Gustaf Bernhard Skar
Norges teknisk-naturvitenskapelige universitet (NTNU)
Lennart Jølle
Norges teknisk-naturvitenskapelige universitet (NTNU)
Arne Johannes Aasen
Norges teknisk-naturvitenskapelige universitet (NTNU)

DOI: http://dx.doi.org/10.5617/adno.7909

Establishing Rating Scales to Assess Writing Proficiency Development in Young Learners

Abstract
Writing assessment scales were developed to include functional aspects of writing proficiency in contemporary Norwegian teaching toolkits for Grades 1 to 3. This study aims to describe the process of developing empirically based, assessor-oriented writing proficiency scales and of investigating the quality of the scales. We focus on psychometric qualities, professional users' perceptions of their quality, and the teachers' use of the scales. Overall, the first piloted version of the scales showed indications of well-functioning scales. The results from this investigation show that it is possible to develop scales for the assessment of young children's writing proficiency that capture the intended construct and provide a basis for reliable assessment. The investigation also found that users of the assessment tool found that it functioned well.

Keywords: writing assessment, writing development, rating scales, reliability

Upprätta bedömningsskalor för att utvärdera utvecklingen av skrivförmågan hos unga studenter

Sammanfattning
Denna artikel berör arbetet med att utveckla bedömningsskalor för bedömning av funktionell skrivförmåga i årskurs 1–3. Syftet med undersökningen var att dels beskriva arbetet med att skapa empiriskt baserade bedömar-orienterade skalor, dels att undersöka kvaliteten på dessa skalor. Det senare gjordes genom att studera skalorna psykometriska kvalitet, den kvalitet de uppfattades att ha samt lärares användning av skalorna. Sammantaget visade undersökningarna att skalorna fungerade väl, redan vid första utkast. Vidare indikerade resultaten att det var möjligt att konstruera skalor för bedömning av små barns skrivande som kunder generera reliabel bedömning samtidigt de fångade väsentliga aspekter av funktionellt skrivande. Undersökningarna indikerade också att användare av bedömningsverktygen generellt uppfattade dem som välfungerande.

Nyckelord: skrivbedömning, skrivutveckling, bedömningsskalor, reliabilitet

Introduction

Writing instruction in Norwegian schools begins in first grade. The Norwegian Parliament passed a bill in 2018 to hold schools accountable for helping students in Grades 1 to 4 who are at risk of "being left behind" in terms of writing development; however, no tool for assessing writing proficiency in Grades 1 to 3 is available. There are tools to identify students' letter knowledge (Norwegian Reading Center, 2018) and competence with coding words (e.g., Carlsten, 2016), but there are no tools for the assessment of writing proficiency. In contrast, in the United States (US), WIDA (e.g., 2017)— a similar "no child left behind" act— has for years provided the teaching community with assessment protocols for the evaluating of writing proficiency.

Unfortunately, adopting a similar system in a new context is not as simple as translating writing assessment resources from other contexts. Current thinking and years of empirical evidence suggest that writing proficiency and writing development are contextual, and that resources therefore need to be adapted to the particular assessment context (Camp, 2012; Jeffery et al., 2018; Purves, 1992a; Slomp, 2012).

According to a recent writing intervention project, writing proficiency can be thought of in terms of functional competencies and coding competencies (Skar et al., 2017). Functional competencies are those that a writer uses to adapt a text to a given communicative situation. These competencies include writing the text to fulfil a given purpose, addressing the audience in suitable manner and using precise and appropriate language. Coding competences relate to the technical aspects of writing: spelling, punctuation, legibility, etc. For educators interested in writing as a meaning-making tool, assessing both types of competencies is a functional approach to writing instruction.

There have been partially successful attempts to develop writing assessment tools drawing on a functional approach; however, these tools have been developed for students from fourth to 11th grades (Skar & Aasen, 2018). Writing assessment tools for Grades 1 to 3 have so far been limited to coding competencies. This paper is the result of work driven by the ambition to close this gap. Consequently, we present the development of rating scales for assessing the writing proficiency development of young writers from a functional perspective. We review the concept of functional perspectives on writing and rating scale development before presenting our aim and research questions. The remainder of the paper explains the context of the study and its results. The paper concludes with a discussion of the results and their implications

A functional perspective on young children's writing

According to a functional approach to writing,[1] to write is to act purposefully. These acts include, but are not limited to, learning activities, memorization, and communication. From a functional perspective, writing is thus first and foremost a tool for interaction with oneself and others for varying purposes (Berge et al., 2016; Gee, 2004; Graham, 2018; Ivanič, 2004; Rose, 2016; Russell, 1997; Scribner & Cole, 1978; Vähäpassi, 1988). A proficient writer—one who can achieve the goals of writing—will produce discourse that can be part of a meaningful interaction with a reader, either somebody else or the writer herself (in a near or distant future). A prerequisite for young children entering the world of writing is mastery of basic aspects of writing's foundational techniques, including how to produce letters by hand and/or keyboard, as well as learning the relationships between phonemes and graphemes (Puranik & Lonigan, 2014).

In line with a functional perspective, writing proficiency is understood as a multifaceted construct. Various accounts of what it means to be able to write highlight that writing proficiency consists of several interrelated aspects, including devices for establishing writer-reader interaction, text-structuring devices, grammar, and mechanics (e.g. Bachman, 1990; Bereiter & Scardamalia, 1987; Evensen et al., 2016; Graham, 2018; Jeffery et al., 2018; Kellogg, 2008; Rijlaarsdam et al., 2012; Rose, 2016; Vähäpassi, 1988). The criteria for successful interaction through writing vary, depending on context and frame of reference. While some normative systems deem sequencing of the different parts of a text to be of upmost importance (e.g., "genre pedagogy"), others emphasize the writer's ability to express their "voice" (Elbow, 1973), and yet others focus on the writer's ability to convey given content by means deemed necessary and/or appropriate in a specified interpretive community (Evensen et al., 2016).

This project drew particularly on the work with writing assessment from a functional perspective by Evensen et al. (2016), Berge et al. (2019), and Skar (2017). From these perspectives, the most important criterion that distinguishes successful from unsuccessful attempts to write a text is the fulfillment of a contextually situated purpose; if a text has been written with the purpose of preserving information and manages to do so, the choice of genre and text structure may very well be atypical, as well as, for example, the choice of words. This perspective can be contrasted with writing assessment criteria that highlight form. One such example is genre pedagogy (Rose, 2016), where the purpose of writing entails default-relations to a number of linguistic choices; according to this theory, put in a somewhat extreme form, the merits of a text lie largely in how closely it has adhered to a predefined text structure.

[1] The terms "functional writing" and "functional approach to writing" are somewhat misleading. This is because "functional" is dependent on the normative system in which an activity takes place. For example, in a case where writing proficiency is defined as control of technical aspects, a functional (i.e., purposeful) approach would be to instruct in and assess technical aspects. The term is indeed also used in other ways, for example denoting teaching writing to fulfil highly specific tasks in for example vocational studies (Ivanič, 2004, p. 235).

Purpose and research questions

To meet the needs outlined above, writing assessment rating scales for students in Grades 1 to 3 were developed drawing on the recommendations of Knoch (2007) and the Council of Europe (2001) (see below). The purpose of this study is to describe the development process and to investigate the quality of the empirically based, assessor-oriented rating scales developed to measure writing proficiency from a functional perspective. The study answers three research questions:

RQ 1: What was the psychometric quality of the rating scales?
RQ 2: How did users perceive the quality of the rating scales?
RQ 3: Was it possible to use the rating scales in school settings?

Rating scale development

A rating scale can be defined as "a scale for the description of [writing] proficiency consisting of a series of constructed levels against which a language learner's performance is judged" (Davies et al., 1999, p. 153). The rating scale most often "provides an operational definition of a […] construct" (Davies et al., 1999, p. 153). Rating scale development thus hinges on a clear concept of the construct to be assessed.

