Skrivesenterets skriftserie 2

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for samfunns- og utdanningsvitenskap
Institutt for lærerutdanning

Gustaf B. Skar

The Norwegian National Sample-Based Writing Test 2016: Technical Report

Trondheim, August 2017

Skrivesenteret
Nasjonalt senter for skriveopplæring og skriveforsking

NTNU
Kunnskap for en bedre verden

Skrivesenterets skriftserie 2
The Norwegian National Sample-Based Writing Test 2016: Technical Report
Gustaf B. Skar

Skrivesenteret
Nasjonalt senter for skriveopplæring
og skriveforsking

www.skrivesenteret.no

ISBN 978-82-93194-19-4 (pdf)

Omslagsfoto: Skrivesenteret
Layout: Skrivesenteret

Forord

Nasjonalt senter for skriveopplæring og skriveforsking skal bidra til at den nasjonale utdanningspolitikken blir iverksatt og gjennomført slik at alle barn, unge og voksne kan få en likeverdig og tilpasset opplæring av høy kvalitet i et inkluderende fellesskap. I oppdragsbrevet som Skrivesenteret mottar fra Utdanningsdirektoratet står det at senteret skal «samle, systematisere og formidle resultater fra forsknings- og utviklingsarbeid til sektorene» og at senteret «skal kjenne til og bruke et bredt og relevant eksisterende kunnskapsgrunnlag i sitt arbeid». Som et ledd i å utføre disse oppdragene gjennomfører Skrivesenteret utviklingsarbeid og lager kunnskapsoversikter. Disse presenteres i Skrivesenterets skriftserie, som består av tekniske rapporter og meldinger.

Denne tekniske rapporten presenterer analyser av den nasjonale utvalgsprøven i skriving som grunnleggende ferdighet i 2016. Formålet med rapporten er todelt. For det første skal den vurdere egenskapene ved selve målingen av elevenes skriveferdigheter. Det betyr at den undersøker påliteligheten og gyldigheten til skriveprøvene. For det andre presenterer rapporten vurderinger av kvaliteten på norske elevers skriving på femte og åttende årstrinn. Rapporten er fagfellevurdert.

Resultatene fra utvalgsprøvene i skriving som grunnleggende ferdighet, kan oppsummeres slik:

- En gjennomsnittlig 5. klassing kan skrive en tekst som viser forsøk på å tilpasse teksten til mottakeren som er oppgitt i oppgaveformuleringen. Innholdet er stort sett relevant for oppgaven som er gitt og er ofte stilt opp assosiativt eller uten logisk rekkefølge. Teksten har gjerne lite variasjon i setningsstruktur, og den kan være preget av muntlig språk.
- En gjennomsnittlig 8. klassing kan skrive en tekst som delvis er tilpasset mottakeren som er oppgitt i oppgaveformuleringen. Innholdet er relevant for oppgaven som er gitt. En gjennomsnittlig tekst kan vise begynnende kompleks setningsstruktur. Begreper og formuleringer kan være presise, og i noen tekster kan en finne bruk av språklige virkemidler.

Denne beskrivelsen av hva gjennomsnittselevene presterer, tar imidlertid ikke høyde for den store variasjonen i elevens skriveferdigheter som rapporten dokumenterer. Denne variasjonen er gjennomgående mellom skoler og mellom grupper av elever. Det vil for eksempel si at gjennomsnittsgutten presterer lavere enn hva denne beskrivelsen tilsier, og tilsvarende at gjennomsnittsjentene presterer bedre. Avstandene mellom kjønnene er så store at jentene ligger over et og et halvt års skolegang foran guttene både på femte og åttende årstrinn (se s. 28 i rapporten). Dette indikerer at forskjellene er varige gjennom grunnskolen og videre at en betydelig stor del av guttene har skriveferdigheter som gjør det vanskelig å nå kompetansemålene i norskfaget både etter fire og sju år i skolen. Det betyr igjen at for mange av disse vil skriveferdighetene være så lite utviklet at skriving trolig ikke er et verktøy for læring i alle fag.

Ved siden av store forskjeller mellom kjønnene, dokumenterer rapporten store forskjeller i elevprestasjoner mellom skoler. Det vil si at hvilken skole en elev går på, er en sterk predikator for hvordan eleven presterer på prøven. Imidlertid innebærer den store variasjonen at det finnes skoler der det ikke er påvist signifikante forskjeller mellom gutter og jenter. Tatt i betraktning resultatene fra NORM-prosjektet (Berge et al., 2017) kan en mulig årsak til dette være at skriveopplæringen på disse skolene er annerledes. Dette innebærer at utvikling av kvaliteten på skriveopplæringen vil kunne redusere forskjellene i gutters og jenters prestasjoner betraktelig.

Rapporten er resultat av et mangeårig arbeid som er gjennomført av flere personer. Takk til Hege Kjeldstad Berg, Kjell Lars Berge, Pia Farstad Eriksen, Lars S. Evensen, Anne Holten Kvistad, Siri Natvig og Jorun Smemo.

Arne Johannes Aasen
Senterleder, redaktør

Gustaf B. Skar
Prosjektleder, redaktør

1. The National Sample-Based Writing Test

National tests to measure students' writing were launched in 2005 as a governmental response to concerns that students were not receiving adequate instruction in so-called "key competencies" (Official Norwegian Report (green paper): NOU 2002, p. 10). The 2005 test was administrated to the whole population of grade 5, 8, 10, and 11 students and rated by students' own teachers. An evaluation of the writing test demonstrated low rater reliability (Lie, Hopfenbeck, Ibsen, & Turmo, 2005), and the writing test project was discontinued by 2006. In that year, however, the notion of key competencies was formalized through the school reform "The Knowledge Promotion" (Norwegian Directorate for Education and Training, 2007; cf. Organisation for Economic Co-operation and Development, 2005). Writing was named as one of five key competencies, meaning that it was to be taught within all school subjects (the other four key competencies included Information and Communications Technologies (ICT) skills, mathematical skills, oral skills, and reading). Almost overnight, all teachers in Norway became writing teachers; however, teachers as well as the government lacked the tools to evaluate student progress within these competencies.

To resolve this issue of lacking tools, in 2010, the Norwegian Directorate for Education and Training commissioned the National Writing Center (NWC) to develop the national sample-based writing test (NSBWT) and the formative writing assessment package (FWAP). The NWC was also charged with establishing a national panel of raters (NPR) comprised of teachers; its goal was a panel that would reliably rate the NSBWT. The NSBWT was to be annually administrated to a nationally representative sample of grade 5 and 8 students. The results and material would form the basis for the FWAP, which included the NSBWT tasks, annotated exemplar texts representing different student proficiencies, and information about the "national level" of student writing proficiency. The national level was equal to the results of the NSBWT.

The last NSBWT was administrated in the fall of 2016. The participants were 950 students from 62 schools who answered a total of seven writing prompts. In addition, all of the teachers who administrated the test were surveyed about their perceptions of how well the tasks functioned. did their job.

This technical report describes the test development and results. Specifically, it will answer the following questions:

- What were the results of the teacher survey?
- What was the measurement quality of the NSBWT?
- What were the results of the NSBWT in general and for groups of students?

The report is organized as follows. Section 2 presents a brief note on some of the theoretical and pedagogical underpinnings of the NSBWT and

FWAP. Sections 3 and 4 describe the tasks and the rating scales, respectively. Section 5 presents the data and methodology. Sections 6–8 focus on the results from the teacher questionnaire, the statistical analysis, and the students' scores on the NSBWT. Section 9 ends the report with a brief conclusion concerning the overall results.

2. Theoretical and Pedagogical Underpinnings

The definition that underpins the tasks and rating scales developed here is represented in the theoretical model the wheel of writing (Berge, Evensen, & Thygesen, 2016), which is depicted in Figure 2.1. According to this model, writing should be understood from a functional perspective, meaning that writing can be thought of as a purposeful act of meaning making. The outer layer of the wheel of writing describes six different "acts of writing" (to convince, to describe, to explore, to imagine, to interact and to reflect). The next layer describes six "purposes of writing" (persuasion, knowledge organization and storing, knowledge development, creation of text worlds, exchange of information and identity formation) that commonly give rise to such acts. For example, if the communicative purpose is to convince somebody to think or act in a certain way, it is common to engage in persuasive writing. However, the two layers do not have locked positions. This illustrates that a writing purpose does not necessarily lead to a specific act of writing. There are many examples of persuasive texts that are stories (cf. the writing act to imagine) rather than persuasions, as a writer has deemed it more efficient to tell a story than to engage in the act of persuasion. In the inner circle of the writing wheel lies the resources that writers make use of when writing. These include grammar, morphology, vocabulary, different text-structuring techniques, and manual tools such as pens and paper.

Because of this definition, all NSBWT tasks engage students in specific acts of writing by asking them to fulfill a communicative purpose through an act of writing. Each script is rated on five rating scales. The rating scales are designed to tap into students' control of the reader–writer relationship, content, text structure, and language use, which all relate closely to specific purposes of writing and specific acts of writing. Moreover, a separate rating scale for generic competencies (i.e., coding competencies) has been developed to tap into students' spelling, punctuation, and grammar proficiency.

In the NSBWT, writing proficiency is defined as the proficiency to engage in an act of writing using necessary mediating tools. This definition is further elaborated elsewhere (Skar, Evensen, & Iversen, 2015; Aasen & Skar, 2017). It follows from this definition that in order to adequately estimate a student's writing proficiency (in terms of width and breadth), one needs to administer several tasks focusing on different writing acts (width), with each student script rated on each rating scale (depth). This conclusion is supported by empirical research suggesting that students

perform differently on different writing tasks (Bouwer, Béguin, Sanders, & van den Bergh, 2015; Skar & Berge, 2017).

The FWAP rests on insights from assessments of learning research (e.g., Black & Wiliam, 1998; Hattie & Timperley, 2007); it stresses the need for teachers to have adequate assessment literacy (Brookhart, 2011) so they can gather and interpret information about students' (writing) proficiency, leading to better-informed instruction (Black & Wiliam, 2009).

The close relationship between the NSBWT and the FWAP has impacted how NSBWT scores have been constructed. Most importantly, scripts in the NSBWT are scored analytically rather than holistically (cf. Weigle, 2002), meaning that every script receives a score on each rating scale. It was believed that analytical scoring would boost teachers' skills in assessing text and analyzing the instructional needs of students. For example, analytical scoring could yield results where most students scored highly on one rating scale (e.g., coding competencies), but low on another (e.g., text structure), which could lead to better-informed writing instruction. It was also the belief that the rating scales and scale descriptors in the NSBWT and the FWAP would offer teachers across subject disciplines a common language for talking about writing and writing proficiency.

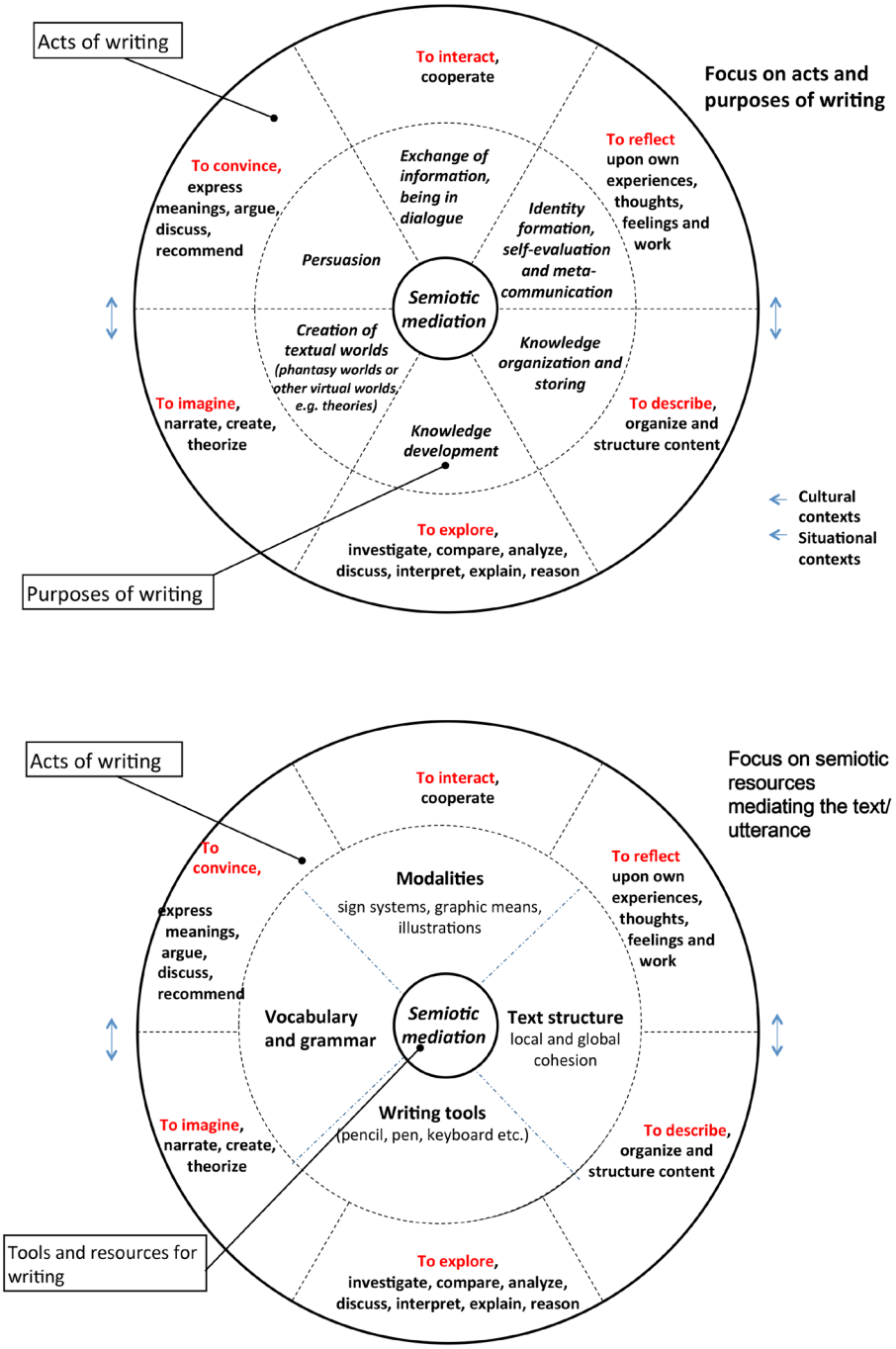


Figure 2.1 The Wheel if Writing. The top shows the entire wheel, while the bottom focuses on acts of writing and semiotic resources.

3. NSBWT Tasks and Administration

The tasks were developed according to the task development directions stated in the (unpublished) test development framework (TDF). They were then piloted and adjusted based on those pilots.

The TDF was jointly written by the Norwegian Directorate for Education and Training and the Writing Centre and was in part constructed to operationalize the abovementioned theoretical and pedagogical underpinnings. The TDF specifies the following:

- All tasks should be thematically related to the national curriculum.
- All tasks should be based on and specify one of five acts of writing and its associated default purpose of writing (cf. the wheel of writing). (The act of writing "to interact" was not part of the TDF.)
- All tasks should state a writing context, one or several recipients, and a content domain.
- All tasks should include discussion points for a "brainstorming session."

For example, the following task was administrated in the NSBWT in 2014. It was based on the writing act to describe and had "school life" as a content domain: "A childhood friend is returning to Norway after having lived abroad since 1st grade. S/he has asked you about everyday life in your school. You answer your friend and describe everyday life in school." (emphasis added). There were several "brainstorming questions" targeting the act of writing (e.g., "What does it mean to describe?") as well as the purpose and recipient (e.g., "What would someone like this recipient need to know about school in Norway?"). The purpose of the task was to structure already known facts about everyday school life in a descriptive text. The recipient was a childhood friend and the context was a personal interaction. The task was related to several parts of the curriculum (cf. the 2015 FWAP).

