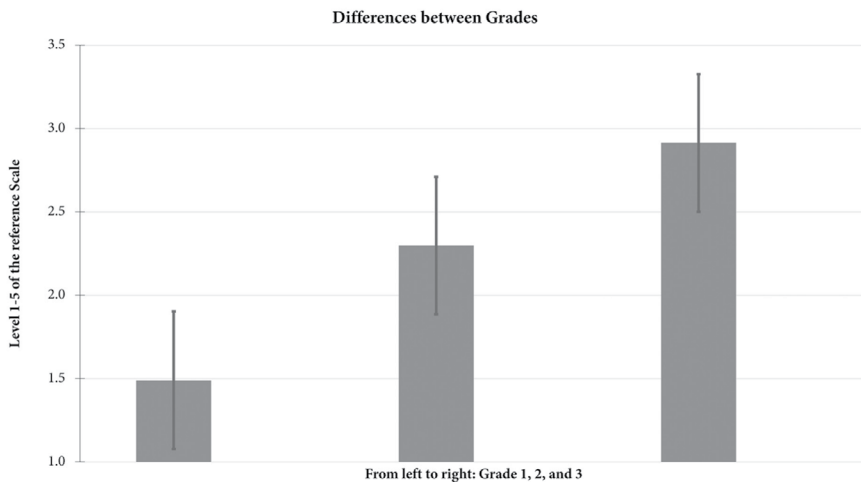


Figure 1. Grade effects: Mean norm scores and confidence intervals per grade according to the panelists.

To assess whether the relevance per grade level for aspects changed in the opinion of panelists, we ran a multivariate analysis of variance with repeated measures with the three grades and the eight text quality aspects as within variables. Two main effects and no interaction were observed. The multivariate effect of text qualities was significant ($F(7,68) = 5,795, p < .001, \eta^2 = .326$) as well as the effect of grade ($F(2,68) = 168,367, p < .001, \eta^2 = .933$). An interaction between the eight aspects and three grades was not observed ($F(14, 168) = 1.117, p = .29, \eta^2 = .089$). Figure 1 presents the effects for grades, and Figure 2 show the effects of the text quality aspects.

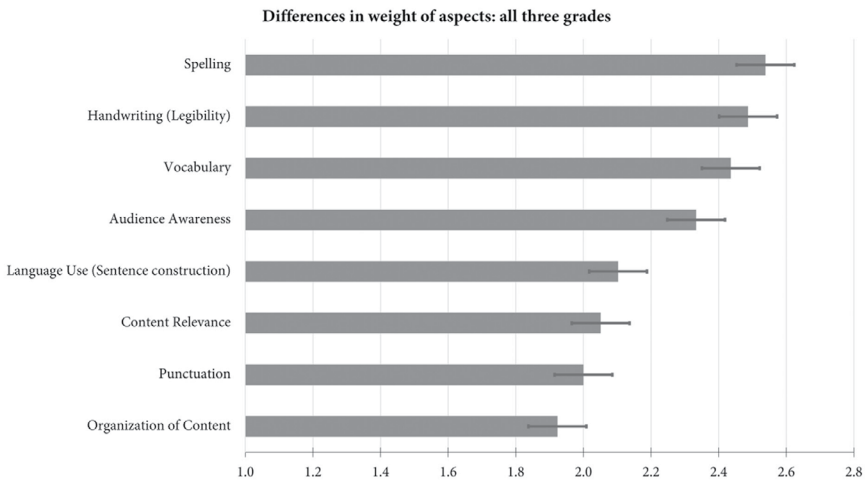


Figure 2. Effects of text qualities. Mean norm scores and confidence intervals per text quality aspect according to panelists aggregated across grades.

Based on Figure 2, it seems that there is a distinct norm for two sets of text quality aspects, each containing four text aspects. Indeed, the difference between the fourth and the fifth aspect, Audience Awareness and Sentence Construction, is statistically significant (p -value = .08). Within the two clusters, no statistically significant difference was found. These results imply that one expects higher achievement for the upper panel cluster – Spelling, Letter Knowledge, Vocabulary, and Audience Awareness – than for the lower panel cluster – Sentence Construction, Relevant Content, Punctuation, and Organization of Content.

Figure 3 presents the development per text quality aspects per grade. Although the analysis did not report an interaction between grade and aspect, Figure 3 shows that for some text quality aspects, the normed growth is not linear, according to the norm of the panelists. The requirements for Sentence Construction in grade 2, for instance, differ less from

grade 3 than from grade 1. For Sentence Construction, the effect size for one grade was 1.4 between grades 1 and 2, and .48 for grades 2–3.

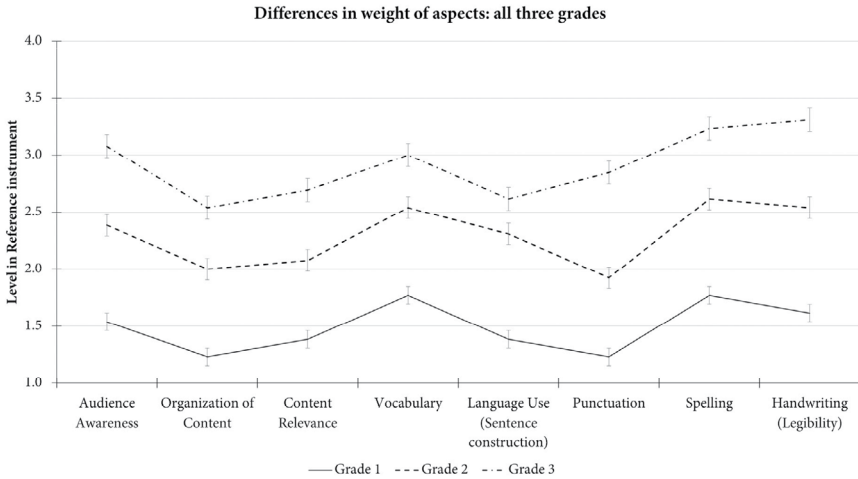


Figure 3. Weight of text qualities. Variation across grades and qualities. Mean scores and confidence intervals per grade for the eight text quality aspects according to the panelists.

