TEACHERS AS RATERS: AN INVESTIGATION OF A LONG-TERM WRITING ASSESSMENT PROGRAM

GUSTAF B. SKAR & LENNART J. JØLLE

Norwegian University of Science and Technology, Faculty of Teacher and Interpreter Education, Trondheim, Norway

Abstract

In 2010, the Norwegian Writing Centre (NWC) was commissioned by the Norwegian Directorate for Education and Training to develop the National Sample-Based Writing Test (NSBWT), which was to be administered annually to a national representative sample of students in primary and lower secondary school (NSBWT-5 for school year 5 and NSBWT-8 for school year 8). The NWC was also commissioned to set up a national panel of raters (NPR), consisting of teachers, with the purpose of 1) establishing a strong interpretive community and 2) having in place a panel that would reliably rate the NSBWT. The first reliability estimates from the autumn of 2010 indicated large variation. However, it was the belief of the NWC that an interpretive community would slowly evolve through rater training over a long period of time. The present study utilized multiple data sources to explore this assumption by investigating potential variation among a sub-sample of NPR members. The data consisted of one quantitative dataset of ratings and one qualitative dataset based on semi-structured interviews and live ratings. The quantitative investigation showed large variation among the raters, as did the investigation using qualitative data. The results are discussed in depth.

Keywords*:* writing assessment, reliability, interpretive community

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

Corresponding author: Gustaf B. Skar, Norwegian University of Science and Technology, NTNU, Faculty of Teacher and Interpreter Education, Pb 8900, 7491 Trondheim, Norway, +4791693654, email: gustaf.b.skar@ntnu.no

1. INTRODUCTION

In educational contexts in which writing is deemed important, writing is assessed using "direct measures" (cf. White, 1984), i.e., actual writing rather than selected response formats. It is assessed by either classroom teachers or external raters. Direct measures and human raters are considered important in stimulating positive washback (cf. Messick, 1996), and many governmental agencies around the world invest large sums on developing tasks, administering tests, and rating student texts using human raters. However, it has proven to be difficult to accurately measure student writing proficiency through direct measures because of the large variation among students, tasks, and raters (e.g., Bouwer, Béguin, Sanders, & van den Bergh, 2015; Coffman, 1971; Schoonen, 2012). In this article, we focus on variation among raters, i.e., "rater effects," or "the systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (Scullen, Mount, & Goff, 2000, p. 157). More specifically, we report the findings from a project in which a nationwide writing assessment program was discontinued, because of rater variation, and reintroduced on the assumption that adequate, generously funded, and extensive training of raters would resolve the problem regarding rater variation.

In 2005, the Norwegian government launched its first writing assessment program with the aim of evaluating the writing proficiency of students through compulsory tests for all students attending school years 4, 7, and 10 (representing primary, secondary, and upper secondary school, respectively). Student texts were rated by the students' own teachers, and a small proportion was rated by another teacher to allow for the estimation of rater reliability. The program was discontinued as early as 2006 following advice from external reviewers that the results be disregarded because of low rater reliability (Lie, Hopfenbeck, Ibsen, & Turmo, 2005). The Norwegian government was faced with the dilemma of either excluding writing from the battery of national tests of key competencies or investing heavily in a new design.

In 2010, the Norwegian Writing Centre (NWC) was commissioned by the Norwegian Directorate for Education and Training to develop the National Sample-Based Writing Test (NSBWT), which was to be administered annually to a national representative sample of students in primary and lower secondary school (NSBWT-5 for school year 5 and NSBWT-8 for school year 8). The NWC was also commissioned to set up a national panel of raters (NPR), consisting of teachers, with the goal of 1) establishing a strong interpretive community (see below), where teachers shared beliefs about writing proficiency and text quality, and 2) having in place a panel that would reliably rate the NSBWT.

1.1 Interpretive communities

The term "interpretive community" was introduced by Fish (1980) and then later advanced by White (1984) and Berge (2002) to denote a group of raters who share beliefs about how to judge pieces of student writing and what features are associated with different levels of writing proficiency, i.e., interpreting texts in a similar fashion with similar norms. In other words, the presence of an interpretive community can be related to the absence of rater effects. Such effects can manifest through reliability estimates and other associated quantitative investigations (Borgström & Ledin, 2014) or by using qualitative data that demonstrate the ways in which raters use the assessment materials (e.g., rubrics and benchmark texts) as well as their perceptions of text quality (cf. Jølle, 2015). As previous research on communities of practice has shown (e.g., Nielsen, 2008), it is also important to include participants' perception of the purpose of being part of a community. For example, different motivations for participating in a rating panel can contribute to explanations of variation in behavior, such as level of effort.

Somewhat discouraging, previous research has shown the difficulty in establishing interpretive communities in assessment writing, i.e., in the sense of raters acting interchangeably (e.g., Björnsson, 1960; Berge, 1996; Borgström & Ledin, 2014; Eckes, 2015). While there are studies, such as that of Brown, Glasswell, and Harland (2004), indicating that rater training may facilitate increased reliability, others have shown a rather marginal effect of such training (e.g., Elder et al., 2007; Goodwin, 2016; Lumley & McNamara, 1995; Purves, 1992; Trace, Meier, & Janssen, 2016; Weigle, 1994, 1998). For example, Weigle (1994, 1998) showed that training had a positive impact on raters' understanding of criteria and intra-rater reliability, but did not eliminate inter-rater variability. Elder et al. (2007) and Purves (1992) reported similar findings.

Related research has shown that there are numerous more or less consistent bias effects in relation to rater practice (cf. Myford & Wolfe, 2003) that may hinder the development of interpretive communities. This includes rater background (Leckie & Baird, 2011; Lim, 2011), rater cognition (Baker, 2012; Eckes, 2012; Zhang, 2016), and rater values and expectations (Baker, 2010). For example, studies have shown a general lack of grammatical metalanguage among teachers, making it difficult both for students to learn to write (Myhill, Jones, & Watson, 2013) and for teachers to discuss and assess student texts (e.g., Matre & Solheim, 2015, 2016). Research on rater decision-making has revealed a tendency for raters to pay insufficient attention to the common framework with which they are presented and to value more highly their own, individually held, and often tacit assessment practices (e.g., Wyatt-Smith, Klenowski, & Gunn, 2010; Jølle, 2014). Related to this, Eckes (2012) linked rater cognition to rater behavior and demonstrated that the perceived importance of the criterion/rating scale had a systematic impact on rater severity. Moreover, Zhang (2016) showed that rater metacognition was associated with rating accuracy. These findings have been corroborated with those from other

fields—for example medicine (e.g., St-Onge et al., 2016)—where similar levels of rater variability have been shown.

There are indications from previous research that rater training in general tends to be overly short (Lim, 2011) and that it requires prolonged investment to develop a professional interpretive community (e.g., Colombini & McBride, 2012; Skar, Evensen & Thygesen, 2017; Smaill, 2012). The present article explores the assumption of time as an important factor in establishing interpretive communities and uses multiple data sources to investigate potential variation among a sub-sample of NPR members. To our knowledge, there are no studies investigating rater programs where the same raters were trained for as long as those in the NWC program (i.e., five years). The first reliability estimates from the autumn of 2010 indicated large variation within the NPR, with an Intraclass Correlation Coefficient (ICC, one-way random single measures) of .46 (Fasting, 2011). In this study, we investigated rater variation five years later.[1] Such an investigation is interesting far beyond Norway; in all educational contexts, researchers and decision-makers require access to case studies that provide insights into ways of trying to reduce rater variation, thereby strengthening overall dependability.

1.2 Research Questions

To investigate potential variation among a sub-sample of NPR members, this study was conducted with both quantitatively- and qualitatively-oriented research questions in line with what is becoming something of a common practice in research on raters of writing (Baker, 2012; Trace et al., 2016; Zhang, 2016). By combining quantitative and qualitative data, we aim to yield a nuanced depiction of potential variation among the investigated raters. The following two research questions were formulated:

- To what extent were the raters consistent in their ratings of student texts?
- To what extent did the raters vary in their reported practice with regard to: the aims of being a part of the rating panel, their understanding of text quality, and their use of the assessment materials?