Seven tasks were developed to comprise the 2016 NSBWT. Two of them ("Home Place" and "Helmet") were piloted on a large sample (see Skar & Iversen, 2016), while the remaining five were piloted on samples of 50–100 students. In the piloting phase, tasks were administrated by teachers and rated by members of the NPR. The teachers and raters were then surveyed (through a questionnaire) for their opinions about how well the tasks functioned. All seven tasks went through several stages of drafting, piloting, and re-drafting before being included in the NSBWT 2016 (the writing acts are bolded):

- *Animal Police*: Norway is now trying out an animal police force that will intervene when people hurt or neglect animals. This will be tried out for a few years before a decision is made about whether Norway should have an animal police force in the future. Write a text to the Prime Minister in which you try to **convince** her that the animal police scheme should continue.

- *Architect:* You have won a competition where the prize is a house. This house will be built exactly the way you want it. Before it can be built, an architect has to draw a plan showing what the inside and outside of the house should look like. Write a text in which you **describe** your dream house in such a way that the architect can draw it.
- *Helmet*: A cycle helmet can help to lessen injuries in the event of an accident. Still, only just over half of all adults use a helmet when they ride a bicycle. Write a text to try to **convince** adult cyclists that they have to wear cycle helmets. Your text is going to be printed in a traffic safety leaflet that will be distributed to the parents of all the pupils in your school.
- *Home Place*: What is special about the place where you live? A family with children your age is going to move there. They have never been there before and they would like to read about the place before they arrive. Write a text in which you **describe** the place where you live to a family that is going to move there.
- *Remark:* You were late for a test because you had to help a boy who had fallen off his bike and hurt himself. Your teacher is of the opinion that it is important to be on time for tests and wants to give you a late mark. Write a text in which you try to **convince** your teacher not to give you a late mark.
- *Substitute Teacher:* Your PE teacher is off sick for a few weeks. The supply teacher who takes over does not normally teach PE and knows little about what you do in PE lessons. There is a particular activity that you enjoy doing. Write a text in which you **describe** this activity to the supply teacher so that he can learn the rules before the next PE lesson.
- *Super Power*: You can choose a superpower. Write a text in which you **imagine** how your superpower works and what you would do with it. It should be an entertaining text for the class to read.

In total, three tasks are based on the writing act to persuade ("Animal Police," "Helmet," and "Remark"), three tasks are based on the act to describe ("Architect," "Home Place," and "Substitute Teacher"), and one task is based on the act to imagine ("Superpower"). While the set of tasks did not fully represent the wheel of writing, the writing acts to reflect and to explore were featured in earlier NSBWTs and were therefore available in earlier editions of the FWAP.

According to the TDF, teachers at participating schools should administrate the NSBWT in the following way. Teachers inform students about the task, including how it is going to be assessed and used. Teachers and students then conduct a brainstorming session lasting for 15 minutes. During this session, teachers use the blackboard to take notes, and students are allowed to use a pen and paper. Students then write by hand or on a computer (depending on what the teacher perceives to be suitable for that particular group of students) for a maximum of 45 minutes. The NSBWT is accompanied by guidelines that explain this and other practical aspects (e.g., how to send their completed texts to the Writing Centre).

4. The Rating Scales

A novelty with the NSBWT 2016 is that the students in grade 5 and grade 8 were given the same tasks. Another novelty was that students were rated on the same rating scales, which was done to accommodate analyses comparing grade 5 and grade 8 students. New rating scales were developed to this end.[1] The rating scales built on the prior NSBWT scales (see Skar & Iversen, 2016), which in turn were related to rating scale development within the project "Developing National Standards for the Assessment of Writing – A Tool for Teaching and Learning" (Berge et al., 2017; Evensen, Berge, Thygesen, Matre, & Solheim, 2016). There are five rating scales in the NSBWT 2016:

- Writer–reader interaction (WRI): This measures to what extent the writer communicates with the reader in a relevant manner and to what extent the writer can accommodate the reader's information needs.
- Content: This measures to what extent the content is relevant to the task specification and to what extent it is elaborated and weighted.
- Text structure: This measures the degree to which a text has a structure that is relevant and logical to the communicative context specified in the task.
- Language use: This measures sentence variation and language precision.
- Coding competencies: This measures task-indifferent coding competencies, namely, grammar, spelling, and punctuation.

The rating scales were developed in the following way (see also Holten-Kvistad & Skar, 2016). In the fall of 2015, a group of senior researchers, teachers, and test developers met for a two-day rating scale development session (RSDS). The researchers, who were leading scholars in the Nordic countries, had backgrounds in applied linguistics, language testing, literature, and language education. The teachers had a great deal of experience teaching and assessing writing in grades 5–10. The test developers worked at the Writing Centre. Researchers, teachers, and test developers were assigned to different groups, each focusing on one particular rating scale.

During the RSDS, each group was presented with several texts written by students in grades 5–8. As much as possible, the texts were sampled from pools of texts where students across grade levels had completed the same writing task; however, additional texts from other tasks were included. Each group read each script and negotiated a ranking of the scripts. Thereafter, the scripts were sorted into seven piles ranging from "best" to "worst." Each group then described the typical features of the texts in each pile. The groups were instructed to reference their personal experiences of students' texts not included in the current group if a particular and important feature was not among those typical in the pile. This work resulted in descriptors for seven bands or proficiency levels.

After the RSDS, the test development team made adjustments to the descriptors so that they would be linguistically and logically harmonized

across rating scales. Then, members of the NPR tried out the rating scales (a trial that was reported to the Norwegian Directorate for Education and Training in an unpublished technical report). A many-faceted Rasch measurement (MFRM) model (see below) showed that not all of the rating scale levels functioned perfectly, which prompted the test development team to condense the rating scales; the new scales had five proficiency levels. Following this, the rating scales were piloted in schools under the supervision of test developers. Additional adjustments were made, ending in a final trialing session that included teachers and researchers. This session led to some minor adjustments before the rating scales in their present form were presented to and used by the NPR in the fall of 2016. All of the rating scales are presented in Appendix A.

To increase the reliability and to enable comparisons between groups of test takers, the final NSBWT score is an average of the scores on all five rating scales. Previous analyses support such an action, given that the rating scales are highly correlated (e.g., Skar & Iversen, 2016). The average score does not match a single scale descriptor. However, the Writing Centre has developed NSBWT proficiency profiles (NPPS) that match NSBWST final scores with features of texts that are typical for that score (the five NPPS are presented in Appendix B).

5. Data and Methodology

5.1 Participants in the NSBWT 2016

The Directorate for Education and Training sampled primary schools (grades 1–7) and secondary schools (grades 8–10) to recruit participants for the NSBWT. The sampling procedure was intended to generate a sample that was representative for the grade 5 and grade 8 population. Schools in the sample were contacted by the Writing Centre and asked to participate. Almost all of the schools agreed to participate (although some schools eventually failed to do so). The participating schools were, in accordance with the TDF, instructed to let students of class "A" (e.g., 5A, 8A) sit for the NSBWT, which included two tasks that were to be administrated within a two-week period. The headmasters of each school were instructed to appoint a teacher that would administrate the test; these teachers received instructions and surveys. The administrating teachers were encouraged to include all of the class members in the test, but they were also instructed to leave out students that for obvious reasons would not be fit to participate under the test circumstances (e.g., students that are unfit for participating in a 45 min long test, students with severe dyslexia).

All of the participating schools sent the final texts to the Writing Centre, which was forced to conduct additional sampling because of economic constraints. A limit of 950 students was set. These 950 students represented 62 schools. Of these, 475 were grade 5 students from 30 schools and 475 were grade 8 students from 32 schools. It was decided to randomly choose 475/30 students per class from grade 5 and 475/32 stu-

Table 5.1. Participants

		Girls		Boys		BM		NN	
		N	%	N	%	N	%	N	%
Grade 5	Sample	248	52	227	48	418	88	57	12
	Population	30,812	49	32,427	51	55,094	88	7,599	12
Grade 8	Sample	245	52	230	48	417	88	58	12
	Population	30,067	49	31,554	51	53,868	88	7,289	12

Note. Population data from grunnskolens informasjonsystem (GSI) (https://gsi.udir.no). There are additional language forms in Norwegian school. GSI also lists Samisk (grade 5, N = 105; grade 8, N = 64) and "others" (grade 5, N = 441; grade 8, N = 413). For grade 8, GSI reports a mismatch of 13 students between N for girls and boys (61,621) and N for language forms (61,634). Bokmål = BM, nynorsk = NN.

dents from grade 8. Because some schools had small classes, other classes were represented by more students than the original quota.

Table 5.1 presents the sample in detail and includes a comparison with population parameters. The sample comprised 52% girls for each grade, while the corresponding population proportion was 49% girls. The next column concerns written forms of Norwegian, of which there are two: Bokmål (BM) and Nynorsk (NN). Students have either BM or NN as their primary written form. There were equally sized proportions of BM users in the sample and the population (88%). Lastly, there were a majority of students in both grades that had Norwegian as their first language (L1) (grade 5 = 87%; grade 8 = 86%). These proportions are not included in the table due to the lack of official data against which to compare them.

While the sampling as such may have resulted in a representative sample, there are at least some factors calling for caution when generalizing the results to the population. The sampling was done in such a way that the overall results are generalizable to the population; any sub-group results need to be interpreted with this in mind.

5.2 NSBWT task distribution and rating design

There were seven tasks in the NSBWT, with each student completing two tasks. Two independent raters scored each paper on all five rating scales, and each student had different raters for each paper. Student texts that were so poor that no descriptions from a particular rating scale matched the quality received a zero (0) on that particular rating scale. Each of the 71 NPR members was given text packages of 60–62 fully anonymous student papers (i.e., with no disclosure of gender, grade, school, and L1/L2) (all rating scales are presented in Appendix A).

To compare the students while controlling for differences in task difficulty and rater severity, one would ideally have employed a fully crossed design where all students sat for all tasks and were rated by all raters on all rating scales. With the "true rating model," this would have resulted in 47,500 ratings (950 students x 2 tasks x 5 rating scales x 5 raters). As this would have been too expensive and too tiresome for the five raters[2],

it was decided to make use of the robust MFRM model, which allows the researcher to control for variables while collecting data in a way that intentionally leaves empty cells in a data set. The only prerequisite is that subjects are linked to each other. Table 5.2 illustrates the linking design. See Appendix C for the full design.

Tables 5.3 and 5.4 display the number of schools and number of students per task. It should be noted that the distribution was not perfectly even across tasks. The number of students per task was somewhat lower than expected from the design, but there was still a substantial and sufficient number of students per task (nine students completed only one task, seven students completed only task 1, and two students completed only task 2).

An "incomplete" but "connected" design was used for the ratings (Eckes, 2015). Figure 5.1 illustrates the basic principles of the rating design. Raters were given text packages with 60–62 texts. Seven of these texts were "linking texts" ensuring that all of the raters were comparable. The other 53–55 texts were randomly drawn from the pool of texts from grade 5 and grade 8 students. In other words, the texts were randomly distributed to the raters, and the raters scored the texts from both grades. In addition, the raters were blind to grade, gender, and L1/L2, but they

Table 5.2 Number of Schools Per Task

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
School A	X	X					
School B	X		X				
School C	X			X			
School D	X				X		
School E	X					X	
School F	X						X
School G		X	X				
School H		X		X			
School I		X			X		
School J		X				X	
School K		X					X
School L	X	X					
...							
School $_{x1}$	X						X
School $_{x2}$		X					X
School $_{x3}$			X				X
School $_{x4}$				X			X
School $_{x5}$					X		X
School $_{x6}$						X	X

Table 5.3 Number of Schools Per Task

	Animal Police	Architect	Helmet	Home Place	Remark	Substitute Teacher	Super Power	Total
Grade 5	8	9	9	9	9	10	6	60
Grade 8	8	9	11	10	8	8	10	64
Total	16	18	20	19	17	18	16	124

Table 5.4 Number of Students Per Task

	Animal Police	Architect	Helmet	Home Place	Remark	Super Power	Substitute Teacher	Total
Grade 5	116	137	142	135	152	103	162	947
Grade 8	120	135	168	143	118	148	112	944
Total	236	272	310	278	270	251	274	1891

Note. Nine students completed only one task, seven students completed only task 1, and two students completed only task 2. See Appendix C for the full design.

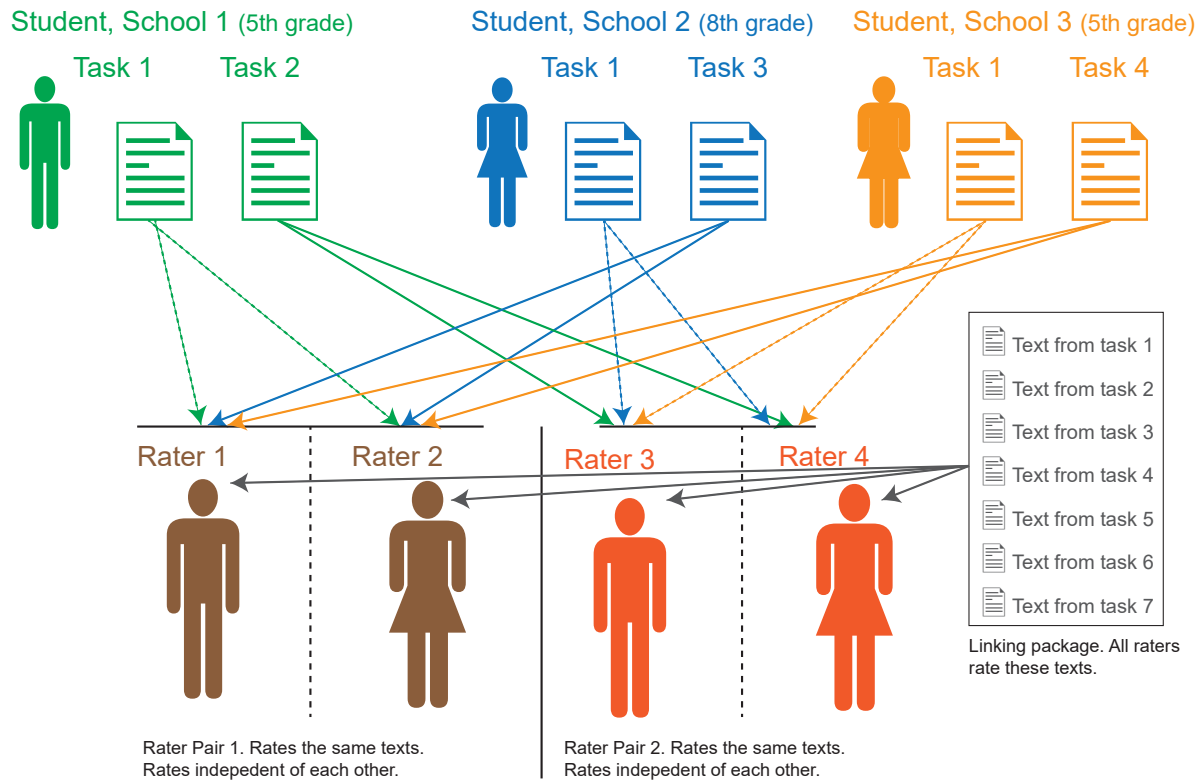


Figure 5.1. Principal rating design.

were presented with the candidates' written form of Norwegian in order to use the coding scale correctly.

5.3. Teacher survey: Participants and analysis

Teachers from 60 schools participated in the teacher survey. Teachers handed in one questionnaire per task. In two schools, more than one teacher handed in a survey. The questionnaire contained claims that were rated on a four-point scale Likert-scale with the following categories: "totally disagree," "somewhat disagree," "somewhat agree," and "totally agree." The questionnaire was identical to questionnaires administrated in previous years and included the following claims:

- Q1: This particular task gives students the opportunity to display writing proficiency.
- Q2: The oral instruction that I am to read to students is understandable.
- Q3: The written instruction that students receive is understandable.
- Q4: The students start to write quickly.
- Q5: The theme of the task is relevant.
- Q6: My students were motivated to write about this theme.
- Q7: There is enough time for task completion.
- Q8: The teacher instruction has an appropriate length.
- Q9: The teacher instruction informs me about how I shall administrate the test.
- Q10: The NSBWT material is readily understood.