Table 4 presents the effect sizes per grade per Text Quality. These effect sizes represent the ‘normative jump’ per grade per text quality aspect and the expected growth pattern. The mean “jump” from grade 1 to 2 is about 1.20, and from grade 2 to 3, it is about .80. Individual text aspects show a somewhat different growth pattern. Audience Awareness shows a relative linear pattern (1.19 + 1.07), Sentence Construction a steeper decline (1.44 + 0.48), and the pattern for Punctuation is somewhat upward (1.16 + 1.27).

Table 4. Effect sizes or ‘normative jumps’

	Grade 1 to 2	Grade 2 to 3
Audience Awareness	-1.19	-1.07
Organization of Content	-1.34	-0.79
Extent of Relevant Content	-1.01	-0.72
Vocabulary	-1.02	-0.58
Sentence construction	-1.44	-0.48
Punctuation	-1.16	-1.27
Spelling	-1.23	-0.83
Handwriting (Legibility)	-1.20	-0.82

Sub-Study 3: What is the Variation in the Text Profiles of Texts from the National Sample that Fall into the Warning Zone with Respect to the Eight Text Quality Aspects?

To create benchmark texts representing the Warning Zone per grade, a new set of texts was selected. Three texts from each grade were marked, which resulted in the profiles partly presented below and presented in full in Appendix B. The marked texts concretize the norms elicited through the standard setting seminar and the online questionnaire. Table 5 shows the scores (levels) of all nine texts for all eight text quality aspects.

Table 5. Text profiles as score levels for texts marked by researchers.

Grade	Text	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	Scale 6	Scale 7	Scale 8
G1	T1	2	1	1	2	1	1	2	2
	T2	2	1	1	2	1	1	2	3
	T3	3	2	2	2	2	1	2	2
G2	T1	3	3	2	3	3	1	3	3
	T2	3	2	2	3	3	1	3	3
	T3	3	3	3	3	3	3	3	2
G3	T1	3	3	2	3	2	2	3	4
	T2	4	3	2	3	2	2	3	3
	T3	3	2	4	3	3	4	3	4

Note: Scale 1 = Audience Awareness, Scale 2 = Organization of content, Scale 3 = Relevance, Scale 4 = Vocabulary, Scale 5 = Sentence construction, Scale 6 = Punctuation, Scale 7 = Spelling, Scale 8 = Handwriting.

To illustrate the variation of text profiles per grade, we present and discuss the three texts that represent the Warning Zone in grade 1 (texts for second and third grades are presented in Appendix B). Figures 4–6 present abridged versions of the text, where drawings have been left out (the full text can be found in Appendix B). Tables 6–8 present the marking of the texts and the norms elicited in Sub-Study 2.

The translation of text 1 is: ‘Football I I like swing’. It has received markings on level 1 and level 2 (see Table 6). Looking closer at the text, one can see that the text consists of individual words that make sense in interaction with each other, which matches the description for level 2 of audience awareness (see Appendix A). This marking does require some generous interpretation, though. The latter part of the text can be interpreted to read ‘Jeg liker disse’ (I like to swing), if ‘Lei’ and ‘LiK’ are interpreted to be

attempts at writing 'jeg' (I) and 'liker' (*like*). The current spelling ('Lei' and 'LiK'), however, gives the words for *sad* and *corps*, respectively. The organization matches the descriptor for level 1 (a text containing individual words), and the amount of relevant text also matches the descriptor for level 1 (a sentence or less). The vocabulary matches level 2 because the texts consist of different words. There is no punctuation. Spelling reaches level 2 because the text contains letter combinations and single words. The word 'Jeg' is spelled phonetically (level 3), with a rotated 'L'. The letters are decipherable but are not crafted in a conventional manner.

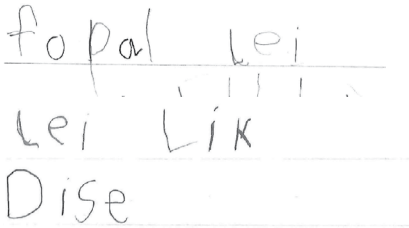


Figure 4. Text 1 (1st Grade): I like to swing

Table 6. Scoring profile text 1

	Grade 1 norms	Level 1	Level 2	Level 3
Audience Awareness	1.54			
Organization of content	1.23			
Relevance	1.38			
Vocabulary	1.77			
Sentence construction	1.38			
Punctuation	1.23			
Spelling	1.77			
Handwriting	1.62			

Note: 'Grade 1 norms' are the norms yielded from Sub-Study 2.

Text 2 is translated to: 'Jump jump rope swing'. The text consists of individual words that make sense in interaction with each other, which matches the descriptor for level 2 on the audience awareness scale. As with text 1, this judgement is based on a generous interpretation of what the words are attempts at. For 'håpe' to be 'jump', it should have read 'hoppe', and for 'håetou' to be 'jump rope', it should have read 'hoppetau'. Finally, for 'dis' to be 'swing', it should have read 'disse'. The organization and relevance, sentence construction, punctuation, and spelling are also very similar to text

1. The handwriting was marked as level 3 because the letters were crafted in a conventional manner (with the exception of 'd' in 'dis').



Figure 5. Text 2 (1st Grade): Jump rope

Table 7. Scoring profile text 2

	Grade 1 norms	Level 1	Level 2	Level 3
Audience Awareness	1.54			
Organization of content	1.23			
Relevance	1.38			
Vocabulary	1.77			
Sentence construction	1.38			
Punctuation	1.23			
Spelling	1.77			
Handwriting	1.62			

Note: 'Grade 1 norms' are the norms yielded from Sub-study 2.