When developing a rating scale, the developer faces several choices. First, one needs to define the construct to be assessed, as well as the primary audience for the rating scale. Alderson (1991) distinguishes between user-oriented (specifying to test users what a score means), assessor-oriented ("guiding the rating process," p. 73) and constructor-oriented (guiding the item development process) rating scales. A parent, or other stakeholder, would typically be most interested in user-oriented rating scales because of the need to understand what a score represents beyond the immediate context of the writing test. A teacher, or another rater, would be most interested in an assessor-oriented rating scale, as it provides information on how to rate or mark features of a text.

In rating scale development, one also needs to decide if the assessment is to be holistic—i.e., one rating scale is used and the text is awarded a single score—or analytical, i.e., a text receives scores on several rating scales. There are other options as well, but they are less common (Weigle, 2002). Furthermore, the developer must decide on what to base the rating scales. Fulcher and Davidson (2007) have identified three such areas: intuition (or experience), empirical data (i.e., student texts), and writing development theory. Rating scale development processes are seldom reported, and intuition-based rating scales, which often are said to be common across the world, have several drawbacks (Knoch, 2007); they may lack empirical support, as may the ordering of descriptors. Some descriptors might comprise aspects that are theoretically, but not necessary empirically related. With empirically based rating scales, however, these disadvantages

disappear, but of course, new ones appear: Features described in rating scale descriptors will be limited to features in the texts at hand. Empirically derived rating scales are, however, recommended (Knoch, 2007); they do, as it were, increase the probability of a match between descriptors and actual features of student texts. Furthermore, a rating scale developer must determine the characteristics of the descriptors. The Common European Framework of Reference set a standard almost 20 years ago, defining satisfactory descriptors as follows: positively oriented (focusing on what the candidate knows and not the opposite), concrete and free from vagueness, transparent, short, and able to function independently (without the need to read other descriptors) (Council of Europe, 2001, p. 205).

There are numerous ways of investigating the qualities of assessment rating scales. Knoch (2009) offers a comprehensive description of statistical analyses appropriate for rating scale validation. Two important aspects include whether rating scales can be used to distinguish between texts of different quality and whether the rating scales allow for reliable assessment (e.g., high inter-rater agreement). It is also important to investigate how the rating scales are perceived by the intended audience (i.e., users, assessors, or developers).

Participants

There were two panels involved in the different steps of the rating scale development (cf. Figure 1). A ranking panel, who performed comparative judgement (see below), consisted of 17 members with a mean age of 45.9 years (SD = 11.5). All had experience from working in school (M = 6.3 years, SD = 6.8, range 1–22 years), although they had more experience from working in teacher education (M = 12.0, SD = 9.5, range 0.3–33 years). A rating panel, who performed ratings based on the first to third drafts of the rating scale (see below), consisted of 16 members with a mean age of 42.0 years (SD = 6.3). All held teaching certificates, one had a bachelor's degree, eight held master's degrees, and seven had doctorate degrees, all in subjects relevant to young children's writing. This group also had experience working in schools (M = 7.8 years, SD = 6.9, range 1–22 years) and in teacher education (M = 7.4, SD = 4.5, range 0.6–18 years).

Data were also collected from teachers who worked in four schools across Norway that participated in piloting the rating scales. These teachers (N = 47) were granted total anonymity, and no data on their backgrounds were collected. Nineteen of the teachers, however, participated in audio-taped recordings while using either of the two versions of the rating scales.

Student texts

There were 1,001 texts used in the rating scale development process. Some texts were collected for the purpose of rating scale development (see technical report; Skar, manuscript), while other texts were collected from corpora that were at the

university's disposal. The texts represented student writing in two genres—informative and narrative—from the first six semesters in school. As can be seen in Table 1, the distribution of texts across genres and semesters was uneven.

The texts were used in different ratings steps (for substantive information on the steps, see section Context of the study and developmental process of the rating scales), with 401 texts used in step two and 600 texts used in rating step seven and step nine. Analysis of variance with subsequent Bonferroni corrected t-tests (excluding texts from the newly arrived students) indicated quality differences between texts written by students at different stages of the first three years of school. For texts used in both comparative judgement and ratings, the semester the text was produced in had a significant effect on text quality, with $F(4,369) = 232.8$, $p < 0.001$ and $F(4,568) = 209$, $p < 0.001$, respectively. The differences between semesters were all in the same direction: Texts from later semesters were constantly deemed to be of higher quality than texts from earlier semesters. In all but three instances (comparative judgement: fourth vs. fifth semester; ratings: third vs. fourth semester, fourth vs. fifth semester), these differences were significant. See Appendix B for specifics.

Table 1. Student texts used in scale development

		First Semester		Second Semester		Total
		Informative	Narrative	Informative	Narrative	
Step #2	1st Grade	130	0	0	0	130
	2nd Grade	92	0	33	0	125
	3rd Grade	58	0	45	16	119
	Newly Arrived*	27	0	n/a	n/a	27
Step #7, #9	1st Grade	73; 76 (3)	0; 0	0	0	152
	2nd Grade	91; 63	34; 32 (2)	0; 11	10; 11	254
	3rd Grade	57; 49 (2)	0; 0	19; 20	7; 12 (1)	167
	Newly Arrived*	3; 21 (2)	1; 0	n/a	n/a	27
Total		747	69	128	57	1001

Note. For Step #7, Step #9 rows: First number = number of texts for step #7; second number = number of texts for step #9; number in parenthesis = number of anchor texts. *Students that recently (≤ 12 months) had arrived in Norway.

Context of the study and developmental process of the rating scales

The development of rating scales was part of a larger research project, Functional Writing in Primary School (FUS), with the first author as principal investigator. The overall aim of the larger project, which included a writing intervention program, was to increase the quality of writing instruction in first and second grades in Norway. The goal of the instructional activities within the program was to promote students' proficiency in using writing to communicate. The rating scales were part of the project in two ways. First, they served as one of several means of evaluating the effect of the project (i.e., as criteria when assessing

student texts using a pre-post design), which made it important to ground them in the functional view of writing presented above. Second, they were important tools in the program, offering teachers following the program criteria for the formative assessment of student texts.

For the rating scales to function properly in relation to these purposes, they needed to account for writing proficiency from first to third grades. They needed to relate both to the communicative force of the text and to the more technical aspects. The rating scales were developed using an empirical approach and several methods. The sections immediately following this one describe the contours of the developmental process—providing necessary information for understanding the investigation—while the sections on data collection and analysis provide the technical details of the process, as well as some essential descriptive statistics.

The development of the rating scales generated five consecutive versions and followed 12 distinct steps (see Figure 1). First, the FUS research group decided to develop a tool for analytical assessment that includes eight rating scales (please refer to Appendix A for the content of these rating scales). The types of scales and number of scales were based on previous scale development work by some of the researchers (Evensen et al., 2016; Skar, 2017). Second, using so-called comparative judgement in the versatile software environment No More Marking® (NNM), 400 student texts were rank ordered by a panel commissioned by the research group. Comparative judgement (Pollitt, 2012), as implemented in NNM, builds on what is known as the Bradley-Terry-Luce model (Wheadon, n.d.), which is a Rasch logistic model (Rasch, 1980):

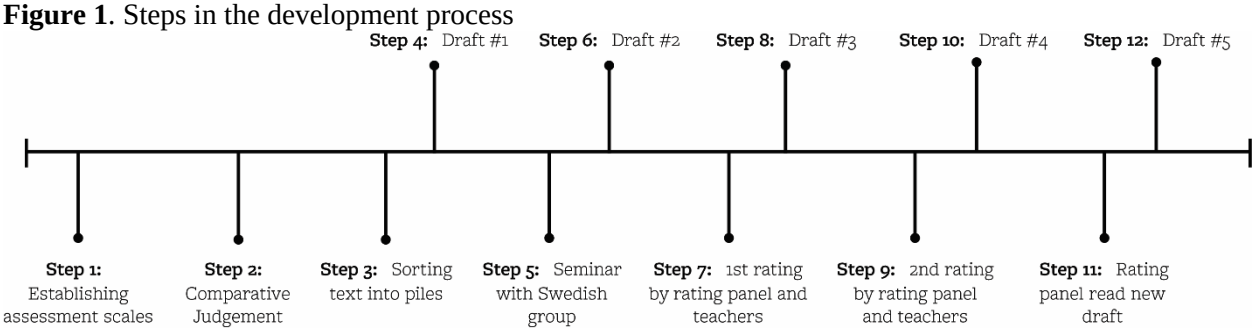
$$probability(A\ beats\ B) = \frac{\exp(x_A - x_B)}{1 + \exp(x_A - x_B)}$$