The results of the questionnaire can be informative of aspects of test administration that threaten the validity of the interpretation and use of the test scores. For example, if teachers found it difficult to agree with Q1 or Q5, they may have failed to engage the students enough for them to make their best effort (which, for example, was the case in one observation during the 2014 NSBWT pilot). Likewise, if teachers were reluctant to agree with all or any of Q4, Q6, or Q7, the results of the NSBWT might have to be interpreted with slightly more caution. If, for example, the time for task completion is perceived to be too little, then the time constraint of 45 minutes might have introduced "irrelevant variance" (Messick, 1996).

The teacher questionnaire was analyzed using descriptive statistics. All subjects were included in the analysis, even though this meant that two schools were represented by more than one teacher per task. Section 6 presents the proportions of respondents opting for each different alternative.

5.4 Analyzing the teachers' ratings

Two main analyses of the teachers' ratings were carried out. First, the raw scores were analyzed using an MFRM model as well as a traditional rater reliability analysis. Second, scores from the MFRM analysis were used to analyze score patterns for the whole test population and for sub-group

scores. The MFRM analysis is presented below. The score analysis is also presented, but in less detail.

In the basic Rasch model (Rasch, 1980), the probability of a correct answer to a dichotomous item is a function of the difference between test taker ability and item difficulty. The MFRM extends this premise, allowing the researcher to model and control for the impact of additional facets such as rater severity and scale step difficulty (Linacre, 2017b); this model is therefore suitable in "messy" (language) assessment situations where scores are contingent on human qualitative judgment (Barkaoui, 2014; Eckes, 2015; McNamara, 1996). In addition, and as mentioned above, the MFRM model is superb whenever data points are missing by design.

The following MRFM model (Engelhard, 2013; Linacre, 2017b) was used in this analysis:

$$\log(Pnmijk/Pnmijk{-}1) = Bn - Dm - Ei - Cj - Fx,$$

where Pnmijk represents the probability of student n on task m, rating scale i, by rater j receiving a score of k, and Pnmijk-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, Dm is the difficulty of task m, 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.[3]

The analysis was carried out in the computer software Facets 3.8 (Linacre, 2017a). When fitting writing assessment data to the MFRM model, Facets performs a logistic transformation of raw scores, creating a linear scale (Engelhard, 2013). This scale, called the logit scale, is common for all elements of all facets (individual students, raters, etc.), and it is graphically depicted in the variable map. Moreover, the facets are disentangled from one another. For example, the severity of a particular rater is not dependent on which students he or she rated.

The Facets output includes several useful graphs and statistics, of which the following will be reproduced in this report. First, Facets produces statistics related to data–model fit. It is possible to investigate "global fit" to get an estimate of overall data–model fit. Since this estimate always shows that data deviates from the model, Eckes (2015) and Linacre (2013) suggest that researchers should instead inspect standardized residuals. According to Linacre (2017b), there is reasonable fit when the proportion of standardized residuals ≥3.0 is less than 1% and when the proportion of standardized residuals ≥2.0 is less than 5%. Another way to evaluate overall fit is to inspect Facets' visual output (see Engelhard & Wind, 2013). This report presents the category probability function graphic and the test characteristic curve (TCC). The category probability function graphic depicts the relationship between category difficulty and person logit values. The TCC illustrates the modeled and empirical relationship between the person logit value and NSBWT score. When the data fit the model, the empirical observation matches the model's predicted values.

Measures of data model fit—infit and outfit—indicate to what extent the data fit the model. A good fit indicates that the model can predict, for example, score patterns and rater behavior. The two indices of infit and outfit indicate to what extent the model can predict raw score observations. The model-expected value is 1.0, while the underfit (i.e., deviation from the MFRM model) is indicated when the fit statistic exceeds this. Underfit may indicate that items are operationalizing another construct or that raters are scoring idiosyncratically. Overfit values indicate less-than-optimal variation (e.g., a rater that is restricted in his or her use of the rating scale or uses a redundant item). Fit values in the range of 0.50–1.50 can be accepted (see Bond & Fox, 2015), but this report pays attention to fit values that drop below 0.75 or exceed 1.30.

The variable map (or Wright map) provides visual information on the extent to which raters share levels of severity. Interpretation of the map is aided by different separation statistics, which estimate the possibility of separating elements of facets (e.g., people/examinees, raters, rating scales) into different severity levels. First, the fixed (all same) chi-square tests the hypothesis that all raters share a certain severity level. Second, strata can be interpreted as the number of statistically distinct classes of severity (Eckes, 2015). Third, reliability provides an estimate of the precision of the separation, with a ceiling value of 1.00. The person reliability measure is analogous to Cronbach's alpha, or test reliability.

The Facets output also generates statistics of category use, which can be used in conjunction with separation statistics to gain insights into how the scale steps function. The Rasch-Andrich thresholds are category thresholds where two categories are of equal probability for a given person logit value. The Rasch-Andrich thresholds should increase with the category number. The outfit statistics report on relationships between expected and observed values. When the data are perfectly modeled, the average person logit measure for a person equals the observed average logit measure for a person in that category. The outfit value should not exceed 2.0.

The output also reports percent agreement and correlations of single rater–rest of raters (SR–ROR), indicating to what extent raters rank students in a similar fashion. In the results section below, this information is supplemented by traditional rater reliability estimates in the form of intra-class correlation coefficients (ICCs) (McGraw & Wong, 1996).

Table 5.5 lists the measurement quality indices used in this report. For technical descriptions of these indices, please refer to Eckes (2015), Knoch (2007), Linacre (2013), Myford and Wolfe (2003, 2004, 2009), Schumacker and Smith (2007), and Skar och Iversen (2015).

Finally, Facets reports the fair average measure, which is a measure that transforms the logit value back to the original NSBWT scale. The fair average is the expected raw score average while controlling for rater severity, item difficulty, etc. As said, in the NSBWT, students received scores on five rating scales (on two tasks by two raters). This amounted to 20

observations, which were used in the MFRM analysis. The fair average is an adjusted average of these observations.

The fair average was used in the major analysis of the results as well as the sub-group analysis. An anchoring procedure was used to describe and analyze student performance on individual tasks or rating scales. In accordance with recommendations in Eckes (2015, pp. 109–110), the analysis used the following procedure: First, logit measures based on all data were produced. Second, an anchor file with logit values for all of the elements of all facets not subject to special interest was used to generate new measures (e.g., to get estimates of student proficiency levels on the WRI scale, an MFRM analysis with anchored values for raters, scale steps,

Table 5.5 Measurement Quality Indices

All Facets	Index	Type	Explanation
	ASR	Absolute standardized residuals	Global estimate of data–model fit. Rule of thumb: standardized residuals >2.0: max 5%, standardized residuals >3.0: max 1%.
	Infit/outfit	Data–model fit	Differences between observed and expected values. Significant infit exceeding 1.30 indicates that data do not fit the model. Significant infit below 0.75 indicates overfit, or too predictable score patterns or items (e.g., a rating scale dependent on another rating scale).
	Q	Homogeneity index	A chi-square statistic that tests the assumption that elements of Facets do not display significant differences. When the statistic is significant, at least two elements are significantly different.
	G	Separation index	The number of statistically distinct classes of elements within a facet.
	R	Reliability index	The reliability of the separation, analogous to Cronbach's alpha. For people, it can be interpreted as test reliability.
Raters	Exact agreement	Percent exact agreement	Percent of all agreement opportunities where raters agree.
	SR–ROR	Correlation (SR–ROR)	How the single raters' ratings correlate with their peers. Values between .30–.70 are generally perceived as acceptable.
Scales	Rasch-Andrich thresholds		Logit value where two adjacent categories are of equal probability. The Rasch-Andrich thresholds should increase with increasing scale values.
	outfit		Relationship between expected and observed average measures for category. Should not exceed 2.0.
	Descriptive statistics		Reports category use. All categories should have at least 10 observations.

and tasks was used, while all raw score observations on the other rating scales were neglected.

5.5 Analyzing the scaled scores

The results section presents descriptive statistics for all students and for groups of students. Differences between groups of students have been computed using t-tests to test for significance and effect sizes to investigate magnitude. More specifically, Cohen's *d* has been computed.

Potential differences between the following groups have been investigated: boys and girls, BM writers and NN writers, and L1 and L2 speakers of Norwegian. In the case of boys and girls, the analysis has been expanded to include a hierarchical multiple regression analysis using fair average as a dependent variable. This analysis builds on school/class as a dummy variable to investigate whether there were gender differences when controlling for writing instruction context (i.e., class/school). The multiple regression analysis met the assumptions of normally distributed residuals, established the linear relationship between residuals, and predicted fair average values. It also met the assumption of homoscedasticity. The diagnostics did not reveal problems with multicollinearity or outliers.

6. Teacher Survey

The results of the teacher survey are presented in a number of figures, with some results explicitly commented on. For repetition, the teacher survey included ten statements about the test and test administration:

- Q1: This particular task gives students the opportunity to display writing proficiency.
- Q2: The oral instruction that the teacher is to read to students is understandable.
- Q3: The written instruction that students receive is understandable.
- Q4: The students start to write quickly.
- Q5: The theme of the task is relevant.
- Q6: My students were motivated to write about this theme.
- Q7: There is enough time for task completion.
- Q8: The teacher instruction has an appropriate length.
- Q9: The teacher instruction informs me about how I shall administrate the test.
- Q10: The NSBWT material is readily understood.

The overall results are presented in Figure 6.1. It can be seen that the vast majority of respondents "agreed" or "totally agreed" with all of the statements. Between 70–75% of all respondents ticked one of these two alternatives. However, items Q1 and Q6, which are of particular interest, show somewhat troublesome tendencies. Twenty-one respondents (or 17.4%) ticked "somewhat agree" to the statement that "this particular task gives student opportunity to display writing proficiency," and 32 respondents (or 27%) ticked "somewhat agree" or "totally disagree" that "my students were motivated to write about this theme."

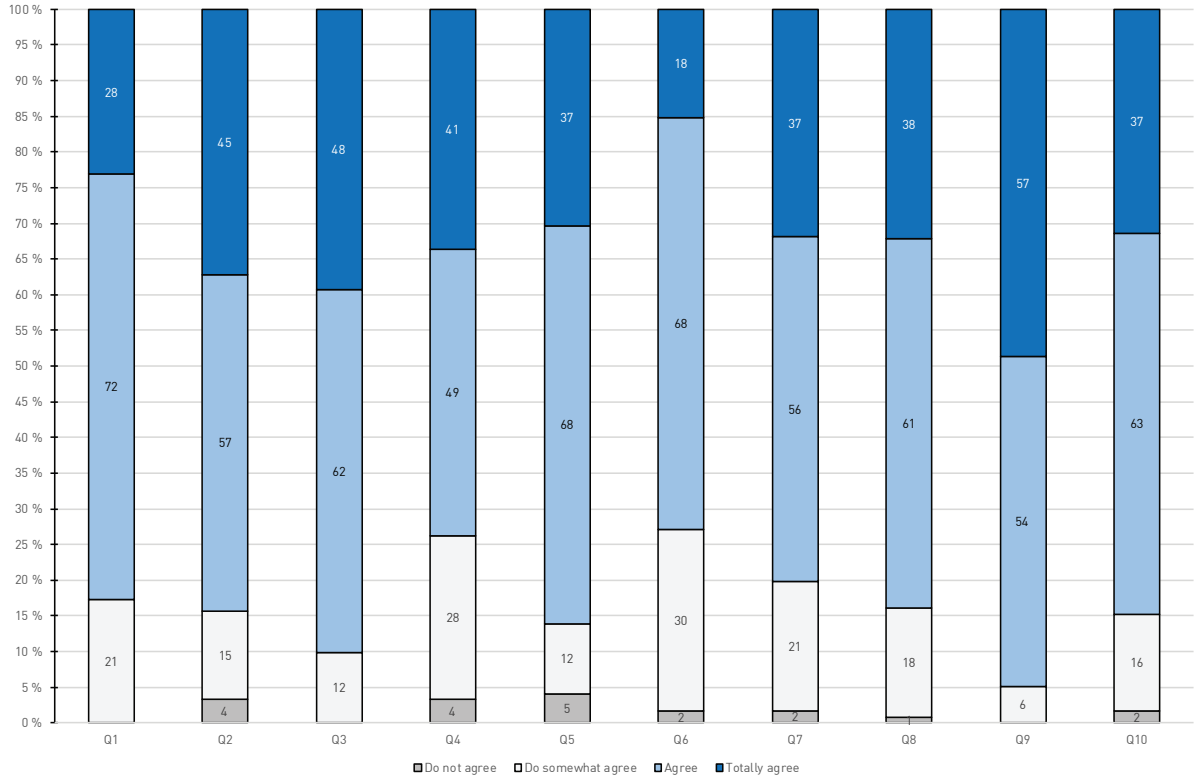


Figure 6.1. The overall results of the teacher survey. The numbers refer to the number of respondents for each category.

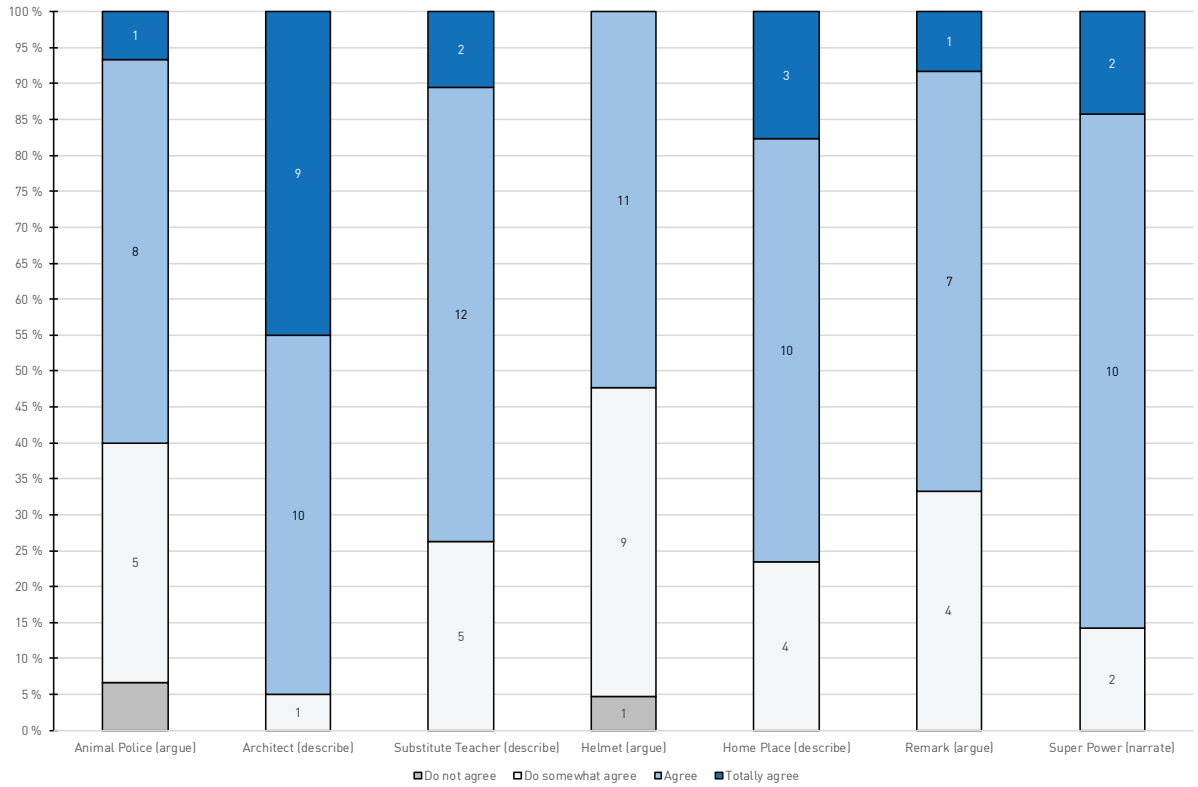


Figure 6.2. Item Q6 at the task level. The numbers refer to the number of respondents for each category.