Text 3 is translated to 'I like and [i.e., to] play dodgeball dodgeball is that one has a ball'. The text consists of individual words that make sense in interaction with each other. The text also offers a rudimentary explanation of the game dodgeball, which can be interpreted as being an instantiation of addressing a reader (see Skar et al., 2022), matching the descriptor of audience awareness for level 3 (see Appendix A). There is an indication of text organization, matching level 2 of the descriptors. The relevant part of the text corresponds to two sentences, also matching level 2. There are a few different words, and there are two complete sentences (i.e., including a subject and a predicate), which both match level 2 of the descriptors. There is no punctuation (matching level 1), and the spelling matches level 2, as does the handwriting because some letters are formed in an unconventional manner. The repetition of 'stickball' (dodgeball) suggests an attempt to correct the first occurrence, which contains mirrored letters.

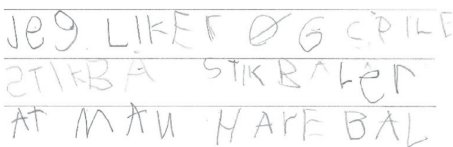


Figure 6. Text 3 (1st Grade): Dodgeball

Table 7. Scoring profile text 3

	Grade 1 norms	Level 1	Level 2	Level 3
Audience Awareness	1.54			
Organization of content	1.23			
Relevance	1.38			
Vocabulary	1.77			
Sentence construction	1.38			
Punctuation	1.23			
Spelling	1.77			
Handwriting	1.62			

Note: 'Grade 1 norms' are the norms yielded from Sub-study 2.

In sum, the following should be considered when reviewing the profiles. First, the profiles are 'uneven' in the sense that texts were not marked at any one level consistently; rather, the texts received marks across the levels, which confirms the investigation of teacher norms. Second, the profiles of texts in the Warning Zone indicate that texts in this zone can take several shapes, and while some may struggle with, for instance, audience awareness, others may struggle with spelling.

Discussion

This investigation was guided by three research questions, which for convenience are cited below. In this discussion, we will briefly discuss the results themselves, and more extensively, we will frame the results as part of a pedagogical instrument. The research questions were:

RQ1: What are the cut-off scores in a nationally representative set of texts for the early writing grades (Grades 1–3) according to relevant experts in the field that indicate a need for extra instruction?

RQ2: What is the expected growth pattern for eight text quality aspects in these grades according to experts in the field?

RQ3: What is the variation of text profiles of texts from the national sample that fall into the Warning Zone with respect to the eight Text Qualities?

Cut-off Scores

The investigation yielded three cut-off scores, which are numerical boundaries, for overall text quality that can be used for the estimation of the proportion of students in first, second, and third grade that write texts that indicate a writer in need of extra support. The cut-off scores were the result of a standard setting seminar in which 14 experts participated and supplied their norms by making holistic judgements. The method for establishing the cut-off scores was based on two premises: (1) a student's text does inform the evaluator about the student's writing proficiency to some extent and (2) teachers and other experts of writing in grades 1–3 can distinguish between text quality features typically associated with struggling students and features typically associated with non-struggling students. The first premise is of course debatable as we know that a single text will fail to reveal whether the struggle is permanent (cf., Verheyden, 2010) and to exhaustively disclose the nature of the struggle, which may be related to writing processes impossible to infer from the text. However, a teacher may counteract this information loss by administering several writing tasks, and if there are consistent indications that a particular student is struggling, a teacher can employ other specialized tools (e.g., for measuring fluency, spelling, and so on) to further investigate *why* the student seems to struggle.

The second premise was confirmed as the experts indeed were able to demarcate the boundaries sought after: the cut-off scores incremented across grades 1–3. The cut-off score for 1st grade was 1.8, the cut-off score for 2nd grade was 2.8, and that for 3rd grade was 3.1 on a scale running from 1–5, but the standard errors suggest that boundaries were overlapping, and that the method used indeed defined warning *zones* rather than definitive boundaries. Of course, the study could have been designed to yield even non-overlapping warning zones (e.g., by forcing consensus), but as stated in relation to the second design principle, we wished the results to represent a (or the) variation in the normative belief systems.

Figure 7 presents a visualization of the Warning Zones with the midpoint of an arrow indicating the cut-off score and arrow points indicating outer boundaries (based on $1SE \times \sqrt{N}$, with $N = 14$). While it might be unsettling to relate to a zone rather than a definitive value, we again stress that the tool should be used for multiple observations. For teachers with students consistently scoring within the grade level Warning Zone, the results are indeed a warning or a signal that these students may need extra support.

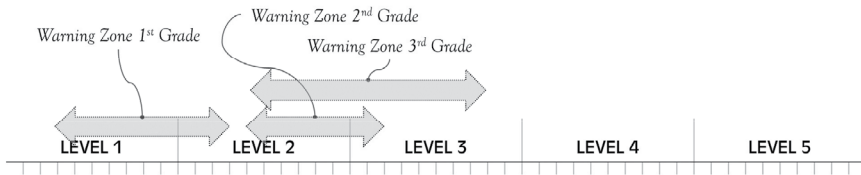


Figure 7. Warning Zones for grades 1–3

Expected Growth Pattern and Variation of Text Profiles

The results of RQ 2 and RQ3 provided further nuances to understand the nature of the Warning Zones for grades 1–3. To summarize the findings from RQ 2, norms differed indeed between and within grades, with a larger gap between grades 1 and 2 than grades 2 and 3. Norms were higher for Spelling, Handwriting (legibility), Vocabulary, and Audience Awareness than for Sentence construction, Content Relevance, Punctuation, and Organization of Content. Growth patterns for the eight Text Qualities that were distinguished varied.