In the model, x_A and x_B are estimates of the quality of two student texts, A and B, which are being compared. Through the implementation of this particular model, the NNM system lets the analyst set up an environment wherein texts are compared with each other until a reliable rank order is achieved. A judge is presented with two texts side by side and judges which is the better one. It should be noted that no other criteria are used in such a case. With enough comparisons and high enough reliability, the resulting measures can thus be interpreted as relative positions—that is, the first text in the pile is better relative to the other 399 other texts in the pile. In this case, 4,104 comparisons were made, and the reliability was 0.92. Compared to a "traditional" method wherein texts are sorted into piles based on intuition, the comparative judgement method offers a more efficient and indeed more reliable initial piling (Aasen & Skar, 2018).

In the third step, the research group sorted the student texts into five equally sized piles (i.e., N/5 = 80 texts per pile), based on the rank ordering, with each pile tentatively representing a proficiency level. The motivation for this was that it was decided beforehand to draft five levels for each rating scale. In step four,

the research group drafted descriptors of proficiency related to the eight rating scales based on texts in the different piles (Draft #1). In step five, the research group presented and discussed the drafts with a sister project (Functional Writing in Early School Years: Assessment, Teaching and Professional Development) in Sweden. Sixth, the rating scales were slightly revised (Draft #2). Seventh, the drafts were piloted on 300 texts by a new panel also commissioned by the research project, as well as by teacher groups across Norway. In step eight, the rating scales were revised based on feedback from the pilot (Draft #3). One rating scale (tentatively called "Content") was dropped altogether, and one was added ("Relevance"). In step nine, the rating scales were piloted again, using 300 texts, the same panel, and a new teacher group. In step 10, the rating scales were revised based on feedback (Draft #4) from those involved in piloting. In the 11th step, the rating scales were piloted a third time by the panel, who provided feedback. In the 12th and final step of the developmental process, the rating scales were finalized (Draft #5). Feedback from people involved in piloting is presented in the results section.

Figure 1. Steps in the development process



In total, eight rating scales were created: Audience Awareness (S1), Vocabulary (S3), Organization of Content (S4), Language Use (S5), Punctuation (S6), Spelling (S7), Handwriting (S8), and Relevance (S9). A content rating scale (S2) was piloted in Draft #2 but dropped in Draft #3, as its aspects were incorporated in other rating scales. S2 will not be discussed further.

As an example of the developmental process, Table 2 delineates the evolution of rating scale descriptors for Level 1 and Level 5 for two rating scales, S1 and S7. S1, in all versions, focused on the writing act, or the text as a communicative utterance. The first version stated that it was "unclear" what the text communicated, and, however redundantly, that "it is necessary to talk to the writer to understand the text" at Level 1. The rating scale was called "interaction," but it did not clearly describe interaction; rather, it described how well the text seemed to convey the writer's intentions. Draft #1 (version for Level 5) contained two descriptors—one that targeted the content aspect and one that applied to the audience of the writing prompt. It also focused on how well the text accommodated the reader's need for information. The latter was retained in later drafts, while the former shifted to focus on how the reader in the writing prompt

was addressed. Drafts #3 and #4 contained a descriptor pertaining to the type of content, which was replaced with a descriptor related to the student's voice. The different versions of S1 focused on how well the text met the inherent requirements of a given writing prompt. In turn, this required prompts that specified purpose and audience.

For S7, we note that the rating scale focused on technical aspects of spelling. The rating scale underwent a major change from Draft #3 to Draft #4, shifting focus from simply correct orthography to frequency of correctly spelled words and the complexity of those words.

Table 2. Evolution of scale descriptors

1st Draft	2nd Draft	3rd Draft	4th Draft	Final Version
Scale 1: Interaction	Scale 1: Writer–Reader Interaction	Scale 1: Writer–Reader Interaction	Scale 1: Audience Awareness	Scale 1: Audience Awareness
Level 1: - It is unclear what the text communicates OR the text consists of isolated meaningful words/expressions/drawings - It is necessary to talk to the writer to understand the text	Level 1: - It is necessary to talk to the writer to understand the text	Level 1: - It is necessary to talk to the writer to understand the text	Level 1: - It is necessary to talk to the writer to understand the text	Level 1: - It is necessary to talk to the writer to understand the text
Level 5: -The text is a meaningful answer to the writing prompt AND/OR the text is understandable without knowledge about the context in which it was created - The text accommodates reader's need for information about participants, context and events in a good way	Level 5: The text addresses the reader specified in the writing prompt in a relevant manner and accommodates reader's need for information about participants, context and events	Level 5: - The text addresses the reader specified in the writing prompt in a relevant manner throughout the text and it accommodates reader's need for information about participants, context and events - The text can contain generalizations, reflections and evaluations	Level 5: - The text addresses the reader specified in the writing prompt in a relevant manner throughout the text and it accommodates reader's need for information about participants, context and events - The text can contain generalizations, reflections and evaluations	Level 5: - The text addresses the reader specified in the writing prompt in a relevant manner throughout the text and it accommodates reader's need for information about participants, context and events - The text *may* contain traces of the student's voice using reflective or evaluating utterances
Scale 7: Spelling				
Level 1: Scribbling AND/OR drawings	Level 1: There may be letters in the text AND/OR scribbling	Level 1: There may be letters in the text AND/OR scribbling	Level 1: There may be letters in the text AND/OR scribbling	Level 1: There may be letters in the text AND/OR scribbling

Level 5: Mainly correct orthography	Level 5: Mainly correct orthography	Level 5: Mainly correct orthography	Level 5: There are numerous instances of correctly spelled words where phoneme and grapheme does not correspond	Level 5: There are numerous instances of correctly spelled words where phoneme and grapheme does not correspond

Data collection and analysis

Table 3 summarizes the data collected. To answer RQ1, the ratings of the texts in steps seven and nine were collected. In each step, the rating panel individually rated 60 texts. All texts were rated by two raters, and to ensure comparability across the panel, some texts were rated by all panelists. Ten texts were included in both step seven and step nine. The ratings of the panel were modelled using the so-called many-faceted Rasch measurement (MFRM-models). More specifically, the following MRFM model (Engelhard, 2013; Linacre, 2017b) was used in this analysis:

$$log(Pnij(k)/Pnij(k-1)) = Bn - Ei - Cj - Fx,$$

where Pnmijk represents the probability of student n, rated on rating scale i, by rater j, receiving a score of k, and Pnij(k-1) represents the probability of the same student under the same conditions receiving a score of k-1. Bn is the ability for person n, Ei is the difficulty of rating scale i, and Cj is the severity of rater j. Finally, Fx represents the point on the logit scale where category k and k–1 are equally probable. Ability, difficulty, and severity are all expressed on the same interval scale: the logit scale (Engelhard, 2013). By convention, the mean of the logit scale is 0.00, and it most often ranges between -4.00 and 4.00. Because it is an interval scale, the distances between, for example, raters or rating scales have substantive meaning (Stevens, 1946).