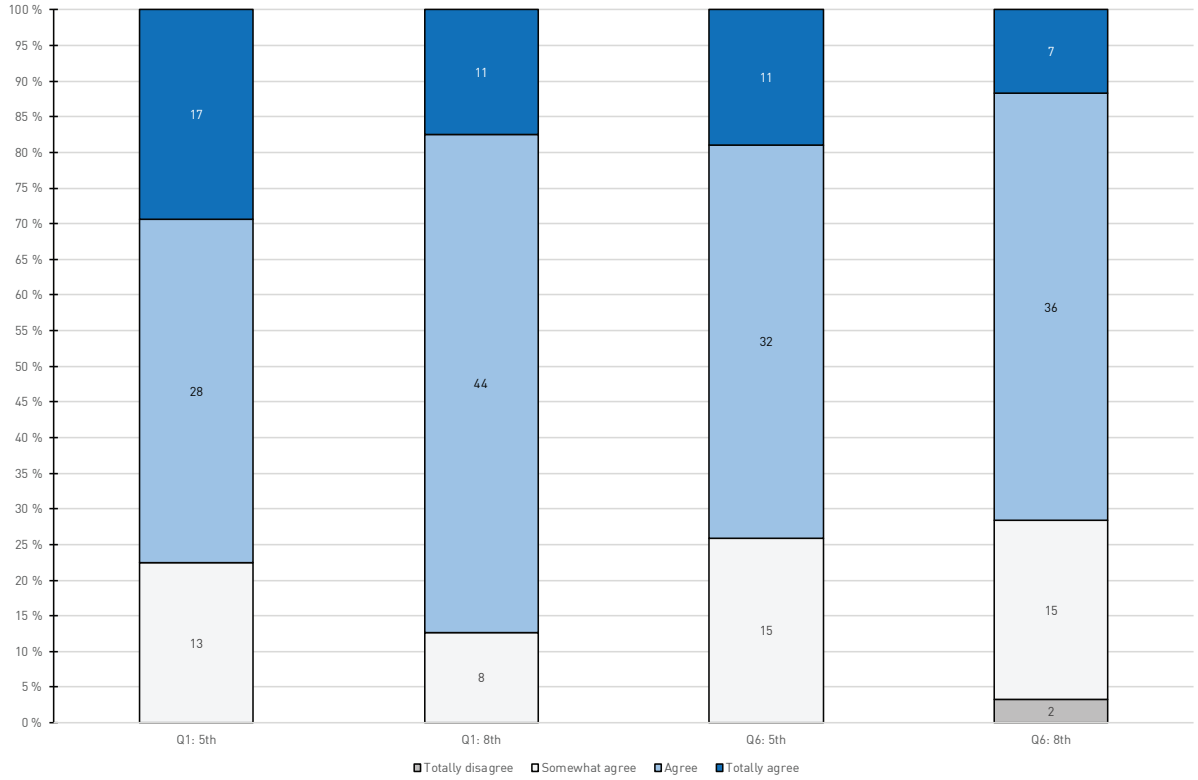


Figure 6.3. Items Q1 and Q6 at the grade level. The numbers refer to the number of respondents for each category.

Figure 6.2 presents the results for items Q1 and Q6 by grade level. Grade 5 teachers seem to be more concerned about each task's possibility to give students the opportunity to display writing proficiency than the grade 8 teachers were (see Figure 6.2). The results do not indicate any noteworthy differences in teachers' perception of students' motivation between grade 5 and grade 8.

Analyzing the motivational aspect further, Figure 6.3 presents the results for Q6 at the task level. The results indicate that teachers perceived students who sat for "Architect" and "Super Power" to be quite motivated. The tasks "Animal Police" and "Helmet" did not seem to be as motivating.

The results from the questionnaire indicate that teachers, on average, perceived that the NSBWT task and administration procedures work. The results also indicate that any interpretation of student proficiency should be done bearing in mind that several factors impact the way students write. Teachers in grade 5 did not agree as much as their grade 8 peers about task relevance for showing proficiency, a result that very well may be related to how familiar grade 5 students are with solving these kinds of tasks. Regarding motivation, one can suspect that "Superpower" and "Architect" fold themselves into a long tradition of creative writing, while the perceived lack of motivation for "Animal Police" and "Helmet" may be related to the fact that few students have any relationship to animal police and that grownups' use of helmets may not be perceived as a problem.

7. Measurement Quality

This section first presents estimates of global data–model fit, and then it continues with the reliability indices from the MFRM analysis. For the raters, the additional results from the ICC analysis and fit statistics are given. Finally, the descriptive and fit statistics for the tasks, rating scales, and categories are provided.

The overall fit was acceptable; of the standardized residuals, 4.4% had a value exceeding 2.0, and 0.5% had a value exceeding 3.0%. This was, in other words, well within the margins (please refer to Table 5.5). Consulting the category probability function and TCC graphs (Figure 7.1), some difficulty with modeling behavior at the ends can be noted. Particularly, this was true at the lower end of the latent trait scale. This difficulty is probably due to a smaller number of observations in category "0." To gain a better picture of this, Figure 7.1 also includes the graphical output from an analysis where the zeros have been removed. As can be seen, this action seems to increase the overall data–model fit. In the main analysis, however, the zeros have been kept.

The reliability indices are presented in Table 7.1. It shows that for all facets—students, raters, tasks, and rating scales—there were significant differences for at least two elements (cf. Q index). In fact, the G index for each facet indicates multiple distinguishable groups of proficiency, severity, and difficulty. The students seem to have formed 4.5 proficiency groups, and the rating scales formed close to 5 difficulty groups. The tasks were also of unequal difficulty, with 6.9 statistical groups. The reliability (which is functionally related to the separation index) indicates high replicability, with values exceeding .90 for all facets. The R-value for students can be interpreted as test reliability and was high, with .95.

That the raters were separated into different groups (5.7) with high reliability (.97) indicates that there were substantial and consistent differences in rater severity. From a traditional rater reliability perspective, this would indicate less-than-useful scores. In the MFRM context, however, this severity difference is modeled and taken into account. Of course, this has consequences for the interpretation of test reliability: While it was possible to produce reliable estimates of students' proficiency within this controlled context, the same would not necessarily be true if rater differences could not be controlled (e.g., in a classroom setting).

In terms of absolute agreement, the raters agreed in 39.8% of the cases. The average SR–ROR was 0.35. The traditional reliability index (the ICC), presented in Table 7.2, indicates acceptable levels of consistency, with an overall mean of .74. The highest rating scale mean was for coding, and the lowest was for WRI. This result is expected because rating with the coding scale would typically involve less judgment than rating with the WRI scale.

The fit analysis, which can be interpreted as an intra-rater reliability estimate, revealed that seven raters exhibited infit and outfit values

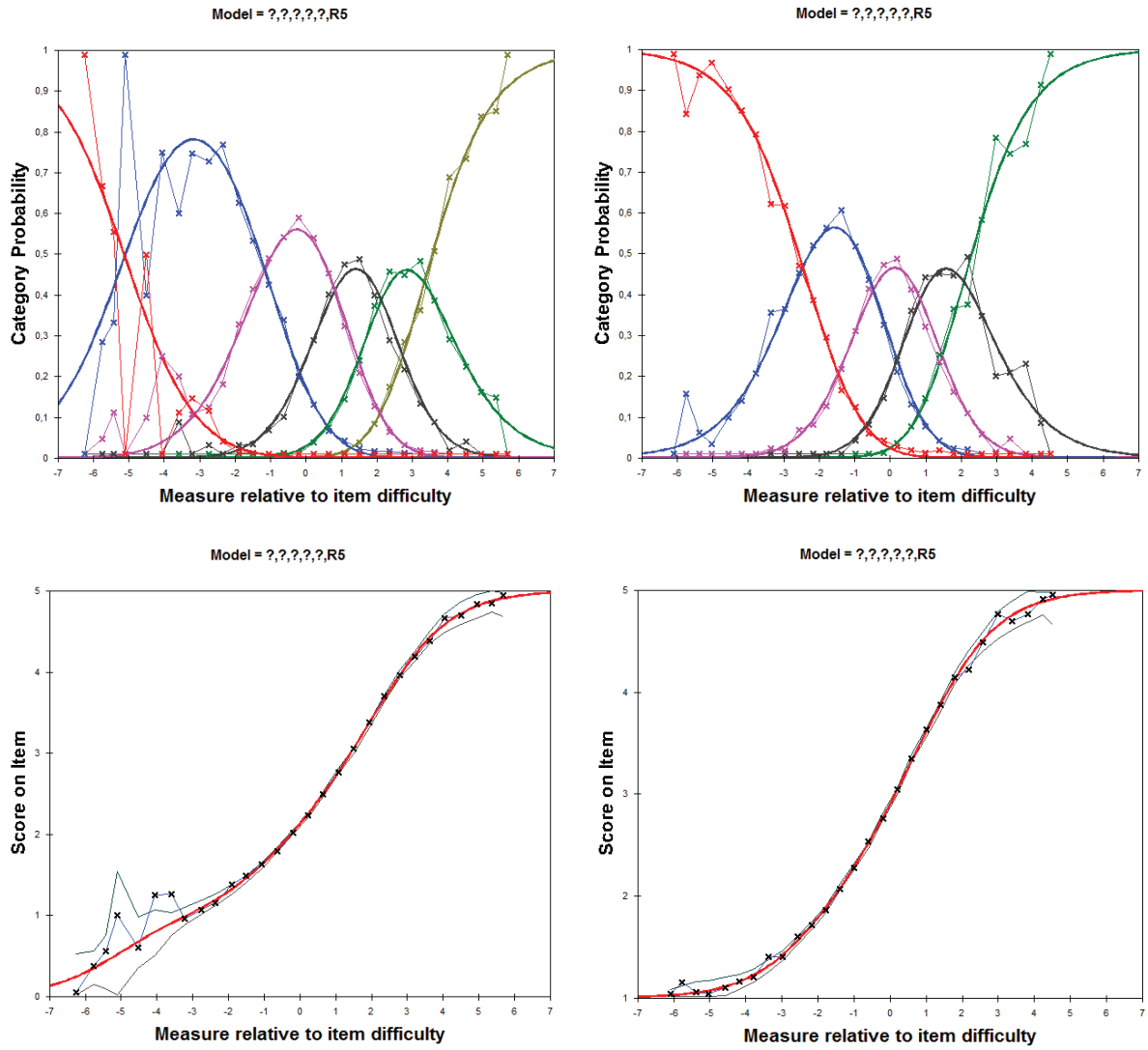


Figure 7.1. The upper left panel depicts the category probability function when all raw score data are included (from left to right, categories 0 [red]–5 [sepia]). The upper right panel depicts the category probability function when all of the zeros are removed (from left to right, categories 1 [red]–5 [green]). Similarly, the lower left and right panels depict the TTC with and without zeros, respectively.

exceeding 1.30. One rater exhibited infit and outfit values exceeding the threshold of 1.50. However, all of the raters were kept in the analysis. See Appendix D for the fit values for all of the raters.

While the G-value indicated distinct and large differences between the tasks, the fair score values indicated that the tasks were of quite similar difficulty. The phenomenon arose because it was possible to measure task difficulty with high precision. Table 7.3 gives the average values along with outfit measures. "Super Power" was the most difficult task (M = 2.41), and "Home Place" was the easiest (M = 2.66). These differences were modeled in the MFRM analysis. No tasks indicated poor fit.

As with the tasks, when inspecting differences in fair averages, the rating scales were also quite close to each other. Coding (M = 2.48) was the most difficult and WRI (M = 2.59) was the easiest. WRI showed somewhat poor fit (outfit = 1.37). This was also the rating scale in which the

Table 7.1. Reliability Indices from the MFRM Analysis

	RMSE	SD	Q (df)	G	R
Students	0.30	1.36	18678** (949)	4.48	0.95
Raters	0.08	0.44	2265.2** (70)	5.65	0.97
Tasks	0.02	0.17	291** (6)	6.90	0.98
Scales	0.02	0.10	99.2** (4)	4.87	0.96

Note. Root mean square error = RMSE, true standard deviation = SD, homogeneity index = Q, degrees of freedom = df, separation index = G, reliability (analogous to Cronbach's alpha) = R.
**p = < .01.

Table 7.2 . Intra-class Correlation Coefficient, Average Measures (Two-Way Mixed Effect Model)

	WRI	Content	Structure	Language	Coding	Overall
Min.	0.37	0.55	0.60	0.59	0.62	0.37
Max.	0.81	0.84	0.85	0.89	0.91	0.91
Mean	0.67	0.70	0.75	0.76	0.81	0.74

Note. WRI = writer–reader interaction, Language = language use, Coding = coding competencies.

Table 7.3. Tasks

	Fair Average	Outfit	Outfit_Z
Animal Police	2.49	0.99	-0.55
Architect	2.43	0.94	-2.67
Helmet	2.63	0.91	-4.04
Home Place	2.66	0.96	-1.8
Remark	2.58	1.17	6.33
Substitute Teacher	2.61	1.04	1.43
Super Power	2.41	1.01	0.52
Min.	2.41	0.91	
Max.	2.66	1.17	
M	2.54	1.00	

Table 7.4. Rating Scales

	Fair Average	Outfit	Outfit_Z
WRI	2.59	1.37	9
Content	2.63	1.1	4.67
Structure	2.49	0.8	-9
Language Use	2.52	0.78	-9
Coding	2.48	0.94	-2.67
Min.	2.48	0.78	
Max.	2.63	1.37	
Mean	2.54	1.00	

Note. WRI = writer–reader interaction, structure = text structure, language = language use, coding = coding competencies.

Table 7.5. Category Statistics

Category	Used	Proportion	Average	Expect	Outfit	Thresh-olds
0	109	1%	-3.09	-3.06	1.0	
1	3104	15%	-0.80	-0.87	1.1	-5.12
2	7236	34%	0.16	0.19	1.0	-1.16
3	6147	29%	1.08	1.10	1.0	0.82
4	3423	16%	2.02	2.02	0.9	2.13
5	1334	6%	2.84	2.84	1.0	3.34

Note. Average = average (logit) measure for category, expect = expected (logit) measure for category, outfit = outfit for category, thresholds = Rasch-Andrich thresholds.

raters exhibited the least amount of consistency (see Table 7.4).

The category statistics indicate that the categories functioned well (although category "0" was not used much). The average person logit value for each category increased as category values increased and was close to the expected value. The outfit values indicate good fit (but see Figure 7.1). The threshold values did also increased as the category values increased.

A potential impacting factor is the time of task administration. It can be assumed that students were less motivated to sit for task 2. However, a paired t-test showed non-significant differences between Time 1 (M = 2.65, SD = 0.85) and Time 2 (M = 2.61, SD = 0.88), with $t(940) = 1.6$, $p < .001$ and an effect size of $d = 0.04$ (SE = 0.05).

In summary, the measurement quality statistics indicate trustworthy and reliable results. All of the differences between raters, tasks, and rating scales have been modeled. The difference in time of task administration–difficulty has not been modeled, but this difference was non-significant.

8. NSBWT Results

This section presents the results of the NSBWT and compares them across groups. Figure 8.1 is the variable map or Wright map and illustrates the distribution of people, tasks, raters, and rating scales on the logit scale (the leftmost column) and the score zones on the raw score scale in the rightmost column. From the Wright map, students' proficiency measure seems to be normally distributed along the latent trait scale with a peak around 0.8 logits. The Wright map also indicates some of the spread in rater severity that was reported in Section 7 as well as some of the spread in task difficulty. The figure also shows that the scale steps of the raw score scale were of unequal length: the width of scale step "3," with logit values from 0.66–2.12, was much narrower than, for example, scale step "1," ranging from -5.15 to -1.37.