2. METHODOLOGY

2.1 Context of the study

The NPR has consisted of some 80 teachers who were recruited from schools across Norway with about 45 raters for NSBWT-5 and about 35 for NSBWT-8. The first year (May 2010–June 2011) there were in average 40 (SD = 3.0) raters at each

[1] *The ICC reported in Fasting (2011) was based on ratings of 156 student texts by 17 raters from the NPR. Each text was rated by two raters. The one-way random model was used because not all raters rated all texts.*

meeting, and the following years there were in average 85 raters (SD = 8.0) at each meeting. The large increase from year one to year two was related to recruitment of new raters. Later variation in number of raters was related to natural circumstances as sickness, parental leave etc., as well as occasional drop-outs. All raters were recruited on the basis of recommendations from headmasters or other teachers. The vast majority were mother tongue teachers or had experience of mother tongue education.

The NPR gathered for two-day or three-day workshops/rating sessions seven times the first two years and thereafter twice a year, during which the teachers underwent extensive training. Consistent with how resources for writing assessment are introduced in schools (but perhaps contrary to common rater training practice (cf. Meadows & Billington, 2005)), part of each workshop was conducted without supervision to allow NPR members to act as relatively autonomous teachers-as-raters.

All workshops were organized in the following way. First the NWC members delivered lectures on each of the following five topics: the writing construct definition (the so-called Wheel of Writing; see Berge, Evensen, & Thygesen, 2016), the rating scales (see Evensen, Berge, & Thygesen, 2016), benchmark texts representing different proficiency levels, texts that had proven to be difficult to assess, and writing instructions. The construct definition, the rating scales, and the benchmark texts formed the "NSBWT assessment material."

Second, the teachers engaged in the assessment of texts in rater pairs (Jølle, 2014). During a workshop, each teacher would be paired with three or four others, with whom he or she had not worked earlier, i.e., engaging in three to four pairs within a rotation system. Each pair would rate the texts of approximately 10 students, which means that each teacher rated between 40 and 50 texts per workshop. Rating in pairs allowed for in-depth discussion about task fulfillment, the construct, and the rating scales. There were three major objectives regarding rater pairing and how it was organized: first, to enhance raters' ability to verbalize text quality by using concepts from the rating scales; second, to strengthen each rater's ability to identify relevant text features using the rating scales; and third, to avoid pairs from developing sub-practices.

2.2 Participants

For practical reasons, and because of economic constraints, it was decided that eight raters from the 33 NPR members rating the NSBWT-8 in 2015 (i.e., five years after the establishment of the NPR) could be invited to rate the same 50 student texts and participate in the study. Because of the small number of participants—eliminating any possibilities for group statistics—and because all members of the NPR had at least two years of NPR experience, the decision was taken to randomly select the eight participants. The mean age was 49.4 years (*SD* = 13.5), and the mean level of teacher experience was 20.2 years (*SD* = 11.0). There were five wom-

en and three men (Table 1). Although all eight raters agreed to participate, one participant (Rater 2) dropped out before the data collection was completed. The rater did, however, allow the researchers to use the assessment data.

Table 1. The participants

Rater	Age	NPR time	Work Experience (yrs)	Gender
R1	66	5	29	Female
R2	52	5	20	Female
R3	36	2	11	Male
R4	64	5	39	Male
R5	30	2	3.5	Male
R6	61	5	25	Female
R7	43	2	16	Female
R8	43	4	18	Female
Mean	*49.4*	*3.8*	*20.2*	
SD	*13.5*	*1.5*	*11.0*	

2.3 Data collection

The data used in this study stem from actual ratings of the NSBWT-8, a test given to students in school year 8, which was administered in 2015. The test consisted of two tasks, and 351 students from 21 schools across Norway participated. Task 1 was an expository essay on why smoking was more acceptable some decades ago than today, and Task 2 was a narrative on what happened one dark night when the main character saw a mysterious light (see Appendix 1). All the students undertook both tasks, and all texts were rated independently by two NPR members. Members of the NPR received text packages and guidance material (descriptors, benchmark texts) electronically and conducted the rating within two weeks at a location of their choosing. The texts were distributed to the NPR in such a fashion that each student faced four raters, two raters for each text. In turn, each NPR member received the work of students from all participating schools. The raters registered the ratings on an NSBWT webpage. Each rater rated 50 student texts and received €10.00 for each rated text.

The ratings were done analytically on six rating scales: *communication*, *content*, *text structure*, *language use*, *spelling*, and *punctuation*, which consisted of descriptors for five proficiency levels. Proficiency level 1 represented the lowest level of mastery and proficiency level 5 the highest level of mastery. There was also a level 0, which represented "impossible to assess." The NWC combined the 24 scores for each student (2 tasks * 2 raters * 6 rating scales) in order to arrive at a "fair average score" (Linacre, 2013), in which task difficulty and rater severity were controlled for (see Appendix 2 for the scales).

The data collection procedures involved the collection of two datasets, the first of which was quantitative and consisted of ratings. All eight raters rated the same

50 student texts, representing 25 students from school year 8 who had completed the two tasks. All texts received ratings on the six rating scales, resulting in 2,400 ratings.

The second dataset was qualitative and consisted of semi-structured interviews and live ratings in which seven of the eight raters participated. Both authors conducted the interviews, with author 2 asking questions and author 1 taking notes (refer to Appendix 3 for the questions included in the interviews). For the live ratings, one of the student texts from the rated text package was presented to the raters. This text was chosen because it represented the average score among the 50 texts. During the live rating sessions, the raters were asked to read and simultaneously think aloud, giving their impressions of the text, following familiar think-aloud procedures (e.g., Cumming, Kantor, & Powers, 2002; Huot, 1993; Lumley, 2005). Even though such an approach has its weaknesses, such as potential reactivity and veridicality, it is deemed to be the best tool to obtain information about raters' rating processes (cf. Barkaoui, 2011). The raters were asked to take their knowledge of the six rating scales as their point of departure, but were otherwise not further instructed on how to read, nor were they handed printed copies of the rating scales (see Appendix 3).

The interviews and live ratings were conducted three months after the ratings were handed in. Preferably, the interviews should have been conducted closer in time to the ratings, but that was not possible due to practical reasons. Each session lasted approximately 30 minutes. All interviews and live ratings were audio-recorded and transcribed by the researchers.

The quantitative and qualitative data collection followed a convergent design (cf. Moeller, Creswell, & Saville, 2016; Tashakkori & Teddlie, 2003). Both data sets were collected and given equal weight, without letting one data set inform the design of the other. Both the quantitative and qualitative data were analyzed independently before being merged for a concluding interpretation.

2.4 Data analysis

To answer research question 1, the quantitative data were first analyzed using an Intraclass Correlation Coefficient (ICC) based on all ratings. Correlation coefficient values exceeding .70 are commonly interpreted as indicating reliability above a minimum level.

Second, the data were fitted to a many-faceted Rasch measurement model (MFRM model) using the computer program Facets 3.71.4 (Linacre, 2014). In the basic Rasch model (Rasch, 1980), the probability of a correct response to a dichotomous item is a function of the difference between the test taker's ability and the difficulty of the item. The MFRM extends this premise, allowing the researcher to model the impact of additional facets, such as rater severity and scale step difficulty (Linacre, 2013). The MFRM is often used in writing assessments because of its

suitability for messy assessment situations where scores are contingent on human qualitative judgment (Barkaoui, 2014; Eckes, 2015; McNamara, 1996).