The result of RQ3 presented different text profiles with grade Warning Zones on the rating scale level. Again, the non-resolute nature of the ensuing tool may be unsettling, but a way to construe it is to acknowledge that text-wise, a student can struggle in many ways. For teachers and other users of the tool, the key to collecting meaningful information would be to collect multiple pieces of evidence to look for patterns.

The Norwegian Early Writers Signal (NEWS)

The aim of the study was to design a screening instrument that supports teachers in identifying texts of students in the first three grades as belonging to the Warning Zone: a text that indicates insufficient overall text quality that calls for extra instructional support for the student. We formulated four design principles for such an instrument (please refer to the introduction), the *Norwegian Early Writers Signal (NEWS): Practicality, Variability in Norms, Content Relevance, and Validity*. Below we discuss to what extent the overall attainment of the goals implied by the design principles.

Practicality refers to the substance and useability of the instrument. The final instrument consists of four elements:

1. Descriptors of eight text qualities, representing a ‘theory of text’, each with five empirically based quality levels, reflecting the national curriculum in Norway (see Appendix A)

2. Cut-off scores for three grades to identify texts that might signal a need for extra instruction
3. Three texts per grade to indicate the variation of configurations of the eight text qualities in the Warning Zone
4. Annotations per text, explaining the eight text quality aspects per text

Together, these four elements – we believe – will provide practitioners with a tool that can be used for screening; however, as noted, its use will require work from a teacher, both in terms of the number of administrations of writing tasks and in terms of interpretation. In comparison with the current Norwegian situation with no screening tools, this work may be worthwhile.

Variability in Norms refers to the extent to which NEWS represents the variation in normative belief systems. This was determined by inviting a group of experts (teachers and teacher educators) to share their norms. The group was small ($N = 14$) in comparison to the whole population of teachers and teacher educators in Norway, so we do not claim that the norms elicited are representative for the whole population; however, the participants were carefully selected, and ultimately, the users of the NEWS must assess to what extent they will trust a tool based on the norms of participants with these characteristics.

Content Relevance refers to the principle that texts in the reference scale of the NEWS must represent qualities that represent the collective understanding of qualities that have to be taught in the first three grades. The measure taken to ensure this was the empirical grounding of the NEWS. Because the instrument must function in a specific context, with a specific national curriculum on literacy and a specific culture of beliefs of what good and weak texts entail in the teacher community and based on pre-service education and practice in schools, all empirical grounding refers to the Norwegian context.

Validity. Overall, we tend to trust the validity of the NEWS. The texts that indicate the Warning Zone are different constellations of the eight text quality aspects from the national curriculum. This is the variation, even in a small compartment of the text quality scale, that one may expect in practice. It is not one type of text and not one aspect of text that is representative of the Warning Zone. The text scale also validly represents the progress between grades 1–3, with more progress between grades 1 and 2 than between 2 and 3. A sign of validity is that the belief system of norms from Sub-Study 2 corresponds with the cut-off scores of the selected texts from Sub-Study 1, showing a larger jump from grade 1 to grade 2 than from grade 2 to 3, with some indication that the growth pattern, as expected,

may differ per text quality. Texts selected from the national database (Sub-Study 3) that met the criteria for belonging to the Warning Zone showed the variability in text profiles that that was expected from Sub-Study 2.

Implications for Screening

The stability of writing performance in general is low: research points to the need for ample evidence of student writing to reliably estimate the writing competence of a student at a given time (Schoonen, 2012; Van den Bergh et al., 2012; Verheyden et al., 2010) due to student-task interaction. The performance of a student depends strongly on the topic, as a function of topic interest and topic knowledge. This hinders the screening of students, and more performances are needed to indicate that students need extra instruction. This implies that students must produce quite a few texts in the same genre, instead of a wide variety of genres, and that students must have the opportunity to acquire topic knowledge before the writing task; however, as soon as a text of a student falls within the Warning Zone, more evidence should be sought. Further studies in the accurate and efficient use of the NEWS must be set up. Most importantly, a follow-up study could provide evidence on the concurrent validity of scores by correlating the writing scores from the assessment used in this investigation with standardized measures of writing, especially the Norwegian obligatory tests for reading comprehension that include two subscales relevant for writing (see above).

We also must investigate to what extent the type of texts selected – narrative descriptions – limited the generalization to other texts. Further investigations would need to show to what extent teachers can use the benchmark texts and norms to evaluate texts written for different topics and purposes to ensure that the cut-off scores are valid for other types of texts than the narrative descriptions used in this study.

A third extension is to study the consistency of performance in lower grades. We know that writing performance can vary due to task and genre. The question then is how many performances a teacher needs to decide that a student needs extra support. This might be a variable number, depending on the student. There is therefore considerable variability in the teacher's decision-making process. Here, the Body of Work, or other sophisticated ways for standard setting, might be helpful (Cizek & Bunch, 2007). Student portfolios (see Bay, 2012; for an example of an implementation) might become inevitable.

Implications for Instruction

Sub-Study 2 further shows that norms for Text Qualities vary: higher achievement on the scale from 1–5 is expected for Spelling, Letter knowledge, Vocabulary, and Audience Awareness than for Sentence Construction, Relevant Content, Punctuation, and Organization of Content. This finding, and the suggestion that the weight of text quality aspects varies across grades, deserves extended studies, with a larger group of respondents to increase power and precision. Such a pattern would imply that the needed attention in instruction moves from grade to grade. The issue, then, is whether this perception of needs reflects practice. Writing *instruction* is not quite common in the lower grades in Norway (Graham et al., 2021), and one may wonder whether explicit choices in what text quality aspects to instruct are made. Another issue for instruction is that teachers must deal with large variation of performances within a class. The national assessment study in Norway (Skar et al., 2021) reported that about 14.1 percent of the students in grade 1 scored already at or above the mean level of grade 2, and that, for instance, in grade 3, 25.2% of the students scored lower than the average of grade 2. This implies that a teacher must provide feedback on various levels of performance and internally set different goals for different students, such as based on the eight text qualities that we used for the NEWS.