Using fair scores, Tables 8.1–8.4 report on the overall results for all of the test takers (M = 2.63, SD = 0.79) and for the sub-groups (group differences are presented further down). The overall result indicates that the average student performs somewhere between NSBWT proficiency profile

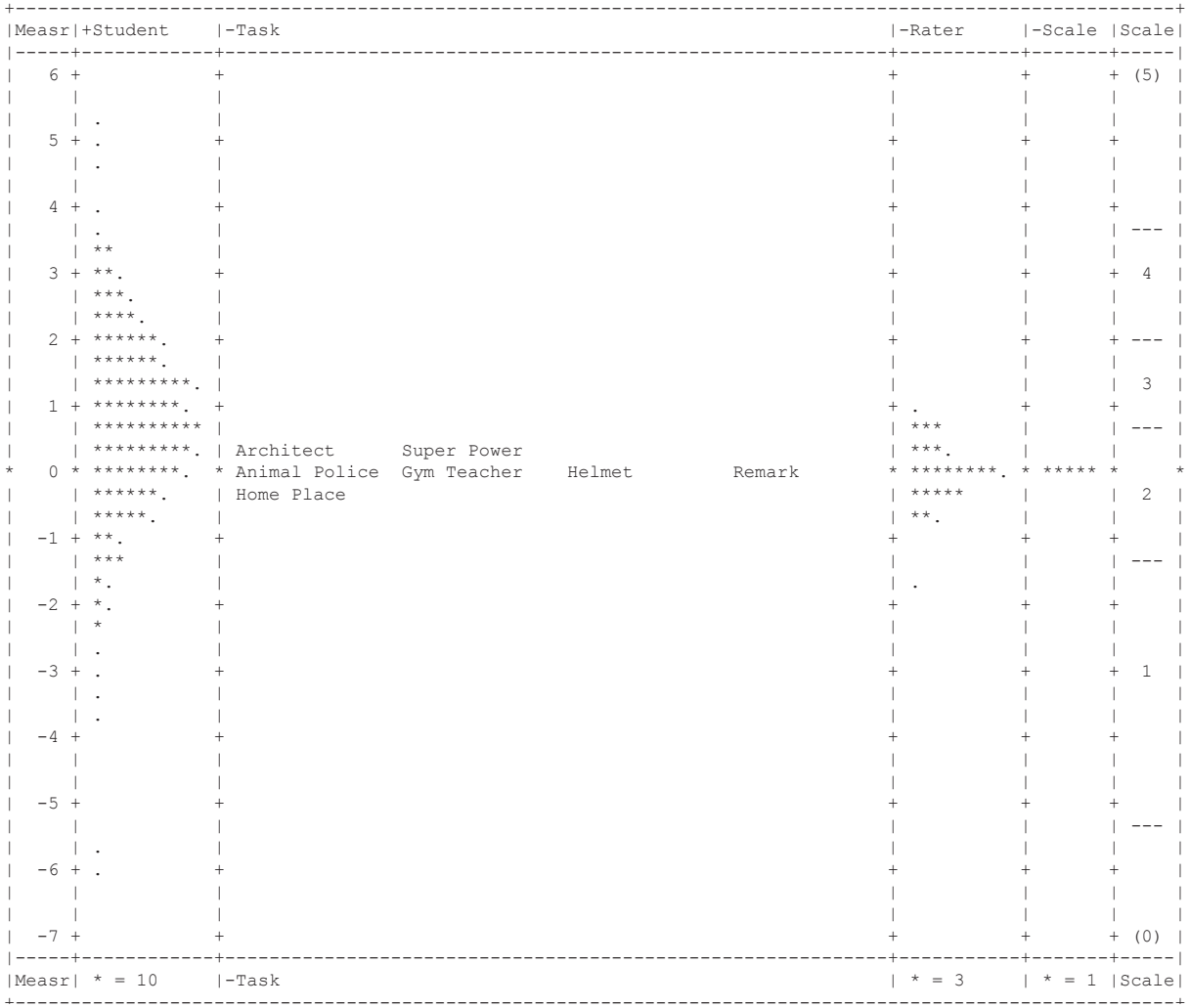


Figure 8.1. From left to right: the logit scale, the people (students), the tasks, the raters, the rating scales, and, in the rightmost column, the original scale.

2 and NSBWT proficiency profile 3 (NPP2 and NPP3, respectively, for short), but upon inspecting Table 8.1, one can see that grade 5 students are closer to NPP2 and grade 8 students are closer to NPP3. This becomes even more apparent when studying the score distribution in Figure 8.2, which shows how the two groups have separate means.

Based on the results, one can use the NPP wording to describe the average grade 5 and grade 8 students. Using the NPP2, the average grade 5 student text is as follows:

> The text shows attempts to adapt the text to the recipient mentioned in the assignment text. The content is mostly relevant to the assignment set. The text may show attempts at structuring, for example, by using bullet points, headings, an introduction, and/or conclusion. The content, however, is often structured in an associative manner or lacks a logical order. There is often little variety in syntax in the text, and it may be characterized by colloquial language. Elements of dialect may occur. More non-phonetic words are spelt correctly. The pupil tries to use punctuation marks other than full stops, for example question marks and exclamation marks. Commas are often used in lists, and the text mostly has capital letters in proper names and at

Table 8.1. Total Results and Results for Grade 5 and Grade 8 Students

		N	M	SE	SD
All students	Total	950	2.63	0.03	0.79
	Grade 5	475	2.27	0.03	0.61
	Grade 8	475	2.98	0.04	0.79

Note. N = number of students, M = arithmetic mean, SE = standard error of the mean, SD = standard deviation.

Table 8.2. Results for Girls and Boys

		N	M	SE	SD
All students	Girls	493	2.83	0.04	0.78
	Boys	457	2.41	0.03	0.75
Grade 5	Girls	248	2.46	0.04	0.63
	Boys	227	2.07	0.03	0.52
Grade 8	Girls	245	3.21	0.05	0.73
	Boys	230	2.74	0.05	0.78

Note. N = number of students, M = arithmetic mean, SE = standard error of the mean, SD = standard deviation.

Table 8.3. Results for Bokmål and Nynorsk

		N	M	SE	SD
All students	BM	835	2.63	0.03	0.79
	NN	115	2.60	0.08	0.83
Grade 5	BM	418	2.29	0.03	0.61
	NN	57	2.19	0.08	0.64
Grade 8	BM	417	2.98	0.04	0.79
	NN	58	2.99	0.11	0.80

Note. N = number of students, M = arithmetic mean, SE = standard error of the mean, SD = standard deviation, BM = bokmål, NN = nynorsk.

Table 8.4 Results for Students with Norwegian as First Language (L1) or Second Language (L2)

		N	M	SE	SD
Total	L1	819	2.68	0.03	0.78
	L2	131	2.31	0.07	0.78
Grade 5	L1	413	2.31	0.03	0.62
	L2	62	2.06	0.07	0.52
Grade 8	L1	406	3.06	0.04	0.75
	L2	69	2.54	0.11	0.91

Note. N = number of students, M = arithmetic mean, SE = standard error of the mean, SD = standard deviation, L1 = Norwegian as first language, L2 = Norwegian as second language.

the beginning of new sentences.

The average grade 8 student text is as follows:

The text is partly adapted to the recipient mentioned in the assignment text. The content is relevant to the assignment set. The text shows attempts at using structuring principles. For example, a non-fictional text can be structured in a way leading up to a main point. Paragraphs are often grouped by topic. The text may show the beginnings of complex syntax. Some of the more complex sentences may contain syntactic flaws. Wording and concepts may be precise. Most of the words in the text are spelt correctly, but the text contains og/å mistakes (confusing the infinitive marker "å" with the word for "and": "og"). The major punctuation marks (full stop, question mark, exclamation mark) are correctly used most of the time. In addition to commas in lists, the text may also use commas between complete sentences.

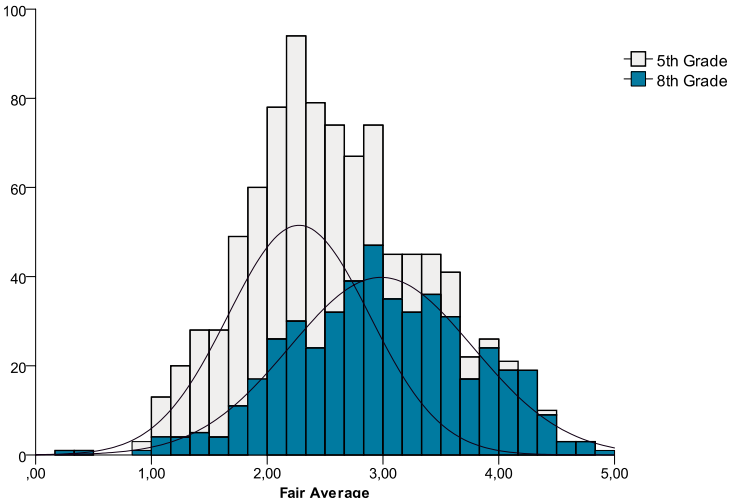
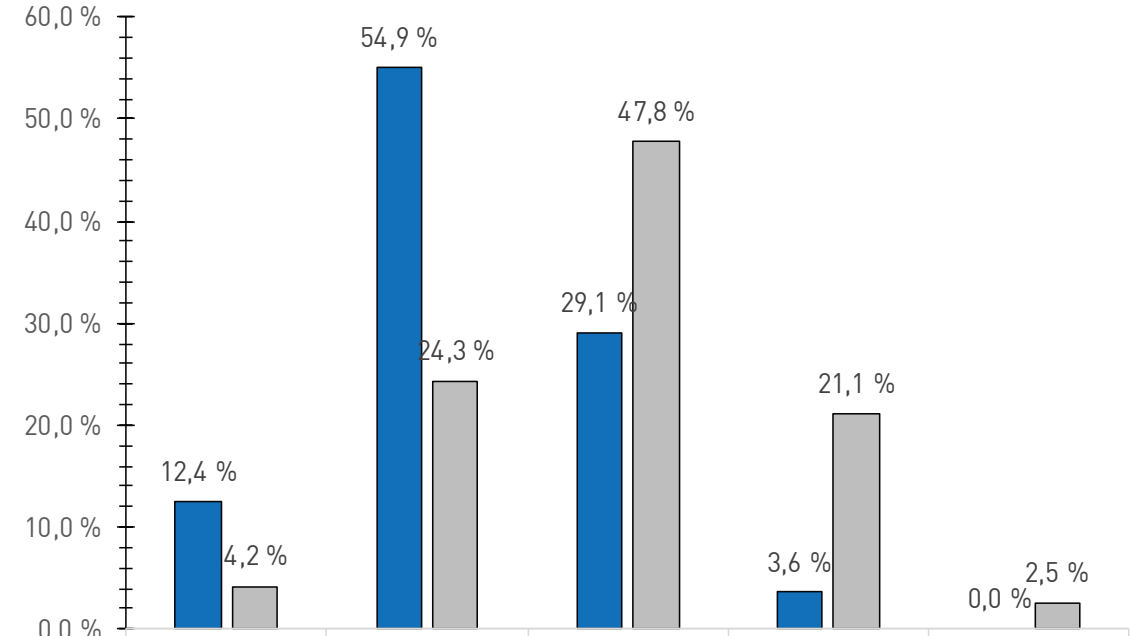


Figure 8.2. Score distribution for grade 5 and grade 8 students along with normal distribution.

As can be seen in Figure 8.3, there are differences at the ends of the NPP scale; grade 5 has proportionally more students at NPP1 and NPP2 than does grade 8, which in turn has more students at NPP4 and NPP5.

Table 8.2 indicates that there were consistent differences between girls and boys, with girls outperforming boys both on average and in each grade. Table 8.3 indicates no substantial differences between BM and NN users. Table 8.4 indicates that there were large differences between L1 and L2 writers, with L1 writers outperforming L2 writers.

Table 8.5 presents the results from t-tests and effect size computations that were done to investigate the potential significance and magnitude of the differences between sub-groups. The largest difference was between grade 5 students and grade 8 students, amounting to d = 1.00 (please refer to table for t-values and significance levels). On average, girls outperformed boys with a difference equaling d = 0.55. On average, the difference between girls and boys was half of that between grade 5 students and grade 8 students. Such a difference would equal 1.5 years of schooling (if



	M1	M2	M3	M4	M5
5. trinn	12,4 %	54,9 %	29,1 %	3,6 %	0,0 %
8. trinn	4,2 %	24,3 %	47,8 %	21,1 %	2,5 %

Figure 8.3. The percentage of students at each NPP based on grade level.

d = 1.00 equals three years of schooling). This difference is slightly bigger than that between LI and L2 writers, which may be related to female L2 writers (N = 72, M = 2.54, SD = 0.79) on average outperforming male LI writers (N = 398, M = 2.47, SD = 0.74). The latter difference was, however, non-significant, with $t(468) = 0.80$, $p = 0.44$, and $d = 0.10$. There was a trivial and non-significant difference between BM and NN users.

The differences between girls and boys on the one hand, and between LI and L2 writers on the other hand, were also significant within grades. In grade 5, the girls–boys difference amounted to $d = 0.66$, and in grade 8, the difference was $d = 0.62$. Bearing in mind the sampling procedure and that this was not repeated across years, this finding still indicates that the gender difference was more or less consistent. The LI–L2 difference was greater in grade 8 ($d = 0.67$ to $d = 0.47$ in grade 5), but for reasons of sample sizes, one should be extra careful with speculations about consistency.

Comparing gender difference across grades, there was a significant difference between grade 5 girls and grade 8 boys ($t(476) = -4.38$, $p < .001$), with an effect size of $d = -0.40$ (SE = 0.09). This difference was slightly smaller than the overall gender difference of d = 0.55 would suggest it to be, and it was quite a bit smaller than the overall difference between grade 5 and grade 8 students ($d = 1.00$). If the latter difference in this case can be said represent 3 years of schooling, then there was a difference of only 1.2 years between grade 5 girls and grade 8 boys, which was smaller than expected. There was also a significant difference between grade 5 boys and

Table 8.5. Differences Between Groups

		Levene's Test		Difference					Effect Size	
		F	p	M diff.	SE diff	df	t	p	D	SE
All	5 > 8	32.38	0.000	-0.71	0.05	891.48	-15.45	0.000	-1.00	0.07
	Girls > Boys	1.47	0.225	0.42	0.05	948	8.52	0.000	0.55	0.07
	BM > NN	0.53	0.467	0.04	0.08	948	0.49	0.623	0.05	0.62
	L1 > L2	1.01	0.314	0.37	0.07	948	4.97	0.000	0.47	0.09
5	Girls > Boys	4.644	0.032	0.39	0.05	468.13	7.28	0.000	0.66	0.09
	BM > NN	0.087	0.768	0.09	0.09	473	1.08	0.283	0.15	0.16
	L1 > L2	1.825	0.177	0.24	0.08	473	2.96	0.003	0.40	0.14
8	Girls > Boys	0.107	0.744	0.47	0.07	473	6.73	0.000	0.62	0.09
	BM > NN	0.08	0.777	-0.01	0.11	473	-0.07	0.945	-0.01	0.14
	L1 > L2	4.111	0.043	0.52	0.12	84.5	4.51	0.000	0.67	0.13

Note. p = significance level, M diff. = mean difference, SE diff. = standard error of the difference, df = degrees of freedom, t = t-value, d = Cohen's d, SE = standard error of d.

grade 8 girls ($t(470) = -19.28$, $p < .001$), with an effect size of $d = -1.77$ (SE = 0.11). With reference to the school year metric above, this difference represents 5.3 years; in other words, the girls from grade 8 performed as grade 10 students when compared to grade 5 boys.

To further explore potential differences, two one-way between-subject analyses of variance (ANOVAs)—one for each grade level—were conducted to test whether the variance was greater within or between schools/classes. Tables 8.6 and 8.7 present the school averages for grade 5 and grade 8. The analysis showed that variance was greater between schools for both grade 5 ($F(29.445) = 3.21$, $p < .001$) and grade 8 ($F(31.443) = 3.71$, $p < .001$). For grade 5, the school factor accounted for 17.3% of the variance in the fair average measure, and for grade 8, it accounted for 20.6%. In other words, there was a significant "school effect" or "class effect" (cf. sampling procedure), suggesting that factors relating to school or class can explain a substantial proportion of the results. These factors include, but are not restricted to, socio-economic factors and teaching factors.

A simple mean analysis indicated that gender differences were not consistent across schools/classes (see Tables 8.8 and 8.9). In order to gain more nuanced information about gender differences, simple hierarchical multiple regression analysis were conducted using gender and the schools as dummy variables.