To fit the writing assessment data to the MFRM model, the Facets program 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, and so on). This is graphically depicted in the so-called 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 has rated. In that way, the analysis of rater reliability permits results in which known sources of variability (i.e., students, texts) are controlled for. The following MRFM model (Engelhard, 2013; Linacre, 2013) was used in this analysis:

$$log(P_{nmijk}/P_{nmijk}–1) = B_n – D_m – E_i – C_j – F_x,$$

where P*nmijk* represents the probability of student *n* on task *m*, rating scale *i*, by rater *j* receiving a score of *k*, and P*nmijk*-1 represents the probability of the same student under the same conditions receiving a score of *k*-1. B_n is the ability for person *n*, D_m is the difficulty of task *m*, E_i is the difficulty of rating scale *i*, and C_j is the severity of rater *j*. Finally, F_x represents the point on the logit scale where category k and k-1 are equally probable.[2]

The Facets output yields a number of useful graphs and statistics. The variable map provides visual information on the extent to which raters share levels of severity. The interpretation of the map is aided by different "separation statistics," which estimate the possibility of separating raters into different severity levels. First, the "fixed (all same) chi-square" tests the hypothesis that all raters shared 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 "reliability" measure is analogous to Cronbach's alpha or "test reliability."

The Facets output also generates descriptive statistics of *category use*, which can be used in conjunction with separation statistics to gain insights into the raters' category use across rating scales. The output also reports *percent agreement* and *correlations of single rater–rest of raters* (SR–ROR), indicating the extent to which raters rank students in a similar fashion.

Measures of data model fit, infit, and outfit indicate the degree of internal rater consistency. Good fit indicates that the model can predict rater behavior, which in turn implies intra-rater consistency. The two indices, infit and outfit, indicate the

[2] *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, therefore, 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).*

extent to which the model can predict raw score observations. The model-expected value is 1.0, and underfit (i.e., deviation from the MFRM model) is indicated when the fit statistic exceeds this. Low fit values indicate less than optimal variation, i.e., a restricted use of the rating scale. However, fit values in the range from 0.50 to 1.50 are generally acceptable (see Bond & Fox, 2015; McNamara, 1996).

Finally, the so-called differential rater function statistics, or bias analysis, enables the researcher to investigate the possible interaction between raters and rating scales. The analysis builds on the relationship between observed and expected values (for technical details, see Linacre, 2013). Non-technically speaking, a significant bias measure indicates that a rater systematically awarded unexpected ratings to whichever facet is part of the analysis. For example, a rater might be unexpectedly harsh with spelling.

To answer research question 2, the qualitative data were subjected to two different procedures. First, the individually held semi-structured interviews were transcribed and coded for instances within the following categories: a) motivation for joining the NPR and perception of the task, b) perception of student text quality, and c) understanding of assessment materials. Accordingly, the researchers allowed the interview guide to set the premise for the different coding categories (see the interview guide, Appendix 3), establishing what Seale (1999) calls "low-inference descriptors." The coding was done by each author individually and was subsequently compared. In the very few cases of discrepancy, consensus was reached through discussion.

The live ratings were analyzed, inspired by Green's (1998) notion of an "idea unit" as "a single or several utterances with a single aspect of the event as the focus." This meant that the think-aloud protocols were segmented into "meaning units," focusing on aspects of the text that were coded as being related to either holistic assessment or the different rating scales. The meaning units were distinguished by shifts in focus, often initiated after extended pauses. For example, Rater 3 started by focusing on spelling and then made a lengthy pause before continuing with the content of the text. This was coded as instances of two separate meaning units whereby the rater focused on different aspects of the text in each unit. The authors did this coding collectively.

3. RESULTS

3.1 RQ1: To what extent were the raters consistent in their ratings of student texts?

The ICC estimate indicated that the raters were not fully interchangeable. The overall ICC single measures was .61 (see Table 2). This neither exceeded the threshold of .70 nor demonstrated a substantial increase from the above-mentioned 2010 measure of .46 reported in Fasting (2011), which had included

other raters. The results therefore suggest that a score from a single rater was not overly reliable (see Table 2).

Table 2. ICC consistency estimates

	ICC Single		ICC Average	
	Estimate	**CI 95%**	**Estimate**	**CI 95 %**
Two tasks (N = 300)	.61	[.57, .66]	.93	[.91, .94]
Expository (n = 150)	.59	[.53, .66]	.92	[.90, .94]
Narrative (n = 150)	.63	[.57, .69]	.93	[.92, .95]

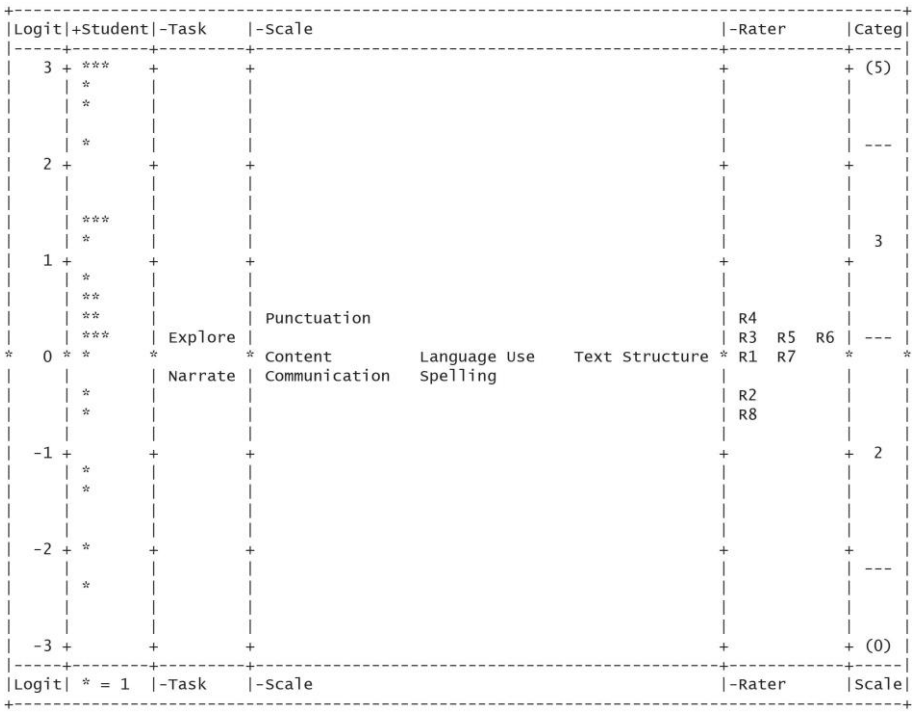
Note. ICC: Intraclass Correlation Coefficient, two-way random model, CI 95 %: 95 % Confidence interval. N: number of ratings (N = 50 texts * 6 scales).

Looking more closely, the results suggest that consistency was somewhat dependent on task type.[3] The values for ICC single measure were slightly higher for the narrative task than for the expository task. When the confidence intervals are taken into account, however, this difference is negligible. It can also be mentioned that the average measure was .93, indicating the expected result that the combined ratings from all eight raters were satisfactorily reliable.

Turning to the MFRM output, the variable map in Figure 1 depicts the logit spread in each facet. Here, the 25 students had different logit values and were separated by a large degree of reliability (Reliability = .99), indicating that the teachers as a group were able to separate students with a high degree of precision (cf. value for ICC average). The variable map also shows that tasks, rating scales, and raters demonstrated a certain amount of spread, indicating that the tasks and rating scales were of varied levels of difficulty and that the raters were not equally severe.

[3] *The ICC (single measure) for the six rating scales (all n = 50) were: communication: .60 [.50, .71]; content: .64 [.54, .74], text structure: .66 [.56, .76]; language use: .65 [.54, .75]; spelling: .60 [.49, .71]; punctuation .65 [.55, .75].*

Figure 1. The Variable Map. From left to right, the logit scale, students ability (more ability higher logit value), the tasks (more difficult, higher logit value) scale (more difficult, higher logit value) raters (more severe, higher logit value) and category (dashed horizontal lines, where probability being observed in category above begins to exceed probability of being observed in category below). CC: consistency estimates



The severity measures for the raters demonstrated a gap of 0.74 logit units between the most lenient rater, Rater 8, and the most severe, Rater 4 (see Table 2). Both exhibited a distance of approximately 1 standard deviation from the group mean. Converted back to the original scale length, the gap between Rater 8 and Rater 4 represented 0.44 score points. Rater 1 and Rater 7 were the most mainstream, deviating only 0.01 points from the average ratings. With regard to the separation statistics (Table 3), the fixed (all same) chi-square, which tests the hypothesis that all raters shared the same severity level, yielded a value of 129.6, with 7 degrees of freedom and $p < .01$. This indicated that the difference in severity did not occur by chance. This result was corroborated by a strata value of 5.51 and the reliability of the separation of 0.94. In other words, it was possible to detect, with a high degree of consistency, more than five distinct classes of rater severity.