Future Studies

Next, we must design and test guidelines for teachers and teams of teachers to use the NEWS in practice throughout the school year. Following these tests, when the pedagogic tool is implemented in practice, to serve the amendment to the law that students have a right to extra support, new studies should be conducted. Naturalistic studies should be conducted following students' performances during a time span of the first three grades, to determine whether they received instructional support and which type. This would provide us with insights regarding which teacher interventions are effective and efficient. Experimental studies researching the effects of teacher training in using the NEWS, with or without provided support for instruction, may reveal what kind of instructional support in different grades might be most effective. In all such studies, differentiation in instruction is implied.

Authors' Note

We have no conflicts of interest to disclose. Our work was funded by Grant #288795 from the Norwegian Research Council. Correspondence concerning this article should be addressed to Gustaf B. Skar, Department of Teacher Education, Norwegian University of Science and Technology, 7491 Trondheim, Norway. E-mail: gustaf.b.skar@ntnu.no

Notes

- 1 <https://www.udir.no/eksamen-og-prover/prover/kartlegging-gs/#137404>
- 2 <https://www.udir.no/eksamen-og-prover/prover/kartlegging-gs/#137404>

References

- Bay, L. (2012). *Developing achievement levels on the 2011 National Assessment of Educational Progress in grades 8 and 12 writing: Technical report*. <https://www.nagb.gov/content/dam/nagb/en/documents/publications/achievement/developing-achievement-levels-2011-naep-grade8-grade12-writing-technical-report.pdf>
- 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>
- Berge, K. L., Skar, G. B., Matre, S., Solheim, R., Evensen, L. S., Otnes, H., & Thygesen, R. (2019). Introducing teachers to new semiotic tools for writing instruction and writing assessment: Consequences for students' writing proficiency. *Assessment in Education: Principles, Policy and Practice*, 26(1), 6–25. <https://doi.org/10.1080/0969594X.2017.1330251>
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications.
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.
- Evensen, L. S. (2002). Convention from below: Negotiating interaction and culture in argumentative writing. *Written Communication*, 19(3), 382–413. <https://doi.org/10.1177/074108802237750>
- Foorman, B. R., Francis, D. J., Shaywitz, S. E., Shaywitz, B. A., & Fletcher, J. M. (1997). The case for early reading intervention. In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia* (pp. 243–264). Lawrence Erlbaum Associates.
- 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., & Harris, K. R. (2018). Evidence-based practices in writing. In S. Graham, C. A. MacArthur & M. Hebert (Eds.), *Best practices in writing instruction* (pp. 3–29). The Guilford Press.
- Graham, S., & Rijlaarsdam, G. (2016). Writing education around the globe: Introduction and call for a new global analysis. *Reading and Writing*, 29, 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>
- Harsch, C., & Kanistra, V. P. (2020). Using an innovative standard-setting approach to align integrated and independent writing tasks to the CEFR. *Language Assessment Quarterly*, 17(3), 262–281. <https://doi.org/10.1080/15434303.2020.1754828>
- 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. <https://doi.org/10.1037/a0018329>
- 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>
- Jeffery, J. V., & Parr, J. M. (Eds.). (2021). *International perspectives on writing curricula and development*. Routledge. <https://doi.org/10.4324/9781003051404>
- 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>
- Opplæringslova [The Education Act]. (1998). *Lov om grunnskolen og den vidaregående opplæringa ({LOV}-1998-07-17-61)*. Norwegian Ministry of Education and Research. <https://lovdata.no/lov/1998-07-17-61>
- 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.
- Rijlaarsdam, G., Jansen, T., Braaksma, M., Van Steendam, E., Van den Branden, K., & Verheyden, L. (2013). Learning and instruction in writing. In C. A. Stone, E. R. Silliman, B. J. Ehren & G. P. Wallach (Eds.), *Handbook of language and literacy*, Second Edition (pp. 545–566), Guilford Press.
- 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>
- Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In *Measuring writing: Recent insights into theory, methodology and practice* (pp. 1–22). Brill.
- Skar, G. B. (2017). *The Norwegian national sample-based writing test 2016: Technical report*. Nasjonalt senter for skriveopplæring og skriveforskning. <http://www.skrivesenteret.no/uploads/files/Skriveproven2017/NSBWT2017.pdf>

- Skar, G. B., Aasen, A. J., Kvistad, A. H., & Johansen, M. B. (2022). Audience awareness in elementary school students' texts: Variations within and between grades 1–3. *Writing & Pedagogy*, 13(1–3), 155–180.
- 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., & Huebner, A. (2022). *The predictability of first grade students' writing proficiency [Submitted]*, Department of Teacher Education, Norwegian University of Science and Technology & Department of Applied and Computational Mathematics and Statistics, University of Notre Dame.
- Skar, G. B., & Jølle, L. (2017). Teachers as raters: Investigation of a long-term writing assessment program. *L1*, 17(Open Issue), 1–30. <https://doi.org/10.17239/L1ESLL-2017.17.01.06>
- 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., 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, 509–538. <https://doi.org/10.1007/s11145-021-10185-y>
- Van den Bergh, H., De Maeyer, S., Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In *Measuring writing: Recent insights into theory, methodology and practice* (pp. 23–32). Brill.
- Verheyden, L. (2010). *Achter de lijn. Vier empirische studies over ontluikende stelvaardigheid*. [The story behind the line. Four empirical studies on writing by third and fourth graders of primary school.] Diss KU Leuven.
- Verheyden, L., Van den Branden, K., Rijlaarsdam, G., Van den Bergh, H., & De Maeyer, S. (2010). Written narrations by 8- to 10-year-old Turkish pupils in Flemish primary education: A follow-up of seven text features. *Journal of Research in Reading*, 33(1), 20–38. <https://doi.org/10.1111/j.1467-9817.2009.01430.x>
- 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>