The MFRM was used to allow for detailed analysis of rating scale tools, as well as analysis of rater behavior. More specifically, the MFRM allowed us to investigate the quality of the rating scales from five perspectives, following the standard for writing rating scale investigations established in Knoch (2009). The following five characteristics were investigated. (1) Discrimination of rating scale as measured by "student separation" expressed as separation index was investigated. The separation index (H-index) is "the number of statistically distinct levels" (Eckes, 2015, p. 62) of a given facet (e.g., raters, students). A higher separation index is perceived to be superior to a lower separation index, as it indicates greater ability to discriminate between candidates when using the rating scale. (2) Rater separation was investigated using the H-index. Contrary to student separation, the fewer classes the better result, as few classes indicated

small differences in severity and leniency. (3) Rater reliability as expressed in the measure "single rater-rest of raters" (SR-ROR) was investigated. Knoch explains the measure as expressing to what extent a single rater's ratings are consistent with all other raters' ratings. Guidelines suggest that correlations in the interval 0.30–0.70 are acceptable, and values below that are to be regarded as low, while values above that are to be regarded as high. To complement the SR-ROR measure, we also computed the intraclass correlation coefficient (ICC), expressing conventional reliability measures between pairs of raters (see Skar & Jølle, 2017 for details). (4) Variations in ratings were investigated. The so-called infit statistic expresses rater variability. The expected value is 1.0, and values exceeding or falling below this indicate more or less variation in the ratings, respectively (Eckes, 2015, p. 77). The statistic has previously been used as an indication of intra-rater reliability (Weigle, 1998), which is a measure of a rater's ability to use the rating scale in a consistent manner. In keeping with Knoch (2009, p. 204), significant values above 1.3 were considered troublesome, indicating large variation. Significant values falling below 0.7 were also considered somewhat troublesome, as they indicated a tendency to use only parts of the rating scale (e.g., overusing certain scale steps). (5) Scale step functionality was investigated using three measures. First, we controlled for the so-called "average measures"—that is, the average logit value associated with a certain scale step, advanced monotonically. Second, we investigated the so-called outfit measure, which, much like infit for raters, expresses estimates of expected and less expected variation. With an expected value of 1.0, values exceeding 1.4 indicated troublesome scale steps requiring further investigation. Third, we investigated the rating scale category thresholds, which are the points at which a student text with the corresponding logit value has a 50% chance of being observed in either one of two adjacent categories. With five categories, these values should increase monotonically by at least 1.0 and no more than 5.0. In addition, adhering to advice from Eckes (2015), we also investigated frequencies across categories, checking (a) that all categories had a minimum of 10 responses, (b) the distribution characteristics, and (c) that no categories were unobserved.

To answer RQ2, we surveyed the rating panel on their perceptions of the rating scales, asking them to rate how they perceived the quality and usefulness of the rating scales. Specifically, the panelists were asked to rate on a six-point Likert scale to which extent they agreed that the individual rating scale was relevant, sufficient, and if the rating scale should be a candidate for deletion— regardless of quality—to save time and effort in rating. The questions were inspired by Bachman's (2005) list of quality traits in language assessment. Readers should refer to Appendix C for a description of the items used in the survey. The surveys were analyzed using descriptive statistics.

Last, to answer RQ3, the rating scales were piloted in school settings by teachers at four schools. At three of the schools, 19 teachers agreed to participate in the audio recording of live assessments wherein teachers used the rating scales

to assess texts. All teachers assessed the same texts, and each assessment lasted an average of 35 minutes (range: 25–40 minutes). The teachers worked in groups (4–5 teachers in each group), and while they were given a short introduction to the project, they received no training in using the criteria. The main reason for this was that we wanted to leave the settings as realistic as possible; in many cases teachers will receive material in textbooks and the like without training in how to use it. Three groups used Draft #2 rating scales and two groups used Draft #3 rating scales. The result was five recordings of teachers talking while testing the rating scales.

The audio data was analyzed by noting instances of "using" or "questioning" the rating scales. The hypothesis was that uncommented usage (e.g., verbally applying a criterion without commenting or questioning it) would indicate that the scales functioned or at least were accepted, while questioning would indicate problems with the questioned part of the rating scale. For each recording, we noted the number of instances of questions and further categorized them into one of three types of questioning, namely (i) questioning related to difficulty of using the rating scale for reasons of ambiguity, (ii) questioning related to difficulty of using the rating scale for reasons of non-communicative descriptors (descriptors hard to grasp), and (iii) questioning of the appropriateness of the descriptors. The categories were set beforehand, based on the authors' work with assessment panels (e.g. Skar & Jølle, 2017), and the first and second author together coded all audio data.

Table 3. Overview of data in relation to draft of scales

	Ratings	Panel Survey	Teacher Audio Data	
			Participants	Recordings in min.
Draft #1	-	-	-	
Draft #2	295+10	16	3 (10)	40; 25; 32
Draft #3	295+10	16	2 (9)	39; 40
Draft #4	-	16	-	
Draft #5	-	-	-	
Total	600	48	5 (19)	176

NB. Ratings: There were 295 unique texts used for rating and ten texts that were constant across drafts. Audio data: number of groups with number of individuals in parenthesis.

Results

RQ 1: What was the psychometric quality of the rating scales?

Tables 4, 5, and 6 summarize the findings from the quantitative investigation using ratings from steps seven and nine. As can be seen in Table 4, the rating instrument and context of judgement produced reliable ratings, and statistically there was room to separate the students into roughly seven performance levels across the two drafts. The simplifications associated with the redrafted rating

scales used in step nine did not seem to hinder meaningful separation of students. The raters were also separated with high precision into eight and five groups, in each step. While the number of student groups remained almost the same, the number of rater groups saw a non-trivial decrease in groups, indicating fewer marked differences in severity when rating Draft #3. The single rater-rest of raters (SR-ROR) correlation was within boundaries (0.58 and 0.57) on both occasions, and of 16 raters, only one displayed a somewhat high infit, indicating unexpected ratings.

Table 4. Separation, correlation, and infit from MFRM-analysis

	H-index	H-index	SR-ROR (*M*)	Infit
	Students	Rater		*n* > 1.3 (value)
Draft #2	7.13	7.99	0.58	1 (1.45)
Draft #3	6.70	5.18	0.57	1 (1.34)

Note. Infit values: number of raters with significant (i.e. $z \geq 2.0$) infit ≥ 1.3.

Table 5 presents the intraclass correlation coefficient (ICC). In this context, where all student texts were rated twice, the ICC average measure is of greatest interest. For all rating scales, the correlation exceeded 0.7, which has traditionally been regarded as a minimum value for ratings to be acceptably reliable (McNamara, 2000). For most rating scales, the value exceeded this, and the average correlation across rating scales was 0.86 and 0.86 for the two drafts, respectively. The most problematic rating scale was spelling, with an average correlation of 0.76 and 0.71, respectively, and a single correlation of 0.62 and 0.56, respectively. The qualitative investigations presented in the rating scale development section indicate that the descriptors were not distinct enough, and several raters complained that it was difficult to know how to score very short texts with no spelling errors (please refer to Table 2). The last version of the rating scale included a requirement to display a repertoire of correctly spelled words at the highest level (see Appendix A).

Table 5. Classical Test Theory Reliability Measures

	Draft #2			Draft #3	
	ICC Single	ICC Average		ICC Single	ICC Average
S1: Audience	.75	.85		.70	.82
S3: Vocab.	.71	.83		.77	.87
S4: Org.	.75	.85		.75	.85
S5: Lang.	.78	.88		.82	.90
S6: Punct.	.84	.92		.85	.92
S7: Spell.	.62	.76		.56	.71
S8: Handw.	.78	.88		.74	.85
S9: Rel.	n/a	n/a		.91	.95
Average	.75	.86		.76	.86

Note. ICC = Intraclass correlation coefficient average across rater pairs.

Table 6 summarizes the scale step functionality investigation. The average (logit) measure associated with each step advanced monotonically on both rating occasions. Also common to both drafts was category five having a relatively high outfit. It did not, however, exceed the critical value of 1.4, and, as has been noted elsewhere, so-called extreme categories are more likely to have large outfit than central categories (Linacre, 2017). In addition, the category threshold values indicated well-functioning scales. The only exemption was the increase from category two to category three on Draft #3, which equaled 0.72, slightly less than 1.0. However, all categories were exclusively "most probable" for some areas of the logit scale. The number of observations in each category indicated a pattern, wherein category three was most popular, followed by categories two and four. Category five was least popular, with 50.9% and 52.5% as many observations as category one in each draft, respectively. No category included fewer than 10 observations.