Table 3. Raters' severity, agreement, and consistency

Rater	Fair ave-rage	Logit	Model S.E.	Infit	Infit_z	Outfit	Out-fit_z	Agree %	SR-ROR
R1	2.72	0.03	0.08	0.94	-0.60	0.93	-0.80	45.8	.32
R2	2.98	-0.50	0.08	0.59	-6.10	0.57	-6.30	47.4	.32
R3	2.61	0.24	0.08	1.00	0.00	0.99	0.00	44.1	.30
R4	2.57	0.34	0.08	0.74	-3.40	0.75	-3.30	47.6	.32
R5	2.59	0.28	0.08	1.30	3.30	1.29	3.20	40.3	.26
R6	2.64	0.18	0.08	0.98	-0.10	0.96	-0.40	45.6	.26
R7	2.72	0.03	0.08	1.36	4.00	1.36	3.90	41.5	.27
R8	3.04	-0.60	0.08	1.04	0.50	1.03	0.40	45.0	.28
Min	*2.57*	*-0.60*	*0.08*	*0.59*	*-6.10*	*0.57*	*-6.30*	*40.3*	*.26*
Max	*3.04*	*0.34*	*0.08*	*1.36*	*4.00*	*1.29*	*3.90*	*47.6*	*.32*
Mean	*2.73*	*0.00*	*0.08*	*0.99*	*-0.30*	*0.98*	*-0.40*	*44.7*	*.29*
SD	*0.18*	*0.36*	*0.00*	*0.26*	*3.30*	*0.26*	*3.30*		*.03*

RMSE: 0.08, Adj. SD: 0.32, Strata: 5.51, Reliability (not inter-rater): .94
Fixed (all same) chi-square: 129.6, Degrees of Freedom: 7, Significance: .00

Note. Fair average: Facets-generated conversion of logit values to original reporting scale; Model S.E.: Standard error of measurement for element or facet. RMSE: Root Mean-Square Standard Error, Adj. SD: corrected standard deviation.

The category use analysis presented in Table 4 illustrates how the raters used the six categories across all rating scales. The results show that, as might be expected, the raters used the categories differently. For example, the most lenient rater, Rater 8, used categories 4 and 5 for 37% of the ratings, while the most severe rater, Rater 4, used these categories for 15% of the ratings. The table shows some large individual deviations from the group total, for example, Rater 6 used category 2 a total of 42% of the time (compared with 34% for the group total). Rater 1, however, who had a mainstream average score, deviated a maximum of only 4 points (category 4: 15% compared with the 19% group total). Considering the raters as a group, categories 2 and 3 were chosen most of the time, with 87% of the ratings.

Table 4. Category use

	Category 0	Category 1	Category 2	Category 3	Category 4	Category 5	Total
R1	0 %	9 %	36 %	31 %	15 %	8 %	100 %
R2	0 %	2 %	28 %	38 %	26 %	5 %	100 %
R3	0 %	13 %	34 %	32 %	16 %	6 %	100 %
R4	4 %	4 %	31 %	46 %	14 %	1 %	100 %
R5	4 %	7 %	33 %	32 %	20 %	3 %	100 %
R6	0 %	4 %	42 %	36 %	15 %	3 %	100 %
R7	0 %	11 %	36 %	25 %	20 %	8 %	100 %
R8	0 %	3 %	32 %	29 %	27 %	10 %	100 %
Total	1 %	7 %	34 %	34 %	19 %	5 %	100 %

Note. Row "total" equals use of category for group. N = 2400.

The rater agreement statistics indicated that the raters showed suboptimal inter-rater consistency (see also Table 3). The overall exact agreement was 44.7%, and the mean single rater–rest of raters (SR–ROR) correlation was .29. In a Rasch environment, an SR–ROR value below .30 is considered "somewhat low" (Myford & Wolfe, 2003, p. 410). Rater 5 exhibited the least agreement (40.3%) and the lowest SR–ROR (.26), while Rater 4 exhibited the most agreement (47.6%) and a high SR–ROR (.32).

With regard to the fit statistics, the raters also varied in consistency (see Table 2). Two raters exhibited somewhat high (and significant) infit and outfit values, namely, Rater 5 (infit 1.30) and Rater 7 (infit 1.36). However, both would have gone unnoticed had the generally accepted threshold of 1.50 been used. Two raters exhibited somewhat low (and significant) infit and outfit values, namely, Rater 2 (infit 0.59) and Rater 4 (infit 0.74). Rater 2 used categories 2, 3, and 4 for 93% of the ratings, and Rater 4 used the same categories for 91% of the ratings (cf. Table 4). These raters too would have gone unnoticed had the usual benchmark been applied. Rater 1, who displayed mainstream severity and category use, had an infit value of 0.94, indicating an almost perfect fit to the MFRM model.

Table 5 summarizes some findings from the bias analysis. For ease of interpretation, only significant bias has been included, and instead of reporting on bias measures, the table presents the observed–expected average. A positive value indicates that when all known information about the rater and the rating scale is taken into account, a rater has systematically awarded higher ratings than those expected in the model. Conversely, a negative value indicates that the rater has awarded lower ratings than those expected in the model.

Table 5. Bias analysis: observed-expected average associated with significant bias

	Communication	Content	Text Structure	Spelling	Punctuation
R1	-	-	-	-	-
R2	-	-	-	-	-
R3	-	-0.22	-	-	-
R4	-	-	-	-	-
R5	-	-	-0.28	0.48	-
R6	-	0.29	0.27	-0.25	-0.23
R7	-	-	-	-	0.44
R8	0.37	-	-	-	-0.36

Note. *Observed-expected average* are the total score for rater on scale minus the expected score based on the MFRM model, divided by number of ratings. Only observed-expected average associated with significant bias ($t \geq 2.0$) has been included in the table.

According to Table 5, three raters showed non-significant bias (R1, R2, and R4). For Rater 3 and Rater 7, there was significant bias related to one rating scale only, content and punctuation, respectively. For Rater 5 and Rater 8, significant bias was found on two rating scales, text structure and spelling, and communication and

punctuation, respectively. Rater 6 demonstrated significant bias on four rating scales, namely, content, text structure, spelling, and punctuation. On content and text structure, the ratings were higher than expected, and on spelling and punctuation, the ratings were lower. Rater 6 was found to have a severity bias toward formal aspects and a leniency bias toward the more functional aspects of content and text structure. Put informally, Rater 6 seemed to pay less attention to whether a text had accurate content than to evidence of good spelling and punctuation.

The results of the statistical investigation demonstrate that the eight participants varied substantially in their ratings. They varied in the severity and use of the rating scale and demonstrated unsatisfactory inter-rater consistency. The nature of the variations indicated no single outlier but, rather, different "rater profiles." Rater 1 stood out as the mainstream rater, Rater 4 as the consistently severe rater, and Rater 8 as the lenient rater with a leniency bias toward the functional rating scale. Rater 6 seemed to be a mainstream rater in terms of severity and displayed high intra consistency, but had a leniency bias toward functional aspects. As a group, the raters demonstrated sufficient intra-rater reliability, and the combined ratings of the group were trustworthy. However, the bias analysis indicated that five of the participants had a tendency to rate inconsistently harshly or leniently on one or more rating scales, in turn indicating bias toward specific rating scales.

3.2 RQ2: To what extent did the raters vary in their reported practice with regard to: the aims of being a part of the rating panel, their understanding of text quality, and their use of the assessment materials?