APPENDIX A – RATING SCALES

	Level 1	Level 2	Level 3	Level 4	Level 5
Audience Awareness	To understand the text, a conversation with the writer is required.	The text contains words/ characters/ drawings that make sense in interaction with each other.	The text contains elements that indicate that the text addresses a reader.	The text addresses the reader in the assignment in a fairly relevant manner and takes into account to some extent the reader's need for knowledge of participants/ characters, circumstances, and events.	The text addresses the reader in the assignment in a generally relevant manner and takes into account the reader's need for knowledge of participants/ characters, circumstances, and events. The text may contain traces of the student's voice with reflective or evaluating utterances.
Organization of content	The text consists of individual letters/ words/ characters/ drawings.	The text may indicate a structure, such as in the form of a list with a marked thematic headline or letter structure. The additive connector "and" may appear.	The text has a global structure with elements arranged in a logical order. In some cases, the introduction or ending may not be explicit. The text contains primarily additive and temporal connectors (e.g. 'and', 'so'.	The text has a global structure with some elaborated elements arranged in a logical order. In some cases, the introduction or ending may not be explicit. The text may show examples of comparisons, classifications, chronology. The text includes different connectors (e.g., 'but', 'also', 'because').	The text has a complete global structure with several elaborated elements arranged in a logical or otherwise appropriate order. The text contains connectors that are used suitably and purposefully.

	Level 1	Level 2	Level 3	Level 4	Level 5
Content Relevance	The part of the verbal text that is a relevant answer to the task corresponds to a sentence or less.	The part of the verbal text that is a relevant answer to the task corresponds to approx. two to three sentences.	The part of the verbal text that is a relevant answer to the task corresponds to approx. half an A4 page (25–49 words).	The part of the verbal text that is a relevant answer to the task corresponds to approx. an A4 page (50–74 words).	The part of the verbal text that is a relevant answer to the task corresponds to approx. one and a half A4 pages or more (75+ words).
Vocabulary	The text consists of individual letters/ words/ characters/ drawings.	The text contains some few (different) words.	The text contains several different words (a lot of them theme-related).	The text contains a repertoire of words and expressions (a lot of them theme-related).	The text contains a repertoire of words and expressions (a lot of them theme-related). In some cases, there is use of specialized and abstract words and/ or creative forms of expression.
Language use (Sentence construction)	The text consists of individual letters/ words/ characters/ drawings.	There may be complete sentences.	The sentences show little variation in structure (in texts where variation is relevant).	Parts of the text shows appropriate variation in sentence structure.	The text has for the most part appropriate syntactic variation, and it has some developed phrases and/ or subordinate clauses.

	Level 1	Level 2	Level 3	Level 4	Level 5
Punctuation	The text has no punctuation.	Some punctuation can occur and/or there is exploratory use of punctuation.	Occurrences of functional use of punctuation (especially dot).	Mostly correct use of periods. Occurrences of functional use of question marks and/or exclamation marks (in texts where relevant).	Functional use of various forms of punctuations. The use of a comma may occur.
Spelling	There may be letters in the text and/or there may be scribbles (imitating writing).	The text contains letter combinations and single words.	The words are spelled phonetically, and some high-frequency words related to primary school students' sphere are written correctly.	There are examples of non-phonetic words that are correctly written. There may be examples of overgeneralization (for example, silent 'h' first in words starting with 'v' – hvært).	There are a number of examples of non-phonetic words written correctly.
Handwriting (legibility)	Letters may be difficult to decipher (if any).	The text contains letters that are not crafted in a conventional manner.	The letters are mainly crafted in a conventional manner.	The letters are crafted in a conventional manner. Instances of conventional use of the 'handwriting house'. [*] Occurrences of alternating use of upper- and lower-case letters.	The letters are drafted in a conventional and legible manner. For the most part, there is conventional use of the 'handwriting house'. Usually follows conventions for use of upper- and lower- case letters.

^{*}The 'handwriting house' denotes the relative position of letters. A lower-case 'g' and a lower-case 'h' will – if correctly written – be placed in the 'ground floor' and 'basement' (g) and in the 'ground floor' and 'addict' (h).

APPENDIX B – MARKED TEXTS

Marked texts: First grade

Text 1: (#952038 [original score: 1.51])

	Grade 1 norms	Level 1	Level 2	Level 3
Audience Awareness	1.54			
Organization of Content	1.23			
Relevance	1.38			
Vocabulary	1.77			
Sentence Construction	1.38			
Punctuation	1.23			
Spelling	1.77			
Handwriting	1.62			

Note: 'Grade 1 norms' are the norms yielded from Sub-Study 2.

Translation: Football I I like swing

Annotation: The text consists of individual words that make sense in interaction with each other, which matches the description for level 2 of Audience Awareness (see Appendix A). This marking does require some generous interpretation, though. The latter part of the text can be interpreted to read 'Jeg liker disse' (I like to swing) if 'Lei' and 'LiK' are interpreted to be attempts at writing 'jeg' (I) and 'liker' (*like*). The current spelling ('Lei' and 'LiK') gives the words for *sad* and *corps*, respectively. The organization matches the descriptor for level 1 (a text containing individual words), and the amount of relevant text also matches descriptor for level 1 (a sentence or less). The vocabulary matches level 2 because the text consists of different words. There is no punctuation. Spelling reaches level 2 because the text contains letter combinations and single words. The word "Jeg" is spelled phonetically (level 3) with a rotated 'L'. The letters are decipherable but not crafted in a conventional manner.

fopal lei



lei Lik

Dise



Marked texts: First grade

Text 2: (#851012 [original score: 1.80])

	Grade 1 norms	Level 1	Level 2	Level 3
Audience Awareness	1.54			
Organization of Content	1.23			
Relevance	1.38			
Vocabulary	1.77			
Sentence Construction	1.38			
Punctuation	1.23			
Spelling	1.77			
Handwriting	1.62			

Note: 'Grade 1 norms' are the norms yielded from Sub-Study 2.