For both grades, the first model (Model 1) was significant (grade 5: $F(29.445) = 3.21$, $p < .001$; grade 8: $F(31.433) = 3.71$, $p < .001$), with 17.3%

Table 8.6 Results by Grade 5 Schools

	N	M	CI 95%	SE	SD
School 501	5	2.97	2.10–3.83	0.31	0.70
School 519	5	2.84	2.08–3.59	0.27	0.61
School 521	15	2.79	2.49–3.10	0.14	0.56
School 502	19	2.70	2.32–3.08	0.18	0.79
School 524	18	2.62	2.29–2.95	0.16	0.67
School 509	14	2.53	2.26–2.79	0.12	0.45
School 508	13	2.51	2.14–2.88	0.17	0.61
School 516	17	2.48	2.14–2.82	0.16	0.67
School 505	18	2.48	2.19–2.77	0.14	0.58
School 522	18	2.42	2.11–2.72	0.15	0.62
School 512	6	2.38	1.99–2.77	0.15	0.37
School 520	21	2.34	2.01–2.67	0.16	0.72
School 517	20	2.32	2.08–2.55	0.11	0.50
School 530	13	2.27	2.01–2.53	0.12	0.43
School 507	10	2.26	1.69–2.83	0.25	0.79
School 515	18	2.26	2.01–2.51	0.12	0.50
School 527	15	2.25	1.94–2.57	0.15	0.57
School 531	18	2.23	1.95–2.51	0.13	0.57
School 534	16	2.21	1.80–2.62	0.19	0.77
School 511	20	2.21	1.98–2.43	0.11	0.49
School 503	16	2.17	1.91–2.43	0.12	0.49
School 532	18	2.15	1.87–2.42	0.13	0.55
School 518	19	2.14	1.92–2.36	0.10	0.46
School 525	18	2.10	1.92–2.29	0.09	0.37
School 529	18	2.08	1.76–2.40	0.15	0.64
School 533	18	2.04	1.80–2.28	0.11	0.48
School 523	19	1.95	1.67–2.24	0.14	0.60
School 513	13	1.93	1.71–2.16	0.10	0.37
School 504	18	1.88	1.67–2.10	0.10	0.44
School 514	19	1.81	1.54–2.08	0.13	0.56

Note. In the process of making the schools anonymous, they were arbitrarily assigned numbers (e.g., 501). In the table, the schools are sorted by their descending mean values.

Table 8.7. Results by Grade 8 Schools

	N	M	CI 95%	SE	SD
School 824	15	3.94	3.65–4.23	0.14	0.52
School 833	13	3.67	3.05–4.29	0.28	1.02
School 823	15	3.55	3.29–3.82	0.12	0.47
School 821	15	3.55	3.12–3.98	0.20	0.78
School 811	15	3.34	2.98–3.70	0.17	0.65
School 815	15	3.33	2.93–3.73	0.19	0.72
School 830	15	3.33	2.98–3.67	0.16	0.63
School 802	15	3.29	2.95–3.63	0.16	0.61
School 828	16	3.22	2.88–3.56	0.16	0.64
School 814	15	3.14	2.73–3.54	0.19	0.73
School 820	14	3.11	2.65–3.57	0.21	0.79
School 816	15	3.03	2.67–3.40	0.17	0.66
School 822	15	3.03	2.59–3.47	0.20	0.79
School 803	14	3.01	2.56–3.46	0.21	0.78
School 826	15	3.00	2.74–3.26	0.12	0.47
School 817	15	2.95	2.56–3.34	0.18	0.71
School 818	15	2.88	2.49–3.28	0.18	0.71
School 829	13	2.85	2.41–3.28	0.20	0.72
School 813	17	2.84	2.43–3.25	0.19	0.80
School 825	14	2.78	2.49–3.07	0.13	0.50
School 805	16	2.77	2.41–3.13	0.17	0.67
School 801	14	2.73	2.43–3.03	0.14	0.52
School 809	15	2.72	2.24–3.19	0.22	0.86
School 807	15	2.71	2.21–3.20	0.23	0.90
School 832	15	2.69	2.32–3.06	0.17	0.67
School 804	15	2.69	2.03–3.34	0.31	1.18
School 810	15	2.67	2.24–3.10	0.20	0.77
School 812	15	2.63	2.37–2.89	0.12	0.47
School 808	14	2.61	2.17–3.06	0.20	0.77
School 819	15	2.56	2.13–2.99	0.20	0.78
School 831	15	2.50	2.10–2.89	0.19	0.72
School 806	15	2.46	1.97–2.94	0.23	0.88

Note. In the process of making the schools anonymous, they were arbitrarily assigned numbers (e.g., 801). In the table, the schools are sorted by their descending mean values.

Table 8.8 . Gender Differences Within Schools, Grade 5

	Girls					Boys					Diff.
	N	M	SE	SD		N	M	SE	SD		M
School 507	7	2.68	0.18	0.48		3	1.30	0.24	0.42		1.38
School 501	4	3.19	0.29	0.57		1	2.09	.	.		1.10
School 519	3	3.20	0.27	0.47		2	2.30	0.18	0.25		0.91
School 502	12	3.01	0.17	0.58		7	2.17	0.33	0.86		0.84*
School 522	7	2.85	0.20	0.54		11	2.14	0.16	0.52		0.71*
School 516	8	2.84	0.26	0.74		9	2.16	0.14	0.42		0.67*
School 527	5	2.68	0.25	0.56		10	2.04	0.15	0.46		0.64*
School 524	11	2.86	0.20	0.66		7	2.25	0.19	0.51		0.61
School 523	10	2.22	0.22	0.69		9	1.66	0.10	0.30		0.55*
School 534	8	2.48	0.33	0.92		8	1.93	0.17	0.48		0.55
School 521	8	3.04	0.19	0.53		7	2.51	0.18	0.47		0.53
School 532	10	2.37	0.16	0.52		8	1.86	0.17	0.48		0.51*
School 520	11	2.58	0.23	0.77		10	2.08	0.19	0.59		0.51
School 530	10	2.38	0.09	0.29		3	1.92	0.40	0.70		0.45
School 503	9	2.37	0.16	0.47		7	1.92	0.15	0.40		0.44
School 509	8	2.70	0.17	0.47		6	2.29	0.13	0.33		0.41
School 529	9	2.27	0.18	0.55		9	1.89	0.23	0.69		0.39
School 518	9	2.33	0.18	0.55		10	1.96	0.09	0.27		0.37
School 513	4	2.17	0.21	0.42		9	1.83	0.10	0.31		0.34
School 533	10	2.18	0.09	0.29		8	1.86	0.22	0.61		0.32
School 517	9	2.47	0.21	0.63		11	2.19	0.10	0.34		0.29
School 511	11	2.33	0.14	0.45		9	2.05	0.17	0.51		0.28
School 504	9	1.97	0.17	0.52		9	1.80	0.12	0.35		0.16
School 525	9	2.18	0.13	0.38		9	2.02	0.12	0.36		0.16
School 515	11	2.31	0.17	0.57		7	2.18	0.16	0.41		0.13
School 514	11	1.81	0.17	0.57		8	1.82	0.20	0.58		0.00
School 531	7	2.19	0.21	0.54		11	2.26	0.18	0.61		-0.07
School 508	7	2.43	0.30	0.80		6	2.60	0.14	0.34		-0.18
School 505	8	2.33	0.22	0.62		10	2.60	0.18	0.55		-0.27
School 512	3	2.17	0.22	0.38		3	2.58	0.16	0.27		-0.41
Total	248	2.46	0.04	0.63		227	2.07	0.03	0.52		0.39

Note. Sorted by descending mean differences. *p < .05.

Table 8.9. Gender Differences Within Schools, Grade 8

	Girls					Boys					Diff.
	N	M	SE	SD		N	M	SE	SD		M
School 804	10	3.17	0.27	0.86		5	1.72	0.54	1.21		1.46*
School 833	7	4.27	0.10	0.27		6	2.97	0.47	1.16		1.30*
School 806	7	3.14	0.25	0.66		8	1.86	0.20	0.55		1.27*
School 803	5	3.80	0.19	0.42		9	2.57	0.18	0.55		1.23*
School 805	9	3.16	0.14	0.42		7	2.26	0.22	0.59		0.90*
School 820	9	3.41	0.19	0.58		5	2.57	0.40	0.90		0.83
School 822	8	3.41	0.28	0.80		7	2.60	0.21	0.56		0.81*
School 809	6	3.20	0.28	0.67		9	2.39	0.28	0.85		0.81
School 832	7	3.03	0.21	0.55		8	2.39	0.23	0.64		0.63
School 807	5	3.11	0.44	0.98		10	2.50	0.26	0.83		0.61
School 813	10	3.09	0.25	0.79		7	2.49	0.28	0.73		0.60
School 817	8	3.21	0.21	0.60		7	2.65	0.28	0.74		0.55
School 810	5	3.03	0.17	0.37		10	2.49	0.27	0.87		0.54
School 818	12	2.98	0.20	0.69		3	2.49	0.46	0.80		0.49
School 831	9	2.67	0.28	0.83		6	2.24	0.19	0.45		0.43
School 819	6	2.81	0.36	0.89		9	2.39	0.23	0.70		0.42
School 814	7	3.35	0.27	0.72		8	2.95	0.26	0.72		0.41
School 811	6	3.56	0.27	0.66		9	3.20	0.21	0.64		0.36
School 830	5	3.48	0.43	0.96		10	3.25	0.13	0.42		0.23
School 828	11	3.29	0.18	0.60		5	3.07	0.34	0.77		0.23
School 824	9	4.03	0.15	0.46		6	3.82	0.26	0.63		0.21
School 812	8	2.72	0.20	0.56		7	2.51	0.13	0.35		021
School 801	6	2.83	0.25	060		8	2.66	0.17	0.47		0.17
School 815	10	3.36	0.23	0.74		5	3.26	0.34	075		0.10
School 816	8	3.06	0.26	0.75		7	3.00	0.23	0.61		0.06
School 829	8	2.86	0.28	0.79		5	2.82	0.30	0.66		0.05
School 821	10	3.56	0.28	0.88		5	3.53	0.27	060		0.03
School 808	7	2.63	0.31	0.81		7	2.60	0.30	0.78		0.03
School 802	6	3.28	0.14	0.33		9	3.29	0.26	0.77		-0.01
School 823	10	3.54	0.16	0.51		5	3.58	0.20	0.45		-0.05
School 825	5	2.66	0.12	0.26		9	2.84	0.20	0.60		-0.18
School 826	6	2.79	0.25	0.61		9	3.14	0.10	0.30		-0.35
Total	245	3.21	0.05	0.73		230	2.4	0.05	0.78		0.47

Note. Sorted by descending mean differences. *p < .05.

Table 8.10. Regression Analysis, Grade 5 (Model 2)

	B	SE (B)	β	t	p
School_502	-0.20	0.27	-0.07	-0.74	0.460
School_503	-0.70	0.28	-0.21	-2.53	0.012
School_504	-0.97	0.28	-0.30	-3.52	0.000
School_505	-0.36	0.28	-0.11	-1.29	0.198
School_507	-0.67	0.30	-0.16	-2.24	0.026
School_508	-0.36	0.29	-0.10	-1.25	0.211
School_509	-0.36	0.28	-0.10	-1.25	0.211
School_511	-0.67	0.27	-0.22	-2.45	0.015
School_512	-0.48	0.33	-0.09	-1.45	0.149
School_513	-0.85	0.29	-0.23	-2.95	0.003
School_514	-1.07	0.27	-0.34	-3.91	0.000
School_515	-0.64	0.28	-0.20	-2.32	0.021
School_516	-0.36	0.28	-0.11	-1.31	0.191
School_517	-0.52	0.27	-0.17	-1.90	0.058
School_518	-0.71	0.27	-0.23	-2.58	0.010
School_519	-0.05	0.34	-0.01	-0.15	0.879
School_520	-0.52	0.27	-0.18	-1.92	0.055
School_521	-0.07	0.28	-0.02	-0.25	0.801
School_522	-0.40	0.28	-0.12	-1.44	0.151
School_523	-0.91	0.27	-0.29	-3.33	0.001
School_524	-0.27	0.28	-0.09	-0.99	0.323
School_525	-0.75	0.28	-0.23	-2.73	0.007
School_527	-0.54	0.28	-0.15	-1.91	0.057
School_529	-0.77	0.28	-0.24	-2.81	0.005
School_530	-0.68	0.29	-0.18	-2.39	0.017
School_531	-0.58	0.28	-0.18	-2.11	0.036
School_532	-0.73	0.28	-0.23	-2.65	0.008
School_533	-0.83	0.28	-0.26	-3.03	0.003
School_534	-0.65	0.28	-0.19	-2.31	0.021
Gender	0.38	0.05	0.31	7.42	0.000

Note. B = unstandardized coefficients, SE (B) = standard error of unstandardized coefficients, β = standardized coefficients. For gender, 1 = girl, 0 = boy. For this model, R2 = .264 and ΔR2 = .091.

Table 8.11 Regression Analysis, Grade 8 (Model 2)

	B	SE (B)	β	t	p
School_802	0.57	0.26	0.13	2.20	0.028
School_803	0.31	0.26	0.07	1.18	0.237
School_804	-0.15	0.26	-0.03	-0.58	0.563
School_805	-0.02	0.26	-0.01	-0.09	0.931
School_806	-0.29	0.26	-0.06	-1.12	0.263
School_807	0.02	0.26	0.00	0.07	0.944
School_808	-0.15	0.26	-0.03	-0.57	0.572
School_809	0.00	0.26	0.00	0.00	0.997
School_810	-0.02	0.26	0.00	-0.06	0.949
School_811	0.62	0.26	0.14	2.41	0.016
School_812	-0.15	0.26	-0.03	-0.59	0.559
School_813	0.04	0.25	0.01	0.16	0.874
School_814	0.39	0.26	0.09	1.50	0.134
School_815	0.49	0.26	0.11	1.89	0.059
School_816	0.26	0.26	0.06	1.00	0.320
School_817	0.17	0.26	0.04	0.66	0.511
School_818	-0.01	0.26	0.00	-0.05	0.961
School_819	-0.16	0.26	-0.04	-0.62	0.539
School_820	0.28	0.26	0.06	1.07	0.284
School_821	0.71	0.26	0.16	2.75	0.006
School_822	0.25	0.26	0.06	0.98	0.326
School_823	0.72	0.26	0.16	2.76	0.006
School_824	1.13	0.26	0.25	4.38	0.000
School_825	0.08	0.26	0.02	0.30	0.767
School_826	0.29	0.26	0.06	1.10	0.270
School_828	0.38	0.26	0.09	1.47	0.142
School_829	0.03	0.27	0.01	0.12	0.905
School_830	0.64	0.26	0.14	2.47	0.014
School_831	-0.31	0.26	-0.07	-1.21	0.229
School_832	-0.06	0.26	-0.01	-0.22	0.825
School_833	0.89	0.27	0.18	3.31	0.001
Gender	0.45	0.07	0.28	6.84	0.000

Note. B = unstandardized coefficients, SE (B) = standard error of unstandardized coefficients, β = standardized coefficients. For gender, 1 = girl, 0 = boy. For this model, R2 = .282 and ΔR2 = .076.

Table 8.12. Gender Performance on Specific Tasks

			N	M	SE	SD
Grade 5	Animal Police (argue)	Girls	59	2.43	0.09	0.67
		Boys	57	2.03	0.07	0.53
	Architect (describe)	Girls	78	2.21	0.07	0.59
		Boys	59	2.04	0.07	0.57
	Helmet (argue)	Girls	73	2.57	0.09	0.74
		Boys	69	2.16	0.08	0.64
	Home Place (describe)	Girls	72	2.82	0.09	0.76
		Boys	63	2.16	0.07	0.59
	Remark (argue)	Girls	72	2.48	0.08	0.69
		Boys	80	2.07	0.06	0.57
	Substitute Teacher (describe)	Girls	88	2.63	0.08	0.74
		Boys	74	2.17	0.09	0.73
	Super Power (narrate)	Girls	51	1.98	0.09	0.65
		Boys	52	1.92	0.08	0.58
Grade 8	Animal Police (argue)	Girls	63	3.24	0.11	0.86
		Boys	57	2.70	0.11	0.83
	Architect (describe)	Girls	66	3.03	0.08	0.67
		Boys	69	2.68	0.09	0.73
	Helmet (argue)	Girls	85	3.39	0.09	0.79
		Boys	83	2.78	0.10	0.95
	Home Place (describe)	Girls	66	3.15	0.10	0.79
		Boys	77	2.64	0.09	0.78
	Remark (argue)	Girls	72	3.29	0.09	0.79
		Boys	46	2.99	0.16	1.09
	Substitute Teacher (describe)	Girls	54	3.11	0.13	0.95
		Boys	58	2.73	0.12	0.93
	Super Power (narrate)	Girls	84	3.15	0.09	0.84
		Boys	64	2.84	0.10	0.83

Note. The mean values are based on fair average values after anchoring rating scales, step difficulty, and raters.