The findings from the interviews and live ratings will be presented in the following order: rater motivation and perception of task; perception of text quality; materials, procedures, and facilitation; and live rating.

3.2.1 Rater motivation and perception of task

When asked about motivation, the answers fell into three categories. Some raters saw it as a possibility for *professional development* (R1, R5, and R8). Rater 5, for example, an L1 teacher with a few years of teaching experience, found a level of asymmetry between the minimal rater training during teacher education and the number of rating tasks he was required to undertake as an L1 teacher. He found that being a NPR member enabled him to face the challenges involved in his day-to-day practice as a writing teacher. There were also raters who already saw themselves as *rating experts* and hence found it natural to contribute with their expertise in the NPR (R3 and R4). Rater 4 had worked as an external examiner of national exams for more than 30 years, while Rater 3 had gone through relatively extensive rater training as a result of being involved in a writing development research project. Lastly, there were raters who wanted to be better prepared to *support the students* in their efforts to develop their writing proficiency (R6 and R7). In this re-

gard, Rater 7 expressed that she hoped that her work as a rater in the NPR would better prepare her to get students to understand the necessity of becoming proficient writers.

It is noteworthy that none of the raters mentioned concepts that are essential within an assessment community, such as consistency or reliability, when they explicated their reasons for being a part of the NPR. In fact, only two of the raters related their work to the ratee, and when they did, it was done in a formative manner in relation to students' writing development.

3.2.2 Perception of text quality

When the raters were asked to compare the NSBWT texts with their own students' texts produced outside a test situation, five raters stated that the texts from the test seemed to be about equivalent to the quality of texts used with their own students. "Much the same" was the phrase the raters used. Consequently, in this respect, it is possible to say that the raters shared an overall holistic perception of text quality and that they had a good premise on which to base their communal rating practice. The two raters, Rater 1 and Rater 3, who believed that the NSBWT texts and their own students' texts differed in quality both perceived their own students' performance as better. Their reasoning for this was that their job as raters had made them better writing teachers. They believed that they had managed to invest the knowledge they had acquired as NPR members in their job as teachers who taught students to write. Rater 1 stated that this was because she had "been working systematically with writing for three years now."

However, although the raters seemed to share a perception of text quality at a holistic level, they differed at an analytical level. When asked what the single hallmark of well-written texts was, two raters said text structure (R3, R8), two communication (R1, R7), and two content (R5, R6), while one rater (R4) declined to choose one of the predefined rating scales and answered: "Something that strikes you." The raters also differed about the areas in which students, in general, needed to improve. They proposed text features related to all the rating scales except spelling. For instance, while Rater 7 identified punctuation as the weak spot, Rater 8 stated that formal aspects were not that important and that effort needed to be put on communication. Rater 4, however, mentioned that students in general needed to write longer texts. These differences are indicative of variation in perception of text quality at an analytical level.

3.2.3 Materials, procedures, and facilitation

The two most important materials in the NPR were the rating scale descriptors and the annotated benchmark texts. The descriptors delineated writing proficiency within different proficiency levels, and the annotated benchmark texts were examples from which the raters could see how the descriptors should inform the ratings.

A third tool, an e-forum, was made available to allow for general discussion about ratings.

All raters regarded the descriptors as the single most important assessment resource that assisted them in their rating practice. That said, they used the tool in different ways. Some used the descriptors all the time as their dipping rod (e.g., R6); some claimed to have internalized them (e.g., R4), using them as a sort of tacit knowledge (Polanyi, 1966); some used the tool in specific cases when in doubt about a level of mastery, as a control mechanism (e.g., R3 and R7); while others questioned the quality of the descriptors and hence their usefulness. In this regard, Rater 1 stated that she liked having descriptors of mastery levels but that she did not always use them because she did not "always find them that good."

The benchmark texts were also used in different ways: from raters stating that they used them "more and more" (R1), and raters who used them "when in doubt" (R7), to raters who questioned the quality of the benchmark texts. Rater 3 rather diplomatically stated that he did "not always agree; sometimes, they are more lenient than I am," and that, in such situations, he did "not always follow the resource." Conversely, Rater 8 had a much harsher evaluation of the benchmark texts, saying: "I usually assess the exemplar before reading the 'official' mark—often, my assessment and the official mark do not coincide—sometimes, there's a difference of two points."

When asked in which order they rated the different rating scales, the raters reported different procedures. Three raters (R1, R5, and R7) said that they followed the sequence given in the assessment rubric, i.e., *scheme based* and *analytic*. Two raters (R3 and R6) stated that they had no specific order, but just made notes of what they saw in the text while reading it several times, i.e., *text based* and *analytic*. The last two raters (R4 and R8) reported that they assessed formal aspects first on weak texts and communication and content first on good texts. This finding revealed an unintended practice: these two raters seemed to have made a first evaluation of the text, which determined how they went about assessing the different rating scales, i.e., a prior *text based* and *holistic* assessment and a subsequent *analytic* assessment.

With regard to the practical arrangement of the ratings, the raters were free to do the work as they preferred. All raters reported that they usually rated a considerable number of texts "at one sitting," explaining that they liked "to get it under their skin" or that they had to "work [their] way into [the assessment]." However, if the raters needed to rate some texts to "warm up," this threw up the probability of rating variation depending on where the different student texts were located in the pile.

It is possible to reduce this effect by re-rating and comparing student texts systematically, which is partially what some of the raters did. Three of the raters (R3, R7, and R8) said that they re-rated and compared texts. For instance, Rater 7 stated that she thought it was good that they "have several texts from the same task because *even though you are not supposed to compare*, a very clear pattern gradually

appears—where each text finds its place" (our italics). Others solved this very differently. Several of the raters believed that they were not allowed to compare texts (R4, R5, and R7 (see the quotation above)), a supposition possibly related to an idea of a "pure" criteria-based rating procedure. This understanding accommodates Rater 5's statement that he would "assess each text objectively and just work with that."

3.2.4 Live rating

To illustrate how the think-aloud protocols were segmented into distinguishable "meaning units" for every rater, Table 6 includes excerpts from the first four units. While not all raters explicitly mentioned the names of the rating scales, all utterances have, where possible, been coded using the rating scales as categories. Meaning units related to an overall judgment have been coded as "holistic." Each cell thus represents a meaning unit.

Table 6 indicates both differences and similarities in the ways in which raters read the text. In terms of the differences, four raters started by making a holistic judgment or observation (R4, R5, R7, and R8) and two by reading the text from a communicative perspective (R1 and R6). Rater 3 first noticed spelling. This difference was accentuated when the raters moved through the text. For example, Rater 1 continued to focus on communicative aspects, while Rater 3 moved between spelling, content, text structure, and content again. The differences, however, are recognizable from the interviewees' reported practice. For instance, Rater 3 stated that he had "no specific order" and that it "depends on the texts." The same concurrence between reported practice and practice was found with Raters 1, 4, 7, and 8, indicating a well-developed knowledge about their own rating behavior.

The excerpts also show that Rater 6 and Rater 8 seemed to disagree about the students' use of language. While Rater 6 deemed it to be oral, Rater 8 emphasized the use of "good words," such as *nicotine*. Further, Rater 3 and Rater 5 held opposite views on whether the text had sufficient or poor links between their opening and closing paragraphs (see column "Meaning unit 3").

Both on and beneath the surface level, there were, however, some striking similarities. First, all participants moved through the text using multiple perspectives, i.e., seemingly using internalized versions of the rating scales as different lenses through which the text could be read. Moreover, some features caught the attention of several raters, for example, the absence of a good title (R1, R5, R6, and R7) or a student's inclusion of questions meant to appeal directly to a reader (R1, R4, R5, R6, R7, and R8).

Thus, the live rating showed differences that were expected when reviewing the results from the quantitative analysis of the ratings and from the interviews. The live rating also shed light on some commonalities such as a non-linear reading/rating strategy and clear signs of abilities to view the text from multiple perspectives, something that had been stressed in the NPR workshops.