Translation: Jump jump rope swing.

Annotation: The text consists of individual words that make sense in interaction with each other, which matches the descriptor for level 2 on the Audience Awareness scale. As with text 1, this judgement is based on a generous interpretation of what the words are attempts at. For 'håpe' to be 'jump', it should have read 'hoppe', and for 'håetøu' to be 'jump rope', it should have read 'hoppetau'. Finally, for 'dis' to be 'swing', it should have read 'disse'. The organization and relevance, sentence construction, punctuation, and spelling are also very similar to text 1. The handwriting was marked as level 3 because the letters were crafted in a conventional manner (with the exception of 'd' in 'dis').

HÅPE HÅETØUIS



Marked texts: First grade

Text 3: (#756006 [original score: 2.11])

	Grade 1 norms	Level 1	Level 2	Level 3
Audience Awareness	1.54			
Organization of Content	1.23			
Relevance	1.38			
Vocabulary	1.77			
Sentence Construction	1.38			
Punctuation	1.23			
Spelling	1.77			
Handwriting	1.62			

Note: 'Grade 1 norms' are the norms yielded from Sub-Study 2.

Translation: I like and [i.e., to] play dodgeball dodgeball is that one has a ball.

Annotation: The text consists of individual words that make sense in interaction with each other. The text also offers a rudimentary explanation of the game dodgeball, which can be interpreted as an instantiation of addressing a reader (see Skar et al., 2022), matching the descriptor of Audience Awareness for level 3 (see Appendix A). There is an indication of text organization, matching level 2 of the descriptors. The relevant part of the text corresponds to two sentences, also matching level 2. There are a few different words, and there are two complete sentences (i.e., including a subject and a predicate), which both match level 2 of the descriptors. There is no punctuation (matching level 1), and the spelling matches level 2, as does the handwriting because some letters are formed in an unconventional manner. Repetition of ‘stickball’ (dodgeball) suggests an attempt to erase the first occurrence, which contains mirrored letters.

JEG LIKER ØG SPILLE
STIKBALL STIKBALLER
AT MAN HAR BALL

Marked texts: Second grade

2nd Grade, Text 1: (#751008 [2.50])

	Grade 2 norms	Level 2	Level 3	Level 4
Audience Awareness	2.38			
Organization of Content	2.00			
Relevance	2.08			
Vocabulary	2.54			
Sentence Construction	2.31			
Punctuation	1.92	← LEVEL 1		
Spelling	2.62			
Handwriting	2.54			

Note: ‘Grade 2 norms’ are the norms yielded from Sub-Study 2.

Translation: I like to play in sandbox because I like to make castles I like free play From XXX.

Annotation: The text explains what the writer likes to do. It is short and not varied in structure. There is some logical organization and a closure. The frequent word 'liker' (like) is misspelled (i.e., 'likker'), but 'jeg' (I) is spelled correctly, and 'å' (to) was used as a mark following infinitive verbs. The text contains no punctuation.



Jeg Likker å Leke i Sankose
 fordi Jeg Likket å Lage slott
 Jeg Likker friLeik

Fra



Marked texts: Second grade2nd Grade, Text 2: (#602042 [original score: 2.80])

	Grade 2 norms	Level 2	Level 3	Level 4
Audience Awareness	2.38			
Organization of Content	2.00			
Relevance	2.08			
Vocabulary	2.54			
Sentence Construction	2.31			
Punctuation	1.92	← LEVEL 1		
Spelling	2.62			
Handwriting	2.54			

Note: 'Grade 2 norms' are the norms yielded from Sub-Study 2.

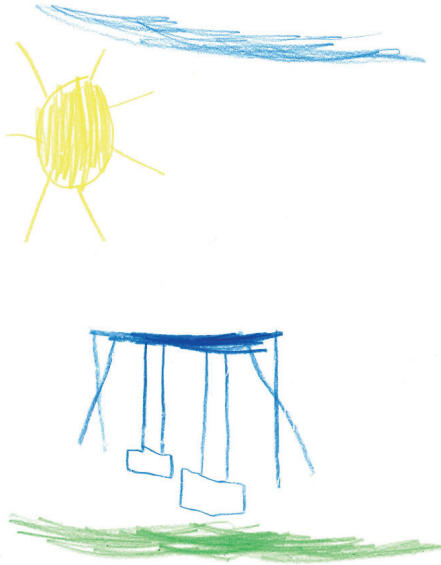
Translation: To the researchers I like to play football and the sand box and swing and jump swing and 'hussura' [tag]

Annotation: The text explains what the writer likes to do. It is short and not varied in structure. Some words are spelled phonetically, but common words, such as 'jeg' (I), are spelled orthographically. The non-use of punctuation matches level 1.