Table 6. Scale Step Functionality

		Ave. Meas.	Outfit	Category Threshold	N observations (%)
Draft #2	Category 1	-4.06	0.9		1062 (14.1)
	Category 2	-1.57	0.8	-3.03	1599 (21.2)
	Category 3	0.11	1.0	-1.33	3005 (39.9)
	Category 4	1.26	1.1	1.42	1329 (17.6)
	Category 5	2.62	1.4	2.94	541 (7.2)
Draft #3	Category 1	-3.05	0.8		1275 (16.9)
	Category 2	-1.22	0.8	-2.08	1310 (17.3)
	Category 3	0.04	1.0	-1.36	2666 (35.3)
	Category 4	1.05	1.0	0.98	1631 (21.6)
	Category 5	1.90	1.4	2.47	670 (8.9)

Note. N observations equals counts used in the MFRM analysis. Actual observations are slightly higher, but MFRM analysis excludes extreme cases.

To gain additional insight into the consequences of re-drafting the rating scales, we used the 10 student texts mentioned in the data collection section above as anchors when analyzing all rating scales at once. The results of the analysis are presented in Table 7, which shows raw score averages, logit measures, and their standard errors. The results indicate that the relative positions of rating scales basically remained between the drafts, although some rating scales became slightly "easier" (e.g., spelling, handwriting) and some rating scales became more difficult (e.g., audience awareness). The exception was handwriting, and consequently audience awareness, which shifted positions. In all, however, re-drafting the rating scales did not seem to alter the relationship between the rating scales.

Summarizing the psychometric investigations of the rating scales, one notices that both Draft #2 and Draft #3 showed signs of functioning well. The changes in

descriptors between the second and third drafts were extensive, but the high reliability remained, and the relationship between the rating scales seemed to be preserved.

Table 7. All Scales Linked. Presented in Descending Order (from "hardest" to "easiest")

	Raw Score Average	Logit Measure	Logit S.E.
S6: Punct. – Draft #2	2.0	1.86	0.05
S6b: Punct. – Draft #3	2.18	1.42	0.05
S4: Org. – Draft #2	2.6	0.53	0.05
S4b: Org. – Draft #3	2.6	0.47	0.05
S5: Lang. – Draft #2	2.69	0.33	0.05
S5b: Lang. – Draft #3	2.73	0.18	0.05
S3b: Vocab. – Draft #3	2.84	-0.06	0.05
S3: Vocab. – Draft #2	2.88	-0.08	0.05
S9b: Rel. – Draft #3	2.91	-0.22	0.05
S1b: Audience – Draft #3	2.97	-0.36	0.05
S8: Handw. – Draft #2	3.03	-0.44	0.05
S1: Audience – Draft #2	3.04	-0.45	0.05
S7: Spell – Draft #2	3.09	-0.56	0.05
S7b: Spell – Draft #3	3.23	-0.91	0.05
S8b: Handw. – Draft #3	3.35	-1.17	0.05

Note. H-index: 22.3.

RQ 2: How did users perceive the quality of the rating scales?

Table 8 summarizes the findings for the three surveys given to the rating panel. For all individual rating scales—Drafts #2 to #4 (cf. Figure 1)—the rating panel was asked to judge the relevance of the rating scale, the sufficiency of the descriptors, and whether a rating scale, regardless of its qualities, should be deleted (to save time). Each panel member was asked to mark on a six-point scale how strongly they agreed or disagreed with claims that rating scales were relevant, sufficient, or should be deleted. A score of one indicated the respondent strongly disagreed, and a score of six indicated the respondent strongly agreed (see Table 8).

Table 8 presents descriptive statistics in the form of means across items and across drafts for the three questions. It also provides an estimate of the consistency of scores across the rating scales in the form of an alpha value. Finally, it indicates what the consistency would be if the most troublesome rating scale was removed. The alpha value shall not be interpreted as an indication of the raters' general perceptions of the rating scales, but rather as an indication of how systematic the pattern of judging the rating scales was. In turn, this indicated a consistent or inconsistent view of the relative merits of the rating scales within the rating panel.

The rating panel generally and moderately agreed that the rating scales were relevant with M = 5.2 (SD = 1.7) for Draft #2, M = 5.4 (SD = 0.75) for Draft #3 and M = 5.1 (SD = 0.9) for Draft #4. Initially, the rating panel members were

inconsistent in their evaluation of the rating scales (α = 0.49). Deleting S7 would have improved the alpha value considerably (α = 0.57). The judgements of the rating scales were more consistent regarding the last draft (α = 0.71), but would have been improved to α = 0.77 if S3 had been deleted.

The rating panel generally agreed, slightly or moderately, that rating scale descriptors were sufficient, with a one-point increase from Draft #2 (M = 4.0, SD = 1.28) to Draft #4 (M = 5.0, SD = 0.80), via M = 4.8, SD = 0.99 for Draft #3. The alpha value indicated good consistency for Draft #2 (α = 0.77) and Draft #4 (α = 0.76), with a temporary drop for Draft #3 (α = 0.59). For both Drafts #2 and #4, deleting S3 would have increased consistency.

Overall, the panel moderately disagreed with the deletion of any rating scale in order to speed up assessment (Draft #2: M = 2.0, SD = 1.3; Draft #3: M = 1.65, SD = 0.96; Draft #4: M = 2.0, SD = 1.2). The alpha values indicated consistency (α = 0.70, α = 0.75, α = 0.65, for the three drafts respectively), albeit not at the same high levels as for sufficiency. As for relevance and sufficiency, S3 stood out as a source for inconsistency. For example, deleting S3 at Draft #3 would have increased the alpha value to 0.84.

In summary, the investigation showed that the rating scales were generally perceived to be relevant across drafts, and each draft (with associated simplifications) increased the impression of sufficiency. Generally, the panel responded negatively to deleting any of the rating scales to save time. However, the vocabulary rating scale was associated with inconsistency.

Table 8. Surveys to the rating panel

		Mean	Std deviation	Alpha	Alpha improve	Item to delete
Relevance	Draft #2	5.16	1.16	0.49	0.57	S7: Spell
	Draft #3	5.39	0.75	0.69	0.75	S3: Vocab
	Draft #4	5.12	0.92	0.71	0.77	S3: Vocab
Sufficiency	Draft #2	4.04	1.28	0.77	0.80	S3: Vocab
	Draft #3	4.84	0.99	0.59	0.61	S8: Handwriting
	Draft #4	4.96	0.80	0.76	0.82	S3: Vocab
Deletion	Draft #2	2.04	1.26	0.70	0.79	S1: Audience
	Draft #3	1.65	0.96	0.75	0.84	S3: Vocab
	Draft #4	2.01	1.22	0.65	0.73	S3: Vocab

Note. 1 = Strongly Disagree, 2 = Moderately Disagree, 3 = Disagree Slightly, 4 = Agree Slightly, 5 = Moderately Agree, 6 = Strongly Agree. Alpha improve: how much would consistency increase if any assessment scale (i.e. item) was deleted?

RQ 3: Was it possible to use the rating scales in school settings?

Three teacher groups at two different schools assessed student texts using Draft #2 rating scales (step six), and two groups at one school did the same using Draft #3 rating scales (step eight). We considered it to be important to pilot the rating scales to get information about the appropriateness of the descriptors at the

different developmental steps. Our main reason is that teachers are, together with assessors, the intended users of the rating scales. The assumption was that both Draft #2 and Draft #3 would function well for the intended purpose, but that Draft #2 would function less well than Draft #3 due to it being an earlier version.