Table 6. Excerpts from live rating protocol

	Meaning unit1	Meaning unit2	Meaning unit3	Meaning unit4
R1	S1: "I think he communicate well. Ask questions to me as reader. Bad title."	S1: "I believe he is exploring. Hm, he is more stating than exploring"	S1: "He turns to-, get the reader's attention from the beginning."	S1: "I believe he is good on Communication."
R3	S5: "What I see immediately is spelling."	S2: "I see that it takes some time before he starts to explore why it was more common before."	S3: "I see that the text structure has an opening and a clear conclusion, with a clear structure"	S2: "He has good content elements here [...] The first part is not particularly relevant."
R4	Holistic: "I first see that this is a pretty good text"	Holistic: "When I look at a text like this the first thing I notice is the layout. You can tell a lot from that. It has paragraphs."	S1: "This text is meant to be a part of a booklet. This is a writer who turns directly to the reader."	S2: "He raises a lot of questions. Many students never answer these questions."
R5	Holistic: "The first thing I notice is that the text looks inviting."	S1: "No heading, but the first paragraph speaks to me [...] I try to see if the text is exploring [...] No title, asks some questions in the beginning"	S3: "Not very strong link between the introduction and the conclusion."	S3: "There is no summary, when it comes to content."
R6	S1: "The text lacks a signal, a title."	S1: "Turns to a reader."	S1: "Very good connection from one paragraph to the next – I am still in communication."	S4: "The language is a bit oral. But if he writes for his peers, then the language is suitable."
R7	Holistic: "I see [...] length, title and structure, as an overall first impression	S1: "It does not have a title that communicates."	S1: "Asks questions, good. Have a reader in mind."	S5: "No big errors when it comes to spelling."
R8	Holistic: "I see paragraphs"	S1: "I see a question mark – a communication thing; I see a fine sentence, 'in this text you will'."	S4: "Long sentence in first paragraph, but good words, like 'nicotine' and 'determine to quite with' – I wonder if the sentences are too long."	S1: "Communicates with reader using 'you'."

Note. *S1* = Scale 1, "Communication"; *S2* = Scale 2, Content"; *S3* = Scale 3, "Text Organization"; *S4* = Scale 4, "Use of Language"; *S5* = Scale 5, "Spelling"; *S6* = Scale 6, "Punctuation".

4. DISCUSSION

When teachers form an interpretive community, they rate student texts in a similar manner. They also express similar understandings of their roles as raters and express and display similar ways of using the assessment materials and perceptions of

text quality. The quantitative investigation showed large variation among the raters. Although several statistics were employed, it was impossible to identify a pattern, in that, no rater stood out as a systematic outlier. Rather, as suggested above, the results indicated different rater types, all contributing noise to the ratings in different ways.

The qualitative investigation also revealed variation. First, there was a tendency among the raters to view themselves as teachers rather than raters. This view seems to have affected the raters' approach to the NPR work (cf. Nielsen, 2008). For example, Rater 7 reported that a main objective for participating in the NPR was to gain knowledge about writing instruction, indicating a strong classroom focus rather than a predominant rating focus. Others observed that membership of the NPR had resulted in systematic writing instruction. For example, Rater 1 and Rater 3 claimed that because they were part of the NPR, their own students performed better than the average NSBWT student. In this connection, Rater 8's ignorance of the benchmark text makes sense: as a professional teacher, it was her prerogative to take a critical stance toward the rating materials included in the NSBWT. In his influential work, Sadler (1987, 2011) has emphasized the need for the use of descriptors and benchmark texts to overcome arbitrary and idiosyncratic ratings. In the present study, when the raters' self-reported use of these rating materials are taken into account, a more complex picture emerges. The rating process appeared not to be a matter of applying or not applying materials. Rather, the raters appeared to perceive the materials in different ways for different reasons.

The raters also expressed divergent understandings of significant text features, listing features related to either one of three rating scales. Conversely, Rater 4 stated that high-quality texts included "something that strikes you." This rater, who also claimed to have internalized the rating scale descriptors, thus made use of concepts that were not included in the official document. His answers indicated a tension between the aim of the rater training, which focused on raters' ability to verbalize text quality using descriptors, and his perception of his use of the NSBWT materials. In addition to Rater 8's disregard for the annotated benchmark texts, Rater 5 described how the descriptors failed to account for actual text quality.

When the results from the two datasets are combined, we get an even better understanding of the often contradictory complexity, as in the following examples. Rater 4 showed a non-significant bias on the different rating scales but, at the same time, expressed quite idiosyncratic opinions about what constitutes text quality. Based on the interview, we might have expected Rater 4 to be an outlier, but the rating data revealed that despite his unusual ideas about text quality, he did not show an alternative rating pattern. The bias analysis also showed that spelling and punctuation were overrepresented when it came to bias, but six out of seven raters in the interviews stated that these two rating scales were the easiest to rate. Again, based on the interviews, we could be tempted to believe that spelling and punctuation are easy to agree upon, yet the rating data tell us another story. Moreover, we have seen that Rater 1, who stood out as the mainstream rater, was the one who

disregarded the descriptors in situations when she found them to be of poor quality. It is no understatement to say that it is a surprise that a rater with a reported tendency to neglect the descriptors can stand out as the mainstream rater. The lesson learned is that a mixed dataset does "provide a more complete understanding" (Creswell & Plano Clark, 2011, p. 8).

Overall, however, although the participants undoubtedly varied, both the quantitative and qualitative results suggested that they functioned reasonably well as a collective. This was indicated by the good data model fit in the MFRM analysis, which strongly suggested that the collective judgment of student texts was reliable and trustworthy. This conclusion was also supported by the high average ICC value. Moreover, most raters demonstrated high internal consistency, which was indicated by the infit/outfit values and the strong links between reported and displayed practice in the interviews and live ratings, respectively. High intra-rater consistency indicates internalization of the assessment rules and the ability to apply them in a reliable manner. The think-aloud procedure revealed common rating competence, in that, each participant was able to use the rating scales to view the text from multiple perspectives.

Where do such results leave us? It has been demonstrated that the NPR members participating in this study varied with regard to the rating of student texts and the reported use and understanding of the NSBWT materials. Despite a generously funded program in which raters had undergone years of training with a focus on the construct, the rating scales, and arenas for negotiating the conception of text quality, the raters did not appear to constitute the interpretive community that NWC had anticipated. In that respect, these results corroborate those of previous studies investigating the rating of student texts, either among professional raters (e.g., Berge, 1996; Eckes, 2015; McNamara, 1996; Weigle, 1998) or teachers (e.g., Björnsson, 1960; Borgström & Ledin, 2015; Skar et al., 2017). The findings also corroborate previous results insofar as the raters displayed a sufficient degree of intra-rater consistency.

In writing assessment research, there are, to date, no good answers to the question: "What kind of rater training works?" It is obvious that the model proposed by NWC did not work as well as intended. The reasons for this might be many. For example, the rater pair discussions might not have challenged the teachers' own understanding of text quality but, instead, might have served as social arenas for demonstrating perceptions rather than participating in negotiations (cf. Jølle, 2015). The NWC may have disproportionately stressed the teachers as experts. Of course, the teachers were and are experts in the sense that student texts are a part of their professional everyday life. However, the teachers were not, and could not have been, experts in the sense of being familiar with participating in such a rating panel, since it was the first of its kind. Moreover, it may be that the NWC model did not fully take into account the need for metalanguage in assessment conversations (cf. Matre & Solheim, 2016), leaving teachers to cope independently with the demanding task of conveying opinions on student text quality

using assessment material devised by an external party. Conversely, being a teacher often involves reacting to and acting with externally produced material and, when necessary, developing adequate skills in the use of such material.

Although these are interesting questions, they are perhaps not properly targeted. It might be appropriate to pose different ones such as: "What can we expect of rater training?" and therefore, "What should be the goal of rater training?" Ultimately, "What is good enough?" If this or any other rater training program fails to produce the outcome that teachers/raters can be used interchangeably, then perhaps we must accept rater variation, just as we acknowledge that there is a large student-by-task variation. If programs establish rater panels in which members rate students' text using the same tools in a similar manner, with a high degree of internal consistency, then multiple rating and statistical procedures can be employed to compensate for deviations in score points. Such a thought is provocative, however, because of its implications for classroom assessment. It dismisses the teacher as representing the collective of teachers.