Table 8.13 . Gender Differences Between Tasks with Girls > Boys

		Levene's Test		Difference					Effect Size	
		F	P	M diff.	SE diff.	df	t	p	d	SE
Grade 5	Animal Police (argue)	2.462	0.119	0.40	0.11	114	3.59	0.000	0.66	0.19
	Architect (describe)	0.056	0.813	0.18	0.10	135	1.75	0.082	0.30	0.17
	Helmet (argue)	0.671	0.414	0.41	0.12	140	3.51	0.001	0.59	0.17
	Home Place (describe)	4.012	0.047	0.66	0.12	131.21	5.67	0.000	0.96	0.18
	Remark (argue)	3.103	0.08	0.42	0.10	150	4.08	0.000	0.66	0.17
	Substitute Teacher (describe)	0.014	0.905	0.46	0.12	160	3.98	0.000	0.62	0.16
	Super Power (narrate)	1.255	0.265	0.06	0.12	101	0.48	0.635	0.09	0.20
Grade 8	Animal Police (argue)	0.127	0.722	0.55	0.15	118	3.53	0.001	0.64	0.19
	Architect (describe)	1.011	0.317	0.35	0.12	133	2.92	0.004	0.50	0.17
	Helmet (argue)	3.02	0.084	0.62	0.13	166	4.58	0.000	0.70	0.16
	Home Place (describe)	0.03	0.863	0.51	0.13	141	3.90	0.000	0.65	0.17
	Remark (argue)	4.357	0.039	0.30	0.19	75.032	1.63	0.108	0.33	0.19
	Substitute Teacher (describe)	1.605	0.208	0.38	0.18	110	2.16	0.033	0.41	0.19
	Super Power (narrate)	0.049	0.825	0.31	0.14	146	2.27	0.025	0.37	0.17

Note. The mean values are based on fair average values after anchoring rating scales, step difficulty, and raters.

and 20.6% of variance explained for grade 5 and grade 8, respectively (this model thus equals the one-way between-subject ANOVA). Adding the gender dummy (see Table 8.10 and Table 8.11), significantly more variance was explained. For grade 5, 26.4% of the variance was explained by Model 2 ($F(1.444) = 55.00$, $p < .001$, adjusted $R2 = .214$, $\Delta R2 = .091$). For grade 8, 28.2% of the variance was explained by Model 2 ($F(1.442) = 46.75$, $p < .001$, adjusted $R2 = .23$, $\Delta R2 = .076$). On average, even when accounting for what seems to be very different writing instruction quality, girls outperformed boys. Taken together, school and gender accounted for more than 25% of the variance irrespective of grade. Differences between schools/classes can be assumed to be caused by known factors such as writing pedagogy, socio-economic background, etc. The systematic differences between girls and boys are hard to explain based on these data.

Continuing to explore gender differences, Tables 8.12 and 8.13 present investigations into differences associated with the tasks. As can be seen, and as can be expected, girls outperformed boys on most of the task across both grades. There were tasks, however, that did not follow this pattern. For grade 5 students, "Super Power" showed non-significant differences, and "Architect" was only significant at the 10% level. For grade 8 students, "Remark" showed non-significant differences. These differences might be related to the distribution of NSBWT tasks. Consulting Appendix C (Tables C.1 and C.2), however, indicates that classes with significant gender differences were administrated tasks with non-significant differenc-

Table 8.14 . Rating Scale Performance

	Grade 5					Grade 8			
	N	M	SE	SD		N	M	SE	SD
WRI	475	2.38	0.03	0.74		475	3.01	0.04	0.90
Content	475	2.38	0.03	0.70		475	3.04	0.04	0.86
Structure	475	2.24	0.03	0.63		475	2.93	0.04	0.83
Language Use	475	2.21	0.03	0.62		475	2.98	0.04	0.85
Coding	475	2.17	0.03	0.74		475	2.97	0.04	0.88

Note. The mean values are based on fair average values after anchoring tasks, step difficulty, and raters.

es overall, and classes with small gender differences were administrated tasks that overall demonstrated large and significant gender differences.

Lastly, this report will present additional descriptive statistics for the rating scales. Table 8.14 presents mean values for the rating scales for each grade level. The rating scale coding was most difficult in both grade 5 and grade 8. The easiest rating scales, across the grades, were content and WRI.

9. Conclusion

This technical report sought to answer three questions: 1) What were the results of the teacher survey? 2) What was the measurement quality of the NSBWT? 3) What were the results of the NSBWT in general and for groups of students?

The answers to the first two questions determine the trustworthiness of the answers to the third. As can be seen, generally, the teachers were positive about the NSBWT task and task administration facilities (such as instructions). Ninety to ninety-five percent of the respondents agreed to some extent with statements about, for example, task relevance, student motivation, and sufficient time for task completion. The teachers' perceptions, as captured in this survey, do not in themselves validate or invalidate the interpretation of student scores. Nevertheless, if most of the teachers would have totally disagreed with the statements, one could suspect that not all of the teachers would have succeeded in making the task administration motivating enough for students.

The measurement quality report demonstrated a good model–data fit and good reliability within the highly controlled MFRM context. Although the fit was good, it would have increased if the zeros had been removed, as was shown in Figure 7.1. However, such an action would have had less positive consequences as well; assuming that students performed differently on the two tasks, treating the zeros as missing could have inflated or deflated the students' scores.

The results showed that the average grade 5 student performed on an

NPP2 level, and that the average grade 8 student performed on an NPP3 level. However, the sub-group analysis revealed major differences within groups and between classes/schools. The analysis showed that girls, on average, outperformed boys to quite a large extent. The gender effect was present even when controlling for class/school. Nevertheless, the gender difference was not constant across tasks or schools; there were tasks and classes/schools where boys and girls performed equally. Regarding the tasks in grade 5, equal scores between girls and boys seem to be associated with girls underperforming rather than boys over performing. The opposite pattern is implied in the scores in grade 8.

In all, this report concludes that these measures of grade 5 and grade 8 student writing proficiency produced trustworthy results. To the extent that the results are generalizable beyond this test context, they offer an important reminder that quite extensive work remains to be done. A general goal is to have as many students as possible at high NSBWT proficiency profile levels; for example, the average grade 5 student was able to produce text with content that was only somewhat relevant for the task. Another, and more acute goal, would be to seriously decrease the class/school and gender impact.

Notes

[1]The concept of a "rating scale" can be ambiguous. Rating scales—also known as proficiency scales or scoring rubrics—refer to descriptions of (language) proficiency that are ordered in levels or bands (Davies et al., 1999, p. 153; Knoch, 2009, p. 39). A rating scale will typically describe a number of proficiency levels for one or more skills. In analytical rating, one uses two or more scales for the assessment of writing proficiency. Some authors use "criterion" to denote what is understood as a rating scale in this context.

[2]The "true rating model" states that it is enough for five raters to accomplish "true ratings." Based on then-current rates, this is approximately USD 112,000 (when buying 1 USD for 8.5 NOK).

[3]This MFRM model builds on the rating scale model (RSM), which assumes that category difficulty is common to all items. In essence, the model can answer the question, "How does this set of raters use this set of . . . scales?" (Myford & Wolfe, 2003, p. 28). As a consequence, the items (scales) are treated as parts of unidimensional total scores. The concept of unidimensionality refers to statistical claims about unidimensional patterns (e.g., a one-factor solution) rather than psychological claims about more or less distinct constructs (cf. McNamara, 1996).

Literature

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1301–1322). Chichester, West Sussex: Wiley-Blackwell.

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. http://doi.org/10.1080/09585176.2015.1129980

Berge, K. L., Skar, G. B., Matre, S., Solheim, R., Evensen, L. S., Otnes, H., & Thygesen, R. (2017). Introducing teachers to new semiotic tools for writing instruction and writing assessment: consequences for students' writing proficiency. *Assessment in Education: Principles, Policy & Practice*, 1–20. http://doi.org/10.1080/0969594X.2017.1330251

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. http://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. http://doi.org/10.1007/s11092-008-9068-5

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model* (3rd ed.). New York: Routledge.

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83–100. http://doi.org/10.1177/0265532214542994

Brookhart, S. M. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice, 30*(1), 3–12. http://doi.org/10.1111/j.1745-3992.2010.00195.x

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

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

Engelhard, G. (2013). *Invariant Measurement.* New York: Routledge.

Engelhard, G., & Wind, S. A. (2013). *Rating Quality Studies Using Rasch Measurement Theory.* New York: The College Board. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2013/8/researchreport-2013-3-rating-quality-studies-using-rasch-measurement-theory.pdf

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. http://doi.org/10.1080/09585176.2015.1134338

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112. http://doi.org/10.3102/003465430298487

Holten-Kvistad, A., & Skar, G. B. (2016). Utvikling av skriveoppgaver. Paper presented at Svenska med didaktisk inriktning, Karlstad, Sweden, November 24-25, 2016.

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.* Frankfurt am Main: Peter Lang.

Lie, S., Hopfenbeck, T., Ibsen, E., & Turmo, A. (2005). *Nasjonale prøver på ny prøve* [Putting national tests to the test]. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Linacre, J. M. (2013). *A user's guide to FACETS. Rasch-model computer programs*. Program manual 3.71.0. Hämtad 2015-04-07. Retrieved from http://www.winsteps.com/a/Facets-ManualPDF.zip

Linacre, J. M. (2017a). Facets® (version 3.80.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.
McNamara, T. F. (1996). *Measuring second language performance.* New York: Longman.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. http://doi.org/10.1177/026553229601300302
Myford, C. M., & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.
Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. J*ournal of Applied Measurement, 5*(2), 189–227.
Myford, C. M., & Wolfe, E. W. (2009). Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. *Journal of Educational Measurement, 46*(4), 371–389. Retrieved from http://www.jstor.org/stable/25651523
Norwegian Directorate for Education and Training. (2007). The Knowledge Promotion. Retrieved from http://www.udir.no/Stottemeny/English/Curriculum-in-English/_english/Knowledge-promotion---Kunnskapsloftet/
NOU 2002:10 (2002). *Førsteklasses fra første klasse – Forslag til rammeverk for et nasjonalt kvalitetsvurderingssystem av norsk grunnopplæring*. Downloaded from: https://www.regjeringen.no/no/dokumenter/nou-2002-10/id145378/sec5?q=grunnleggende#KAP3-4-1-P3
Organisation for Economic Co-operation and Development. (2005). *The Definition and Selection of Key Competencies - Executive Summary*. Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/pisa/35070367.pdf
Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.
Schumacker, R. E., & Smith, E. V. (2007). Reliability. A Rasch Perspective. *Educational and Psychological Measurement, 67*(3), 394–409. http://doi.org/10.1177/0013164406294776
Skar, G. B., & Berge, K. L. (2017). *Elevers skrivförmåga och texters kvantitativa egenskaper* [Students' writing proficiency and quantitative text features]. Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking. Retrieved from http://www.skrivesenteret.no/uploads/materiell/LIX-rapport.pdf
Skar, G. B., Evensen, L. S., & Iversen, J. M. (2015). *Læringsstøttende prøver i skriving 2014. Teknisk rapport.* [The National Sample-Based Writing Test and Formative Assessment Package 2014. Technical Report]. Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.
Skar, G. B., & Iversen, J. M. (2015). *Læringsstøttende prøver i skriving 2014. Teknisk rapport avseende pilotuppgifter HT 2014*. Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.
Skar, G. B., & Iversen, J. M. (2016). *Teknisk rapport: Pilotering av uppgifter till den nasjonale utvalgsprøven i skriving 2016*. Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.
Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Aasen, A. J., & Skar, G. B. (2017). Skriving og dybdelæring [Writing and Deep Learning]. Paper presented at NOLES, Copenhagen, March 27-29, 2017.

Appendix A

Writer–reader interaction

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
• The relationship the text establishes between the writer and reader is unclear. Access to the assignment text is needed in order to understand the pupil's text.	• In the non-fiction text, the pupil attempts to establish a relevant relationship between the writer and reader (see the assignment text), but why the writer is addressing this particular reader may be somewhat unclear.	• The non-fiction text indicates a relevant relationship between the writer and reader (see the assignment text), but does not consistently relate to this reader's perspective.	• The non-fiction text establishes a largely relevant relationship between the writer and reader (see the assignment text), but does not always relate to this reader's perspective.	• The non-fiction text establishes a relevant relationship between the writer and reader (see the assignment text), and consistently relates to this reader's perspective.
	• The non-fiction text can sometimes attempt to address the reader's need to know the participants, concepts and circumstances (e.g. by providing explanations, which are not, however, adapted to suit the reader given in the assignment text).	• The non-fiction text to some extent addresses the reader's need to know the participants, concepts and circumstances (e.g. by providing explanations which in some cases are adapted to suit the reader given in the assignment text).	• The non-fiction text largely addresses the reader's need to know the participants, concepts and circumstances (e.g. by providing explanations, which are often adapted to suit the reader in the assignment text).	• The non-fiction text consistently addresses the reader's need to know the participants, concepts and circumstances (e.g. by providing explanations which are adapted to suit the reader in the assignment text).
		•The non-fiction text can include a very few signposts, but these are not always necessary or functional.	•The non-fiction text can include signposts, which are mainly functional and help to keep the reader focused.	•The non-fiction text can include signposts, which are used in a balanced and systematic manner.
	•The fiction text stages a fictional world that lacks credibility and coherence.	•The fiction text stages a fictional world which has participants, circumstances and atmosphere, but which is not consistently credible and may include some faulty logic.	•The fiction text stages a mainly credible fictional world with participants, circumstances and atmosphere.	•The fiction text stages a consistently credible fictional world with participants, circumstances and atmosphere.

Content

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
•The content of the text is partly relevant or has some parts that are relevant to the assignment.	•The content of the text is mainly relevant to the assignment.	•The content of the text is relevant to the assignment.	⇢	⇢
•Parts of the text focus on the topic.	•The topic is the main focus of the whole text.	⇢	•The topic is the focus of the whole text. The content is expediently balanced.	•The topic is in focus and can be developed in the text. The content is expediently balanced.
•The non-fiction text contains a small number of simple examples, explanations and reasoning, but these may not be particularly relevant.	•The non-fiction text contains simple examples, explanations and reasoning, most of which are relevant.	•The non-fiction text contains some relevant examples, explanations and/or reasoning.	•The non-fiction text contains several good and relevant examples, explanations and/or reasoning.	•The non-fiction text contains consistently good and relevant examples, explanations and/or reasoning.
•The fiction text is action-driven or comprises disconnected incidents.	•The fiction text is primarily action-driven and can show some tendencies towards developing motives, topics, characters, environments or events.	⇢	•The fiction text may be partly action-driven, but also contains parts in which motives, topics, characters, environments or events are well developed.	•The fiction text shows the development of motives, topics, characters, environments or events, and these aspects of the text are coherent.