Overall, the impression was that both drafts functioned well. The teachers seemed to accept the rating scales as valid tools for assessing young writers' texts. The following example shows the unquestioned use of the rating scale that completely dominated all five group assessments:

> Teacher: "Complete sentences may occur" [citing S5]. Yes, one complete sentence occurred in this text. We do not need to go to the next level. The text belongs here.

This excerpt was coded as "using" the rating scale. As can be seen above, the teacher cited the descriptor and made an evaluation about whether the descriptor accurately described a specific feature in the text. When the rating scale descriptor(s) reflected the teacher's perceived quality of the text, they were ready to move on to the next rating scale.

Fifteen times during the almost three hours of assessing across the five groups, teachers questioned the rating scales: seven times while using Draft #2 rating scales and eight times while using Draft #3 rating scales. Five times, teachers' questioning was related to perceived ambiguity (questioning category [i]) within and between descriptors on a rating scale. The next example displays how the teachers discussed at which level on rating scale S1 a particular student's text belonged. They read the descriptors from Level 1 and upward. Reaching Levels 4 and 5, they encountered problems, and one teacher stated:

> I find it hard to decide. What is the difference between Level 4 and Level 5? What does it really take to reach Level 5?

For one and a half minutes, four teachers elaborated on this problem, moving between the text and the rating scale descriptors, before concluding that the text had features that allowed them to assess the text as both belonging to Level 4 and Level 5. In other words, descriptors that do not function to discriminate between rating scales are problematic.

Most often, the teachers' questions were related to non-communicative aspects of some of the descriptors (questioning category [ii]). Nine times, teachers had comments of this sort, and the next excerpt, where the teachers assessed the text's quality related to rating scale S3, serves as an example:

> Teacher A: I am at Level 4. "The text displays variation in "tema rema-binding" [theme-rheme]. What does that mean?"
> Teacher B: I have not heard that expression before.
> Teacher A: Good to hear.
> Teachers: [Laughter].

> Teacher C: We have to Google that.
> Teachers: [Laughter]
> Teachers A: Is it that it rhymes? No?

As can be seen, the teachers were unfamiliar with the expression "tema rema-binding." The chosen strategy in these situations was to ignore the non-communicative descriptors while trying to find support elsewhere to complete the assessments.

The last category of questioning related to the appropriateness of the descriptors (question category [iii]). Only one instance of questioning was coded as belonging to this category. After citing a relatively complex Level 5 descriptor for rating scale S2, a teacher stated:

> Teacher A: I was thinking—about the level, yes. They are Year 2 students, true. If they had been older students, we could have assessed them using this descriptor. So, to me it seems like a Year 2 text must belong to the lower levels.

The teacher found the descriptors for higher levels within the rating scales too ambitious for the Year 2 text she had in front of her. A rating scale with no or few instances of Level 5 texts would have been problematic, but in the context of this example, relevant information may have helped to explain the excerpt. This was the first text the teacher group assessed using the tool, and as novices in this new context, they were unfamiliar with the rating scales. The first text they assessed was, overall, a lower-level text. Later, they assessed a more advanced Year 2 text, where the same teachers used the higher levels in the rating scales (including content) without problematizing the appropriateness. Finally, since this was the only coded instance in this category, the overall impression was that the teachers found the rating scales and the levels within the rating scales both relevant and appropriate.

Discussion and conclusion

This study aimed to describe the process of developing empirically based assessor-oriented writing proficiency rating scales and to investigate the quality of those rating scales. We did this by focusing on psychometric qualities, users' perceptions of quality, and teachers' use of the rating scales. Overall, the investigation found indications of well-functioning rating scales already from the first draft that was piloted. This investigation's results indicate that it is possible to draft rating scales for the assessment of young children's writing proficiency that can be used for reliable assessment and that are perceived by users to capture a particular construct.

The key purpose of developing the writing assessment rating scales was to operationalize the construct of functional writing in order to make it possible to

assess students' writing proficiency. In our context, functional competencies of writing imply using writing as a meaning-making tool for accomplishing different purposes (i.e., writing to communicate, for learning, for developing identity). In keeping with this approach to writing, we developed eight rating scales focusing on functional competencies and coding competencies. We found that when operationalizing a functional approach, it is important to describe both functional competencies—such as skills in audience awareness and text organization—as well as coding competencies, since the latter represent a prerequisite for writing.

We developed the rating scales using an empirical approach, basing the descriptors on actual texts written by students in Grades 1 through 3. We used the method of comparative judgement as a tool for rank ordering texts to provide a reliable overview of differences in text quality. This method enabled us to sort the texts into five stacks without using interim-descriptors and allowed us to use a true explorative approach when describing characteristics of texts in each stack. The developmental approach adopted by this project created descriptors that represent students' writing rather than rating scale developers' experiences and their notions of writing development. There is, however, a limitation to this method: Too small a text sample always introduces the risk of overlooking key features in texts. Therefore, further studies need to investigate more closely whether the rating scales are applicable for all types of texts in Grades 1 through 3. Any limitations in this regard are important to note, as they affect the possibility of adequately representing writing proficiency and writing development between grades.

Our first research question was as follows: What was the psychometric quality of the rating scales? The statistical analyses indicated satisfactory reliability for the rating scales measuring functional competencies. Provided that future assessments are carried out as this study was (with two raters per text), the reliability could indeed surpass that of previous attempts to formulate such rating scales (Purves, 1992b; Skar et al., 2017; Skar & Aasen, 2018; Thygesen et al., 2007). In turn, this indicated that it is possible to develop rating scales that greatly expand the existing toolkit (Carlsten, 2016; Norwegian Reading Center, 2018) for reliably assessing writing in Grades 1 through 3. That is to say, there are convincing psychometric arguments for including the functional aspects of writing proficiencies even when assessing beginner writers. The statistical investigation also indicated good reliability across two different drafts of the rating scales. The development of rating scales in an iterative process presents a risk that statistical quality will vary between drafts. However, in this case, we suspect that the initial process of comparative judgement allowed us to, from the beginning, base descriptors on texts that were sorted into stackswith distinct differences.

Our second research question ("How did users perceive the quality of the rating scales?") focused on user perceptions of the rating scales. The rating panel seemed to throughout the project find the rating scales satisfactory; they found

them to be sufficiently detailed, and to a large extent the panel refrained from suggesting that any of the rating scalesshould be deleted. The investigation did uncover, however, some disagreement concerning the vocabulary rating scale. This disagreement was not associated with low reliability, but it raised some concern for future investigations, which will need to examine more closely what issues raters may have with vocabulary. Such information will be useful when designing supplementary materials (such as annotated exemplar texts).

Summarizing the teachers' use of the rating scales—which was targeted by RQ3 ("Was it possible to use the rating scales in school settings?")—the analyses showed that use changed very little between drafts; this also indicated that such use should be interpreted as positive, given the few instances of questioning. Even at Draft #2, the teachers' use of the rating scale indicated that they found the rating scales acceptable. Further, the instances of questionings were few and short (cf. the excerpt related to questioning category [ii]), indicating that teachers were able to use the rating scales more than they questioned them. The few instances that did occur indicated two things. First, later revisions should make sure that descriptors in adjacent levels within a rating scale discriminate well (cf., question category [i]), and second, that meta language (i.e., terminology) in the descriptors is known to the teachers (cf., question category [ii]).

The rating scales are intended to serve different but related audiences: professional raters and teachers. The psychometric investigations indicated that the former audience seemed to be able to use the rating scale as it is now. This seemed to be partially true for the latter audience as well, but instances of questions and questioning indicated that there will be a need to include a comprehensive list of concepts and their definitions (e.g., "theme–rheme"), as well as annotated exemplar texts.

There are many choices in rating scale development, and in this particular instance, a total of eight rating scales with varying degrees of extensive descriptors were developed through several steps. This method of producing descriptors has been time consuming, labor intensive, and expensive, but has been worth the effort: Professional raters and teachers can (with the abovementioned caveats) be assured that the descriptors are based on empirically identified features of student texts. In settings where the rating scales are to be used, it is plausible to predict that the rating scales may facilitate reliable ratings.