This study indicates that there is a need for further research into the area of rater-mediated assessment. This is also true when considering the caveats associated with this study. First, the sample of raters was limited to eight raters or 10% of all NPR members. Thus, the scope of the investigation somewhat limits the possibilities for making inferences about the NPR population. Second, the gap between ratings and interviews should preferably have been narrower.

5. CONCLUSION

This article has presented an investigation framed by two research questions: to what extent were the raters consistent in their ratings of student texts? To what extent did the raters vary in their reported practice with regard to: the aims of being a part of the rating panel, their understanding of text quality, and their use of the assessment materials? Findings of inconsistency and variation are not equivalent to labeling the NPR as a failure. Indeed, the NWC has not succeeded in turning NPR members into "rating machines" (Linacre, 2013) or mechanical (and uncritical) users of NSBWT rating materials. While such drastic results were not the aim of the NWC, it was its belief that long-term engagement would make the ratings less heterogeneous. If that were the case, the implications for other rating programs and for K–12 education would be that it is possible to foster an interpretive community of teachers who each can award grades that reflect the opinions of that community. Instead, as the results of this study corroborate the findings of many other rater reliability studies, one might be driven to conclude that students and other ratees deserve to be rated by several independent raters and that appropriate statistical techniques should be utilized to compensate for rater effects (cf. Eckes, 2015).

Importantly however, what this study has also shown is that it is possible to train teachers to be internally consistent and to use the same assessment materials to rate texts, apparently without losing the sense of being a professional teacher

with autonomy. This implies that this type of rater training can be a fruitful endeavor if it is accepted that a single trained rater may not replace a strong community.

ACKNOWLEDGEMENT

The authors would like to thank two anonymous reviewers for constructive comments in the review process. We would also like to thank Lyle F. Bachman, who inspired us to conduct this study.

REFERENCES

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing*, *15*(3), 133–153. https://doi.org/10.1016/j.asw.2010.06.002

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225–248. https://doi.org/10.1080/15434303.2011.637262

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*, 51-75.

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1301–1322). Chichester, UK: Wiley-Blackwell.

Berge, K. L. (1996). *Norsksensorenes tekstnormer og doxa: en kultursemiotisk og sosiotekstologisk analyse*. (Doktorsavhandling, Norges teknisk-naturvitenskapelige universitet).

Berge, K. L. (2002). hidden norms in assessment of students' exam essays in norwegian upper secondary schools. *Written Communication, 19*(4), 458–492. https://doi.org/10.1177/074108802238011

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

Björnsson, C. H. (1960). *Uppsatsbedömning och uppsatsskrivning [The assessment and writing of esssays]*. Stockholm, Sweden: Almqvist & Wiksell.

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

Borgström, E., & Ledin, P. (2014). Bedömarvariation. Balansen mellan teknisk och hermeneutisk rationalitet vid bedömning av skrivprov. *Språk och Stil*, *24*, 133–165. [Rater variation. The balance between a technical and a hermeneutical rationality in writing assessment].

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

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105–121. https://doi.org/10.1016/j.asw.2004.07.001

Colombini, C. B. & McBride, M. (2012). "Storming and norming": Exploring the value of group development models in addressing conflict in communal writing assessment. *Assessing Writing, 17*, 191-207.

Coffman, W. E. (1971). Essay examination. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 271–302). Washington, DC: American Council of Education.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cumming, A., Kantor, R. and Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67-96.

Eckes, T. (2012). Operational rater types in writing assessment: linking rater cognition to rater behavior. *Language Assessment Quarterly: An International Journal*, *9*(3), 270–292.

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

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37–64. https://doi.org/10.1177/0265532207071511

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

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

Fasting, R. B. (2011). *Nasjonale utvalgsprøver i skriving som grunnleggende ferdighet. Vurdererfellesskap og reliabilitet ved vurdering av skriving.* [The National Sample-Based Writing Test. On the issue of interpretive community and reliability of writing assessment.]. Trondheim, Norway: Nasjonalt senter for skriveopplæring og skriveforsking.

Fish, S. (1980). *Is there a text in this class? The authority of interpretive communities*. Cambridge, MA: Harvard University Press.

Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. New Jersey, NJ: Aldine Transaction.

Green, A. (1998). Verbal protocol analysis in language testing research: A handbook (Studies in language testing, Vol. 5). Cambridge, UK: Cambridge University Press.

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing*, 20(1), 37–52. https://doi.org/10.1016/j.asw.2014.01.002

Jølle, L. (2015) Rater strategies for reaching agreement on pupil text quality. *Assessment in Education: Principles, Policy & Practice,* 22(4), 458-474. http://dx.doi.org/10.1080/0969594X.2015.1034087

Leckie, G. & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater Experience. *Journal of Educational Measurement, 48*, 399-418.

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

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*, 543-560.

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

Linacre, J. M. (2014). Facets® (version 3.71.4) [Computer Software]. Beaverton, OR: Winsteps.com.

Lumley T. & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*, 54-71.

Matre, S., & Solheim, R. (2015). Writing education and assessment in Norway: Towards shared understanding, shared language and shared responsibility. *L1 Educational Studies in Language and Literature*, *15*, 1–33. https://doi.org/10.17239/L1ESLL-2015.15.01.05

Matre, S., & Solheim, R. (2016). Opening dialogic spaces: Teachers' metatalk on writing assessment. *International Journal of Educational Research*. https://doi.org/10.1016/j.ijer.2016.07.001

McNamara, T. F. (1996). *Measuring second language performance*. New York, NY: Longman.

Meadows, M. & Billington, L. (2005). *A review of the literature on marking reliability*. Retrieved from https://cerp.aqa.org.uk/research-library/review-literature-marking-reliability

Moeller, A. J., Creswell, J. W. & Saville, N. (Eds.) (2016). *Second language assessment and mixed methods research*. Cambridge, UK: Cambridge University Press.

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.

Myhill, D., Jones, S. & Watson, A. (2013). Grammar matters: How teachers' grammatical knowledge impacts on the teaching of writing. *Teaching and Teachers Education, 36*, 77–91

Nielsen, K. (2008). Learning, trajectories of participation and social practice. *Critical Social Studies, 10*(1), 22–36. http://ojs.statsbiblioteket.dk/index.php/outlines/article/download/1965/1755

Purves, A. C. (1992). Conclusion. In A. C. Purves (Ed.), *The IEA study of written composition* (Vol. 2, pp. 199–203). Oxford, UK: Pergamon.

Rasch, G. (1980). *probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970. https://doi.org/10.1037//0021-9010.85.6.956

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. van Steendam, M. Tilema, & G. Rijlaarsdam (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. 1–22). Leiden, The Netherlands: Brill.

Seale, C. (1999). *The quality of qualitative research*. London, UK: Sage.

Smaill, E. (2012). Moderating New Zealand's national standards: teacher learning and assessment outcomes. *Assessment in Education: Principles, Policy & Practice*, 1–16. http://doi.org/10.1080/0969594X.2012.696241

Skar, G. B., Thygesen, R., & Evensen, L. S. (2017). Assessment for learning and standards: A Norwegian strategy and its challenges. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard setting in education, methodology of educational measurement and assessment* (pp. 225–241). New York, NY: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-50856-6_13

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in Health Sciences Education, 21*(3), 627–642. https://doi.org/10.1007/s10459-015-9656-3

Tashakkori, A., & Teddlie, C. (Eds.) (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage.

Thygesen, R., Evensen, L. S., Berge, K. L., Fasting, R. B., Vagle, W., & Haanæs, I. R. (2007). *Nasjonale prøver i skriving som grunnleggende ferdighet. Sluttrapport*. [National tests of writing as a key competence. Final report]. Nasjonalt senter for leseopplæring og leseforskning, Universitetet i Stavanger.