⇢ = the same as the previous level

Text Sturcture

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
•At an overall level, the structure of the non-fiction text is unclear. •The structure of the fiction text is unclear.	•The non-fiction text shows attempts at structure (e.g. in the form of an introduction and/or conclusion, bullet points or headings). •The fiction text can show attempts at structure.	•The non-fiction text shows attempts at using structuring principles that are suited to the writing situation (e.g. developing towards a main point). •The fiction text can show attempts at using structuring principles in order to develop the plot.	•The non-fiction text shows a use of structuring principles that gives a mostly expedient structure suited to the writing situation. The text may have a conclusion (e.g. a summary or final appeal). •The fiction text applies structuring principles to develop the plot.	•The non-fiction text shows a use of structuring principles that gives an expedient structure suited to the writing situation. The introduction forms a relevant background for the text. The text has an expedient conclusion (e.g. a summary or final appeal). •The fiction text applies systematic and advanced structuring principles to develop the plot.
•The paragraphs are short and often consist of disconnected statements.	•The paragraphs consist of elements that are structured in an associative manner or that lack a logical order. The paragraphs can consist of disconnected statements.	•The paragraphs consist of content elements that are often grouped by topic.	•As a rule, the paragraphs have a logical internal structure (e.g. topic sentences) and are often grouped by topic. This makes the text easy to follow (e.g. explanations are often presented in a functional order). Paragraphs may be marked in a graphically correct manner.	•The paragraphs have a logical internal structure (e.g. use topic sentences). The paragraphs are ordered by topic in a way that makes the text easy to follow (e.g. explanations are often presented in a functional order, such as the general before the specific or the conclusion referring back to the introduction). Paragraphs are usually marked in a graphically correct manner.
•The text can show use of several types of simple connecting words, but they are not always used in a functional manner.	•The text shows use of several types of simple connecting words. There is little variety in their use.	•The text shows use of several types of connecting words. Simple connecting words are often overused. The connecting words are superfluous in some cases.	•The text shows use of several types of connecting words, which are used in a functional manner most of the time.	•The text shows functional use of varied and advanced connecting words. These are not used in a superfluous manner.

Language Use

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
•Most of the sentences begin in the same way. The text has a simple syntax without subordination.	•The sentences show some variety in the grounding field, but the text shows little variety in syntax, e.g. many repetitive parts of sentences.	•The sentences show some variety in the grounding field. The text may demonstrate the beginnings of complex syntax. Some of the more complex sentences may contain syntactic flaws.	•The sentences show some variety in the grounding field. Parts of the text may show complex and varied syntax.	•The sentences show variety in the grounding field. The text has complex and varied syntax.
•The text is characterised by colloquial language, with associative content elements or colloquial wording.	•The text may be characterised by colloquial language, with associative elements, asides and many fillers. •There is little variety in the choice of words. Concepts and wording are often imprecise.	•The text can show some variety in the choice of words. Wording and concepts may be precise.	•The text mostly shows variety in the choice of words. Wording and concepts are mainly precise.	•The text shows variety in the choice of words and precise use of wording and concepts. •Advanced concepts are used in a correct manner in texts where this is relevant.
		•The text may show use of linguistic devises (e.g. metaphors, similes, contrasts, rhetorical questions, repetition and irony).	•The text may show functional use of linguistic devises.	⇢

⇢ = the same as the previous level

Coding competencies

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
•The text shows extensive use of phonological strategy. Some non-phonetic words are spelt correctly.	•The text shows correct spelling of some long phonetic words and several non-phonetic words. The text may contain some dialect words.	•Most words in the text are largely correctly spelt. The text often contains og/å mistakes (confusing the infinitive marker 'å' with the word for 'and' -'og') as a result of being more linguistically advanced.	•Most words in the text are correctly spelt. Some og/å mistakes may occur.	•Most words in the text are correctly spelt.
•The text contains some full stops at the end of complete sentences. The text mostly has capital letters in proper names and at the beginning of new sentences.	•Punctuation other than full stops are attempted: question marks, exclamation marks. The text often uses commas in lists. The text mostly has capital letters in proper names and at the beginning of new sentences.	•The text uses the major punctuation marks correctly most of the time. The text uses commas in lists. The text may contain commas between complete sentences that are connected with connecting words. The text has capital letters in proper names and at the beginning of new sentences.	•The text uses the major punctuation marks correctly most of the time. The text has commas in lists and between most complete sentences that are connected with connecting words. The text has capital letters in proper names and at the beginning of new sentences.	•The text uses the major punctuation marks correctly. Commas are mostly used correctly in the text, but mistakes in more complex syntaxes may occur (e.g. after a relative clause). The text has capital letters in proper names and at the beginning of new sentences.
		•The text may mark direct speech with a dash or colon and quotation marks (in texts where this is relevant).	⇢	•The text shows correct use of punctuation for direct speech.
•Tenses may not be used consistently.	•The text mostly shows functional use of tenses in simple sentences.	•The text mostly shows functional use of tenses. There may be some mistakes in advanced sentences.	⇢	•The text shows functional use of tenses.
	•The text shows correct concord most of the time.	•The text shows correct concord.	⇢	⇢

⇢ = the same as the previous level

Appendix B

NSBWT Proficiency Profiles (NPP)

NPP1

At NPP 1 you need access to the assignment text in order to understand the pupil's text. The content of the text is partly relevant to the assignment set. The structure of the text is unclear. Paragraphs are short and often consist of disconnected statements. Most sentences start in the same way, and the syntax of the text is simple. The text is characterised by colloquial language. The text shows extensive use of phonological strategy, but some non-phonetic words are spelt correctly. The text contains some full stops at the end of complete sentences, and mostly has capital letters in proper names and at the beginning of new sentences.

NPP2

At NPP 2 the pupil attempts to adapt the text to the recipient mentioned in the assignment text. The content is mostly relevant to the assignment set. The text may show attempts at structuring, for example by using bullet points, headings, an introduction and/or conclusion. The content, however, is often structured in an associative manner or lacks a logical order. There is often little variety in syntax in the text, and it may be characterised by colloquial language. Elements of dialect may occur. More non-phonetic words are spelt correctly. The pupil tries to use punctuation marks other than full stops, for example question marks and exclamation marks. Commas are often used in lists, and the text mostly has capital letters in proper names and at the beginning of new sentences.

NPP3

At NPP 3 the text is partly adapted to the recipient mentioned in the assignment text. The content is relevant to the assignment set. The text shows attempts at using structuring principles. For example, a non-fictional text can be structured in a way leading up to a main point. Paragraphs are often grouped by topic. The text may show the beginnings of complex syntax. Some of the more complex sentences may contain syntactic flaws. Wording and concepts may be precise. Most of the words in the text are spelt correctly, but the text contains og/å mistakes (confusing the infinitive marker 'å' with the word for 'and' -'og'). The major punctuation marks (full stop, question mark, exclamation mark) are correctly used most of the time. In addition to commas in lists, the text may also use commas between complete sentences.

NPP4

At NPP 4 the text is largely adapted to the recipient mentioned in the assignment text. The content is relevant to the assignment set. The text is mostly structured in an expedient manner. Most of the paragraphs have a logical internal structure. Paragraphs may be marked in a graphically correct manner. Parts of the text may display complex and varied syntax, and wording and concepts are mostly used in a precise manner. Most words in the text are correctly spelt. Some og/å mistakes may occur. The major punctuation marks are mostly used correctly, and the text has commas between most complete sentences joined by connecting words.

NPP5

At NPP 5 the text is adapted to the recipient mentioned in the assignment text. The content is relevant to the assignment set. The text is structured in an expedient manner. The paragraphs have a logical internal structure. Paragraphs are usually marked in a graphically correct manner. The text has a complex and varied syntax, and wording and concepts are used in a precise manner. Advanced concepts are used in a correct manner in texts where this is relevant. Most words in the text are correctly spelt. Commas are mostly used correctly in the text, but mistakes in more complex syntaxes may occur.

Appendix C: NSBWT-16 Task Distribution Design

Table C.1 Distribution 5th grade

	16131 Home Place (describe)	16161 Helmet (argue)	16132 Substitute Teacher (describe)	16133 Architect (describe)	16151 Super Power (narrate)	16162 Remark (argue)	16163 Animal Police (argue)
501	5	5					
502	19		19				
503	16			16			
504		17	18				
505		18		18			
507			10	10			
508		12	13				
509			14			14	
511				20		20	
512				6			6
513					13	13	
514					19		19
515	18				17		
516						17	17
517	20					20	
518		19				19	
519	5						5
520		21					21
521			15				15
522	18	18					
523		19	19				
524			18	18			
525				18	18		
527						15	15
529	18		18				
530		13		13			
531			18		18		
532				18		18	
533					18		18
534	16					16	
Schools	9	9	10	9	6	9	8
Students	135	142	162	137	103	152	116

Table C.2 Distribution 8th grade

	16131 Home Place (describe)	16161 Helmet (argue)	16132 Substitute Teacher (describe)	16133 Architect (describe)	16151 Super Power (narrate)	16162 Remark (argue)	16163 Animal Police (argue)
801	14				14		
802	15					15	
803	14						14
804		15				15	
805		16					16
806	15	15					
807			15				15
808	14		14				
809		15	15				
810	15			15			
811		15		15			
812			10	15			
813		17			17		
814			15		15		
815				15	15		
816			15			15	
817				15		15	
818					15	15	
819				15			15
820					14		14
821						15	15
822	15				15		
823		15				15	
824		15			15		
825	14			14			
826		15			15		
828				16			16
829	13					13	
830		15					15
831	14		15				
832		15		15			
833			13		13		
Schools	10	11	8	9	10	8	8
Students	143	168	112	135	148	118	120

Appendix D: Rater Statistics

Table D.1 Rater statistics

Rater ID	Obs	Fair	Logit	S.E.	Infit	Infit_Z	Outfit	Outfit_Z	SR-ROR
R29	2,16	2,05	0,88	0,08	1,41	4,56	1,45	4,94	0,26
R57	2,34	2,06	0,86	0,08	0,89	-1,40	0,87	-1,74	0,30
R61	2,16	2,08	0,82	0,08	0,80	-2,65	0,81	-2,45	0,32
R60	2,35	2,08	0,81	0,13	1,05	0,42	0,99	-0,05	0,29
R16	2,22	2,14	0,70	0,08	0,83	-2,14	0,82	-2,27	0,34
R2	2,22	2,15	0,69	0,08	1,12	1,45	1,09	1,06	0,32
R58	2,26	2,16	0,67	0,08	1,14	1,65	1,19	2,21	0,30
R53	2,29	2,17	0,65	0,08	0,89	-1,36	0,94	-0,68	0,39
R64	2,31	2,21	0,57	0,08	0,89	-1,38	1,00	-0,03	0,36
R59	2,48	2,22	0,56	0,08	0,81	-2,43	0,79	-2,73	0,35
R55	2,23	2,23	0,53	0,08	0,81	-2,52	0,85	-1,93	0,35
R30	2,42	2,27	0,46	0,08	0,99	-0,14	1,04	0,48	0,39
R7	2,43	2,30	0,41	0,08	1,22	2,59	1,14	1,71	0,42
R52	2,46	2,31	0,39	0,08	1,09	1,09	1,05	0,67	0,32
R47	2,46	2,34	0,34	0,08	1,35	3,86	1,37	4,06	0,32
R31	2,33	2,37	0,29	0,08	0,94	-0,72	1,01	0,14	0,35
R1	2,29	2,41	0,22	0,10	1,18	1,77	1,19	1,90	0,37
R63	2,67	2,41	0,22	0,07	0,83	-2,22	0,85	-2,02	0,34
R45	2,60	2,43	0,19	0,08	0,99	-0,14	0,97	-0,38	0,37
R69	2,65	2,43	0,19	0,07	0,86	-1,82	0,85	-2,03	0,33
R4	2,59	2,43	0,19	0,06	0,71	-5,47	0,71	-5,55	0,37
R46	2,57	2,44	0,17	0,08	1,16	1,92	1,11	1,30	0,34
R22	2,56	2,44	0,16	0,08	0,92	-0,96	0,94	-0,70	0,34
R32	2,40	2,45	0,15	0,08	1,03	0,37	1,02	0,32	0,32
R24	2,67	2,45	0,15	0,08	0,85	-1,97	0,82	-2,31	0,38
R43	2,38	2,46	0,14	0,08	0,72	-3,78	0,72	-3,81	0,39
R6	2,55	2,47	0,13	0,08	0,84	-2,14	0,81	-2,48	0,35
R20	2,41	2,47	0,12	0,08	1,22	2,59	1,23	2,63	0,28
R25	2,75	2,47	0,11	0,07	0,97	-0,39	0,96	-0,55	0,37
R12	2,67	2,48	0,10	0,08	1,20	2,38	1,13	1,58	0,41
R23	2,67	2,50	0,07	0,08	1,25	2,89	1,17	2,01	0,34
R18	2,61	2,52	0,04	0,08	0,89	-1,35	0,94	-0,76	0,35
R19	2,66	2,54	0,01	0,08	0,72	-3,86	0,73	-3,69	0,37
R65	2,72	2,55	-0,02	0,08	1,04	0,51	1,07	0,85	0,36
R42	2,75	2,56	-0,03	0,08	1,19	2,10	1,12	1,40	0,37
R48	2,75	2,56	-0,03	0,08	1,07	0,85	1,08	1,03	0,35
R37	2,67	2,57	-0,04	0,07	0,57	-6,53	0,58	-6,38	0,38
R13	2,69	2,59	-0,07	0,07	0,86	-1,78	0,92	-0,96	0,33
R68	2,67	2,60	-0,09	0,08	1,16	1,94	1,13	1,55	0,35
R51	2,58	2,60	-0,09	0,08	0,97	-0,39	1,02	0,28	0,34
R5	2,66	2,61	-0,10	0,08	0,89	-1,43	0,88	-1,52	0,33
R38	2,79	2,61	-0,10	0,07	1,11	1,41	1,12	1,57	0,35
R26	2,82	2,61	-0,11	0,07	1,09	1,13	1,05	0,70	0,40
R41	2,59	2,62	-0,13	0,07	0,87	-1,63	0,87	-1,67	0,32

R39	2,73	2,63	-0,14	0,07	0,73	-3,74	0,73	-3,76	0,39
R56	2,73	2,63	-0,14	0,08	0,74	-3,48	0,74	-3,57	0,38
R36	2,79	2,63	-0,14	0,08	0,84	-2,03	0,91	-1,05	0,37
R21	2,82	2,65	-0,16	0,07	0,72	-4,08	0,72	-3,98	0,37
R44	2,73	2,65	-0,17	0,07	1,09	1,09	1,07	0,84	0,36
R10	2,53	2,66	-0,18	0,08	0,97	-0,33	0,94	-0,69	0,39
R67	2,82	2,66	-0,19	0,08	1,32	3,68	1,30	3,39	0,34
R33	2,67	2,67	-0,20	0,08	0,85	-1,88	0,84	-2,15	0,38
R49	2,62	2,68	-0,21	0,08	1,00	-0,02	1,01	0,14	0,35
R15	2,72	2,73	-0,30	0,07	0,93	-0,90	0,92	-0,96	0,37
R40	2,72	2,75	-0,32	0,07	0,78	-2,93	0,77	-3,17	0,40
R50	2,87	2,75	-0,33	0,08	0,86	-1,86	0,85	-1,99	0,37
R66	2,94	2,77	-0,36	0,07	1,49	5,58	1,46	5,33	0,29
R14	2,76	2,77	-0,36	0,07	1,04	0,55	1,02	0,33	0,38
R27	2,73	2,78	-0,37	0,07	1,29	3,39	1,24	2,89	0,38
R71	2,86	2,79	-0,39	0,07	0,87	-1,64	0,92	-0,98	0,35
R8	2,97	2,80	-0,41	0,07	0,82	-2,40	0,81	-2,64	0,39
R35	2,93	2,84	-0,46	0,07	1,15	1,83	1,21	2,49	0,32
R9	2,84	2,85	-0,47	0,08	1,17	2,07	1,17	2,05	0,39
R34	2,94	2,87	-0,51	0,08	1,59	6,31	1,63	6,56	0,32
R70	2,96	2,88	-0,52	0,08	1,46	5,09	1,44	4,61	0,39
R62	2,85	2,90	-0,55	0,07	1,10	1,26	1,08	1,08	0,37
R28	2,79	2,91	-0,56	0,07	0,76	-3,29	0,77	-3,19	0,38
R54	2,80	2,97	-0,65	0,07	0,89	-1,43	0,87	-1,64	0,36
R11	3,04	3,01	-0,70	0,08	1,05	0,67	1,13	1,59	0,39
R3	3,12	3,06	-0,78	0,07	0,86	-1,93	0,91	-1,12	0,30
R17	3,49	3,63	-1,59	0,08	1,31	3,63	1,33	3,64	0,35
Average	2,64	2,55	0,00	0,08	1,00	-0,17	1,00	-0,14	0,35