The development and type of validity investigations presented in this article are merely the starting point for investigations of ratings scales. Future work will need to investigate how the ratings scales function in applied settings, including as a tool to measure the effect of the FUS project and – eventually – as a tool in teachers' everyday practice of formative writing assessment. The former is of upmost importance since the quality of students' texts will be one of the major dependent variables when estimating the effect. The second is equally important since there is a lack of tools for assessing young students' writing. Should future investigations provide robust evidence of the usability of these rating scales in

everyday school practice, there is a real chance that these rating scales can extend the toolbox available to teachers.

Future investigations will also need to focus on to what extent the ratings scales are applicable to all types of student texts (e.g. texts in different genres). This investigation has not had such a focus, and findings from other rating scale development projects do indeed indicate a need for the development of rating scales that are genre specific (Glasswell et al., 2001; Glasswell & Brown, 2003).

About the authors

Gustaf B. Skar holds a Ph.D. in educational science from Stockholm university. His research interests are assessment and testing, writing assessment, writing instruction, writing intervention, writing theory, and school development.
Institutional Affiliation: Department of Teacher Education,
Norwegian University of Science and Technology, NTNU, 7491 Trondheim
E-mail: gustaf.b.skar@ntnu.no

Lennart Jølle (1976) is an associate professor at the Department of Teacher Education, NTNU. He has a Master in literature from UiT and a PhD in applied linguistics from NTNU. He is also the Head of Norwegian at the Department of Teacher Education, NTNU.
Institutional Affiliation: Department of Teacher Education,
Norwegian University of Science and Technology, NTNU, 7491 Trondheim
E-mail: lennart.jolle@ntnu.no

Arne Johannes Aasen is the director at the The Norwegian Centre for Writing Education and Research (The Writing Centre). His research interests include writing didactics and assessment.
Institutional Affiliation: The Norwegian Centre for Writing Education and Research, NTNU, 7491 Trondheim
E-mail: arne.j.aasen@ntnu.no

References

Aasen, A. J., & Skar, G. B. (2018). *Developing Scales and Investigating Writing Proficiency Among Fifth- and Eighth-Grade Students in Norway*.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). Modern English Publications/British Council/Macmillan.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment*

Quarterly: An International Journal, *2*(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

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

Camp, H. (2012). The psychology of writing development—And its implications for assessment. *Assessing Writing*, *17*(2), 92–105. https://doi.org/10.1016/j.asw.2012.01.002

Carlsten, C. T. (2016). *Lærerveiledning 1.–2. trinn VÅR. Norsk rettskrivings- og leseprøve for grunnskolen (Teacher's Guide 1–2 grade SPRING. Norwegian spelling and reading test for grades 1–10)*. Cappelen Damm.

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge University Press.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. F. (1999). *Dictionary of language testing*. Cambridge University Press.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Elbow, P. (1973). *Writing without Teachers*. Oxford University Press.

Engelhard, G. (2013). *Invariant Measurement*. Routledge.

Evensen, L. S., Berge, K. L., Thygesen, R., Matre, S., & Solheim, R. (2016). Standards as a tool for teaching and assessing cross-curricular writing. *The Curriculum Journal*, *27*(2), 229–245. https://doi.org/10.1080/09585176.2015.1134338

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. Routledge.

Gee, J. P. (2004). *Situated language and learning: a critique of traditional schooling*. Routledge.

Glasswell, K., & Brown, G. T. L. (2003). Accuracy in the scoring of writing: Study in large-scale scoring of asTTle writing assessments. In *asTTle Technical Report 26*. University of Auckland/Ministry of Education.

Glasswell, K., Parr, J. M., & Aikman, M. (2001). Development of the asTTle Writing Assessment Rubrics for Scoring Extended Writing Tasks. In *asTTle Technical Report 6*. University of Auckland/Ministry of Education.

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

Ivanič, R. (2004). Discourses of Writing and Learning to Write. *Language and Education*, *18*(3), 220–245. https://doi.org/10.1080/09500780408666877

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

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1–26. https://doi.org/10.17239/jowr-2008.01.01.1

Knoch, U. (2007). Do Empirically Developed Rating Scales Function Differently to Conventional Rating Scales for Academic Writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 1–36.

Knoch, U. (2009). *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale*. Peter Lang.

Linacre, J. M. (2017). *A user's guide to FACETS. Rasch-model computer programs. Program manual 3.80.0*. Hämtad 2017-05-25. http://www.winsteps.com/a/Facets-ManualPDF.zip

McNamara, T. F. (2000). *Language testing*. Oxford University Press.

Norwegian Reading Center. (2018). *Lesesenterets bokstavprøve (the Norwegain Reading Center's Letter Test)*. Lesesenterets Bokstavprøve (the Norwegain Reading Center's Letter Test). https://lesesenteret.uis.no/boeker-hefter-og-materiell/boeker-og-hefter/lesesenterets-bokstavprove-article104746-12686.html

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Puranik, C. S., & Lonigan, C. J. (2014). Emergent Writing in Preschoolers: Preliminary Evidence for a Theoretical Framework. *Reading Research Quarterly*, *49*(4), 453–467. https://doi.org/10.1002/rrq.79

Purves, A. C. (1992a). 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.

Purves, A. C. (Ed.). (1992b). *The IEA Study of Written Composition II: Education and Performance in Fourteen Countries*. Pergamon Press.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. The University of Chicago Press.

Rijlaarsdam, G., Van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Van Steendam, E., & Raedts, M. (2012). Writing. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook, Vol 3: Application to learning and teaching.* (pp. 189–227). American Psychological Association. https://doi.org/10.1037/13275-009

Rose, D. (2016). New Developments in Genre-Based Literacy Pedagogy. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 227–242). The 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

Scribner, S., & Cole, M. (1978). Literacy without Schooling: Testing for Intellectual Effects. *Harvard Educational Review*, *48*(4), 448–461. http://www.metapress.com/content/F44403U05L72X375

Skar, G. B. (2017). *The Norwegian National Sample-Based Writing Test 2016: Technical Report*. Nasjonalt senter for skriveopplæring og skriveforsking. http://www.skrivesenteret.no/uploads/files/Skriveproven2017/NSBWT2017.pdf

Skar, G. B., & Aasen, A. J. (2018). Å måle skriving som grunnleggende ferdighet. *Acta Didactica Norge*, *12*(4), 10. https://doi.org/10.5617/adno.6280

Skar, G. B., & Jølle, L. (2017). Teachers as raters: Investigation of a long term writing assessment program. *L1 Educational Studies in Language and Literature*, *17*(Open Issue), 1–30. https://doi.org/10.17239/L1ESLL-2017.17.01.06

Skar, G. B., Thygesen, R., & Evensen, L. S. (2017). Assessment for Learning and Standards: A Norwegian Strategy and Its Challenges. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education* (pp. 225–241). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-50856-6_13

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing*, *17*(2), 81–91. https://doi.org/10.1016/j.asw.2012.02.001

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science, 103*(2684), 677–680. http://www.jstor.org/stable/1671815

Thygesen, R., Evensen, L. S., Berge, K. L., Fasting, R. B., Vagle, W., & Haanæs, I. R. (2007). *Nasjonale prøver i skriving som grunnleggende ferdighet. Sluttrapport.* Nasjonalt senter for leseopplæring og leseforskning, Universitetet i Stavanger.

Vähäpassi, A. (1988). The Domain of School Writing and Development of the Writing Tasks. In T. P. Gorman, A. C. Purves, & R. E. Degenhart (Eds.), *The IEA Study of Written Composition I: The International Writing Tasks and Scoring Scales* (Vol. 1, pp. 15–40). Pergamon.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Wheadon, C. (n.d.). *A Comparative Judgement Approach to the Large-Scale Assessment of Primary Writing in England*.