Trace, J., Meier, V., & Janssen, G. (2016). "I can see that": Developing shared rubric category interpretations through score negotiation. *Assessing Writing*, *30*, 32–43. https://doi.org/10.1016/j.asw.2016.08.001

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223. https://doi.org/10.1177/026553229401100206

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

White, E. M. (1984). Holisticism. *College Composition and Communication*, *35*(4), 400–409. Retrieved from http://www.jstor.org/stable/357792

Wyatt-Smith, C., Klenowski, V. & Gunn, S. (2010). The centrality of teachers´ judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*(2), 37–53. https://doi.org/10.1016/j.asw.2015.11.001 59-75.

APPENDIX

Appendix 1

Task 1: Expository task

Write a text where you explore reasons to why smoking was more accepted in society before than it is now. You are writing the piece to a booklet about drugs.

Task 2: Narrative

Imagine that you are on your way home one late autumn night and see a strange light. Write a text where you tell about your experiences this night. Imagine that you are going to read the text aloud to your classmates, maybe while you are in a dark room with candles.

Appendix 2

Rating scale	Mastery level 1	Mastery level 2	Mastery level 3	Mastery level 4	Mastery level 5
Communication	- The relationship the text establishes between writer and reader is unclear	- The student attempts to establish a relevant relationship between writer and reader.	- The text displays a relevant relationship between writer and reader	- The text displays a relevant relationship between the writer and reader	- The text displays a relevant relationship between the writer and reader
	- The student implies a position as a writer	- The text can sometimes attempt to address the reader's need to know the participants, concepts and circumstances	- To some extent the text addresses the reader's need to know the participants, concepts and circumstances	- To a large extent the text addresses the reader's need to know the participants, concepts and circumstances	- The text addresses the reader's need to know the participants, concepts and circumstances
		- The student implies an appropriate position as a writer	- To a large extent the student displays an appropriate position as a writer	- The student displays an appropriate position as a writer	- The student displays an appropriate position as a writer

			- To some extent the text displays appropriate style level	- To a large extent the text displays appropriate style level	- The text consistently displays appropriate style level
Contents	- The content of the text is to a small extent relevant to the assignment	- The content of the text is to some extent relevant to the assignment	- The content of the text is relevant to the assignment and the text can display subject matter knowledge	- The content of the text is relevant to the assignment and the text displays subject matter knowledge	- The content of the text is relevant to the assignment and the text displays subject matter knowledge and originality
	- Topics are not elaborated	- A few topics are elaborated	- Some topics are elaborated	- The topics are to a large extent elaborated	- The topics are sufficiently elaborated
Text Structure	- The text has to some extent a structure and may have some framing (e.g. to-from)	- The text has to some extent a structure and have to some extent an introduction and an ending	- The text is structured in a purposeful way and often displays introduction and ending	- The text is structured in a purposeful way and includes introduction and ending	- The text is structured in a purposeful way and includes, for example, extended introduction or extended ending and use of topic sentences.
	- The text is to some extent thematically coherent	- The text is to some extent thematically coherent, but may have a repetitive structure	- The text is thematically coherent; paragraphs may not always be marked graphically	- The text is thematically coherent; paragraphs are almost always marked graphically	- The text is thematically coherent; paragraphs are marked graphically

	- The text demonstrates textual cohesion by simple connectors, like and, so, when, because, if, but	- The text demonstrates textual cohesion by connectors, like and, so, when, because, if, but, or	- The text functionally employs connectors, like and, so, when, because, if, but, or, since	- The text functionally employs connectors, like and, so, when, because, if, but, or, since, besides, for example	- The text functionally employs connectors, like and, so, when, because, if, but, or, since, besides, for example, when it comes to, contrary, that is
Language Use	- The text has sentences that does not convey meaning	- Most sentences in the text convey meaning	- The sentences in the text convey meaning	- The sentences in the text convey meaning	- The sentences in the text convey meaning
	- The text has sentences with a simple syntax	- The text has sentences with a mainly simple syntax	- The text may have sentences that display complex syntax	- The sentences in the text show variety in the grounding field	- The sentences in the text show variety in the grounding field
	- Most of the sentences in the text begin in the same way.	- The sentences in the text show some variety in the grounding field	- The sentences in the text show variety in the grounding field	- The text shows variety in the choice of words. Wording and concepts are precise.	- The text shows variety in the choice of words. Wording and concepts are precise and used in a correct manner.
	-The text shows little variety in the choice of words. The text is characterized by colloquial language	- The text shows some variety in the choice of words. The text may be characterised by colloquial language	- The text shows variety in the choice of words. Wording and concepts may be precise.	- The text may show use of linguistic devises such as highlighting, exaggeration and similes	- The text may show use of linguistic devises such as highlighting, exaggeration, humor, metaphors, similes, and contrasts
		- The text may show use of linguistic devises such as highlighting and exclamation	- The text may show use of linguistic devises such as highlighting and exclamation		

Spelling	- 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.	- The text shows correct spelling of long phonetic words and high frequent non phonetic words. The text may contain some dialect words.	- The text shows correct spelling of long phonetic words and non phonetic words. The text may contain some dialect words.	- The text shows correct spelling of long phonetic words and advanced non phonetic words. The text may contain some spelling errors.
	- The text mostly has capital letters in proper names and at the beginning of new sentences.	- The text mostly has capital letters in proper names and at the beginning of new sentences.	- The text has capital letters in proper names and at the beginning of new sentences.	- The text has capital letters in proper names and at the beginning of new sentences.	- The text has capital letters in proper names and at the beginning of new sentences.
		- The text shows correct concord most of the time.	- The text shows correct concord.	- The text shows correct concord.	- The text shows correct concord.
Punctuation	- The text displays some correct use of full stops, exclamation marks and question marks	- The text often displays correct use of full stops, exclamation marks and question marks	- The text displays correct use of full stops, exclamation marks and question marks	- The text displays correct use of full stops, exclamation marks and question marks	- The text displays correct use of full stops, exclamation marks and question marks
	- The text sometimes uses commas in lists	- The text often uses commas in lists and before “but”	- The text uses commas in lists and before “but”. The text may contain commas between complete sentences that are connected with connecting words.	- The text uses commas in lists and before “but”. The text often contain commas between complete sentences that are connected with connecting words.	- The text uses commas in lists and before “but”. The text almost always contain commas between complete sentences that are connected with connecting words. The text may also display other correct uses of comma.

The text may display correct use of hyphen, parentheses, colon, and quotation mark	The text may display correct use of hyphen, parentheses, colon, quotation mark, dash	The text may display correct use of hyphen, parentheses, colon, quotation mark, dash

Appendix 3

Bio

Name:
Age:
Current workplace and type of school:
Work experience (number of years and years at specific year level):
Language variant: bokmål or nynorsk (both are standard forms of the Norwegian)

Rater motivation and perception of task

Motivation to participate:
Experience in NPR:
Training:

- What has the content been during training?
- What has been stressed, do you think?
- Has the training effected the way you rate? How?

Perception of text quality

Do you think the NSBWT-texts usually are better or worse than the texts you see students produce in the classroom?
Based on your experience with the NSBWT: What do you think students, in general, should work more with to improve their texts?
What is the single feature that makes a text good?

Materials, procedures and facilitation

Do you use the descriptors?
Do you use the annotated benchmark texts?
Which tool do you find the most important?
Which rating scale do you hold to be the most difficult to rate - and why?
Which rating scale do you hold to be the easiest to rate - and why?

Describe how you rate a text. Do you rate rating scale by rating scale in a specific order, do you variate the order?
How do you go about with the pile of texts - do you read and rate consecutive, do you read every text first and then start to rate, or other ways?
When do you rate the texts?
How long time do you use to rate the texts?
Do your read on screen or do you print the texts?
Do you cooperate with anyone?

Live rating

First we would like you to very quickly reread a student text you rated at last NSBWT (fall 2015). Could you now try to activate your NPR rating procedure and re-rate this text? We would like you to think aloud while rating.