Til ForskernE

Jeg liker å spille

Fotball og sathkasa
 Å så disse og hopedise
 og HUS SURa



Marked texts: Second grade

2nd Grade, Text 3: (#701008 [original score: 3.00])

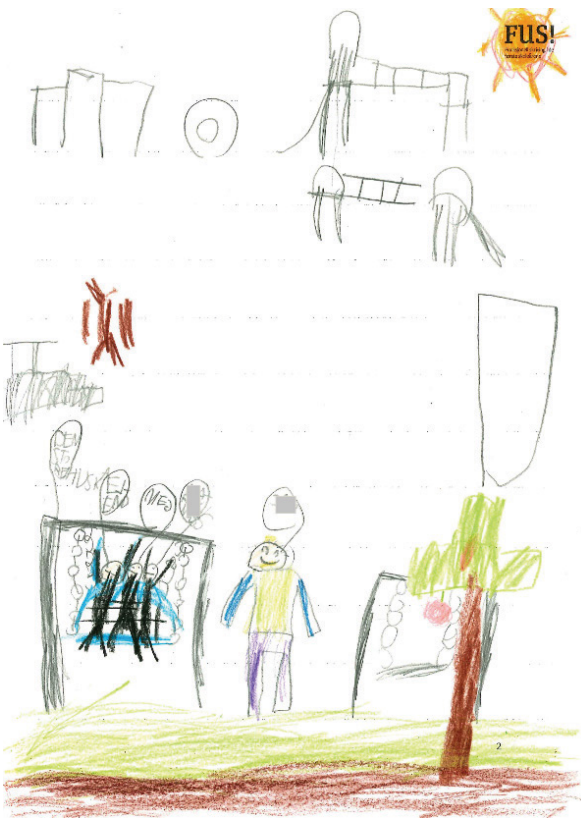
	Grade 2 norms	Level 2	Level 3	Level 4
Audience Awareness	2.38			
Organization of Content	2.00			
Relevance	2.08			
Vocabulary	2.54			
Sentence Construction	2.31			
Punctuation	1.92			
Spelling	2.62			
Handwriting	2.54			

Note: 'Grade 2 norms' are the norms yielded from Sub-Study 2.

Translation: I like to play with XXX and XXX and XXX. I like to play Minecraft. I like to play in the garden of apples. I like to play in the school yard. I like to play on the blue mat. I like to play Minecraft in those places.

Annotation: The text tells who the writer likes to play with, what his/her favorite play is (Minecraft), and where s/he likes to play Minecraft. The structure is somewhat opaque until the last sentence, where it becomes clear that the disclosure of places the writer likes to play are places to play Minecraft. All letters are capitalized, and one letter is consistency mirrored (G), while one (L) is mirrored from the start of the third row.

JEG LIKER Å LEKE MED
 O O O
 O O O? JEG LIKER Å
 JEKE MÅNEDRIFT JEG JIKER Å LEKE
 I EPLEHAJEN JEG JIKER Å
 JEKE I SKOJEÅREN
 JEG JIKER Å JEKE PÅ
 DEN BLÅ MATEN PÅ DE STE-
 DENE JIKER JEG Å JEKE MÅNEDRIFT



Marked texts: Third grade3rd Grade, Text 1: (#525006 [original score: 2.70])

	Grade 3 norms	Level 2	Level 3	Level 4
Audience Awareness	3.08			
Organization of Content	2.54			
Relevance	2.69			
Vocabulary	3.00			
Sentence Construction	2.62			
Punctuation	2.85			
Spelling	3.23			
Handwriting	3.31			

Note: 'Grade 3 norms' are the norms yielded from Sub-Study 2.

Translation: God-day researchers I like to chop trees. And branches. And makes cabin. Regards XXX

Annotation: As an introduction, the text addresses the recipients, and the closure includes a 'regards.' The writer uses full stops with some accuracy. The words are spelled orthographically, but these are high-frequency words, such as 'I.' Mostly, the writer has positioned the letters correctly in relation to each other (hence level 4).

God-dag forskere Jeg liker å sage ned tær.
 OG kvister. Og Lager hytte.
 Hilsen [REDACTED]

Marked texts: Third grade3rd Grade, Text 2: (#575035 [original score: 3.10])

	Grade 3 norms	Level 2	Level 3	Level 4
Audience Awareness	3.08			
Organization of Content	2.54			
Relevance	2.69			
Vocabulary	3.00			
Sentence Construction	2.62			
Punctuation	2.85			
Spelling	3.23			
Handwriting	3.31			

Note: 'Grade 3 norms' are the norms yielded from Sub-Study 2.

Translation: Hi researcher play Pokémon with my friends we fight. I often win regards XXX

Annotation: The text has a letter structure with a greeting and a closure. The sentences are not complete but consist of sentence fragments. Some of the words are spelled correctly, but Pokémon, the only non-high frequent word, is misspelled ('pokmon').

HEI FORSKER
 LEIKE POKMON MED
 VENNENE MINE
 VI KEMPER
 JEG VINNER ÅLE
 HILSEN

Marked texts: Third grade3rd Grade, Text 3: (#709024 [original score: 3.50])

	Grade 3 norms	Level 2	Level 3	Level 4
Audience Awareness	3.08			
Organization of content	2.54			
Relevance	2.69			
Vocabulary	3.00			
Sentence Construction	2.62			
Punctuation	2.85			
Spelling	3.23			
Handwriting	3.31			

Note: 'Grade 3 norms' are the norms yielded from Sub-Study 2.

Translation: I became friend with XXX. Now we are 4 including XXX and XXX. I love to have new friends. My mother said I was good because I got a new friend. I love recess time. The recess times is fun. I love to play in the climbing stand. It is so so so so fun.

Annotation: The text is long but somewhat off-topic as it includes making friends and mothers appraisal. The structure is opaque. The writer uses full stops correctly. Most words are spelled correctly, but they are high-frequency words. Letters are conventional and positioned correctly in relation to each other.

Jeg ble ven med
Tanishka er
vi 4 er sammen
med [redacted] og [redacted]
Jeg elsker å få
nye venner. Mamma
men så at jeg
flek forseg jeg har
tot en ny ven.

Jeg elsker begge
fri minutterne.

Det er gøy på
fri minutterne.

Jeg elsker og
bæli klatter-

Satymet. Det er
så så så så gøy.