WIDA. (2017). *Speaking and Writing Interpretive Rubrics*. Speaking and Writing Interpretive Rubrics. https://wida.wisc.edu/sites/default/files/resource/Speaking-Writing-Interpretive-Rubrics.pdf

Appendix A

	Entering writing education	**Establish familiarity**	**Develop knowledge**	**Expanding knowledge**	**Reaching for the next year levels**
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.
Vocabulary	The text consists of individual letters/words/characters/drawings.	The text contains some few words that are not particularly theme-related.	The text contains several different words, a lot of them theme-related.	The text contains a repertoire of words and expressions that are relevant to the task.	The text contains a repertoire of words and expressions that are relevant to the task. In some cases, there is use of specialized and abstract words, and/or creative forms of expression.

	Entering writing education	**Establish familiarity**	**Develop knowledge**	**Expanding knowledge**	**Reaching for the next year levels**
Organization of content	The text consists of individual letters/words/characters/drawings.	-The text may indicate a structure, for example 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 the additive connector "and" and the temporal connector "so."	The text (verbal and optionally drawing) 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.	The text has a complete global structure with several elaborated elements arranged in a logical or otherwise appropriate order. The text may show examples of comparisons, classifications, chronology. The text contains connecters that are used suitably and purposefully.
Language use	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 show appropriate variation in sentence structure.	The text has for the most part appropriate syntactic variation, and it has some developed phrases and/or paragraphs.
Punctuation	The text does not use 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.

	Entering writing education	**Establish familiarity**	**Develop knowledge**	**Expanding knowledge**	**Reaching for the next year levels**
Spelling	There may be letters in the text and/or there may be scribbles (imitating writing).	The text contains letter combinations and single words. Unstable use of spaces	The words are spelled phoneticallly, and some high-frequency words are written correctly. There is use of space between words.	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	Letters they may be difficult to decipher (if any).	The text contains letters that are not crafted in a conventional manner.	The letters are crafted in a convention al manner.	The letters are crafted in a conventional manner. Instances of conventional use of "bokstavhuset."	The letters are drafted in a conventional and legible manner. For the most part, there is conventional use of "bokstavhuset." Usually follows conventions for use of upper- and lower- case
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 approximately two to three sentences.	The part of the verbal text that is a relevant answer to the task correspond s to appr. half an A4 page.	The part of the verbal text that is a relevant answer to the task corresponds to approximately an A4 page.	The part of the verbal text that is a relevant answer to the task corresponds to approximately one and a half A4 pages or more.

Appendix B: Bonferroni corrected t-tests of differences in text quality with semester as a factor

Texts from main step #2

Texts from step #2: Descriptive statistics

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1st semester	130	-3,5892	1,76394	,15471	-3,8953	-3,2831	-8,99	,96
3rd semester	92	,4965	1,73207	,18058	,1378	,8552	-6,52	3,95
4th semester	33	1,8130	1,43676	,25011	1,3036	2,3225	-2,91	5,16
5th semester	58	1,5047	1,87523	,24623	1,0116	1,9977	-4,96	5,41
6th semester	61	4,1705	2,02889	,25977	3,6509	4,6901	-1,32	8,01
Total	374	-,0519	3,34962	,17320	-,3925	,2887	-8,99	8,01

Note. Ratings expressed on a logit scale with min score = -8.99 and max score = 8.01

Texts from step #2: Multiple Comparisons Bonferroni-corrected t-tests

(I) Semester	(J) Semester	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1st semester	**3rd semester**	-4,08575*	,24445	,000	-4,7761	-3,3954
	4th semester	-5,40226*	,34973	,000	-6,3899	-4,4146
	5th semester	-5,09389*	,28331	,000	-5,8940	-4,2938
	6th semester	-7,75972*	,27845	,000	-8,5461	-6,9734
3rd semester	**1st semester**	4,08575*	,24445	,000	3,3954	4,7761
	4th semester	-1,31651*	,36406	,003	-2,3446	-,2884
	5th semester	-1,00813*	,30082	,009	-1,8577	-,1586
	6th semester	-3,67397*	,29625	,000	-4,5106	-2,8374
4th semester	**1st semester**	5,40226*	,34973	,000	4,4146	6,3899
	3rd semester	1,31651*	,36406	,003	,2884	2,3446
	5th semester	,30838	,39122	1,000	-,7964	1,4132
	6th semester	-2,35746*	,38772	,000	-3,4524	-1,2625
5th semester	**1st semester**	5,09389*	,28331	,000	4,2938	5,8940
	3rd semester	1,00813*	,30082	,009	,1586	1,8577
	4th semester	-,30838	,39122	1,000	-1,4132	,7964
	6th semester	-2,66584*	,32905	,000	-3,5951	-1,7366
6th semester	**1st semester**	7,75972*	,27845	,000	6,9734	8,5461
	3rd semester	3,67397*	,29625	,000	2,8374	4,5106
	4th semester	2,35746*	,38772	,000	1,2625	3,4524
	5th semester	2,66584*	,32905	,000	1,7366	3,5951

* The mean difference is significant at the 0.05 level.

Note. Ratings expressed on a logit scale with min score = -8.99 and max score = 8.01

Texts from step #7, #9

Texts from step #7, #9: Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
first_sem	152	-3,3668	1,89539	,15374	-3,6706	-3,0631	-7,03	,50
third_sem	222	-,0892	1,08463	,07280	-,2327	,0542	-2,39	3,64
fourth_sem	32	,3031	1,13421	,20050	-,1058	,7121	-1,77	3,15
fifth_sem	108	,7939	1,48371	,14277	,5109	1,0769	-3,36	5,05
sixth_sem	59	1,8924	1,66555	,21684	1,4583	2,3264	-2,36	7,19
Total	573	-,5663	2,31373	,09666	-,7561	-,3764	-7,03	7,19

Note. Ratings expressed on a logit scale with min score = -7.03 and max score = 7.19

Texts from main step 2: Multiple Comparisons Bonferroni-corrected t-tests

(I) semester	(J) semester	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1st semester	**3rd semester**	-3,27761*	,15547	,000	-3,7157	-2,8395
	4th semester	-3,66997*	,28723	,000	-4,4794	-2,8605
	5th semester	-4,16073*	,18586	,000	-4,6845	-3,6370
	6th semester	-5,25921*	,22652	,000	-5,8976	-4,6209
3rd semester	**1st semester**	3,27761*	,15547	,000	2,8395	3,7157
	4th semester	-,39236	,27925	1,000	-1,1793	,3946
	5th semester	-,88312*	,17326	,000	-1,3714	-,3949
	6th semester	-1,98161*	,21631	,000	-2,5912	-1,3720
4th semester	**1st semester**	3,66997*	,28723	,000	2,8605	4,4794
	3rd semester	,39236	,27925	1,000	-,3946	1,1793
	5th semester	-,49076	,29723	,993	-1,3284	,3469
	6th semester	-1,58925*	,32422	,000	-2,5029	-,6756
5th semester	**1st semester**	4,16073*	,18586	,000	3,6370	4,6845
	3rd semester	,88312*	,17326	,000	,3949	1,3714
	4th semester	,49076	,29723	,993	-,3469	1,3284
	6th semester	-1,09848*	,23908	,000	-1,7722	-,4247
6th semester	**1st semester**	5,25921*	,22652	,000	4,6209	5,8976
	3rd semester	1,98161*	,21631	,000	1,3720	2,5912
	4th semester	1,58925*	,32422	,000	,6756	2,5029
	5th semester	1,09848*	,23908	,000	,4247	1,7722

* The mean difference is significant at the 0.05 level.

Note. Ratings expressed on a logit scale with min score = -7.03 and max score = 7.19

Appendix C

Related to each scale, the panelist judged the following claims on a six-point scale:

A. I perceive the scale to be relevant.

B. I perceive the scale to be sufficiently elaborated.

C. I believe that the scale should be deleted for time-saving reasons (regardless of how I judge the relevance and sufficiency of the scale).

1 = Strongly Disagree, 2 = Moderately Disagree, 3 = Slightly Disagree, 4 = Slightly Agree, 5 = Moderately Agree, 6 = Strongly Agree.