**Title**

“Digging for Gold” or “Sticking to the Criteria”: Teachers’ Rationales When Serving as Professional Raters

**Authors**

Lennart Jølle

Gustaf B. Skar

**Author Contact Details**

Lennart Jølle, Department of Teacher Education, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, lennart.jolle@ntnu.no

Gustaf B. Skar, Department of Teacher Education, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, gustaf.b.skar@ntnu.no, <https://orcid.org/0000-0002-6486-396X>

**Corresponding Author**

Gustaf B. Skar

**Abstract:**

This paper reports findings from a project called “The National Panel of Raters” (NPR) that took place within a writing test program in Norway (2010-2016). A recent research project found individual differences between the raters in the NPR. This paper reports results from an explorative follow up-study where 63 NPR members were surveyed with 23 items that were dilemma-like in the sense that deviating from the NPR rules would follow another – but socially acceptable – rationale. Four NPR members participated in a follow-up interview in which they motivated why they had agreed or disagreed with certain items. The results indicate two distinctly different stances toward rating work, with one stance threatening the validity of the scoring process.

**Key Words:**

Writing assessment

Reliability

Validity

Rater training

# “Digging for Gold” or “Sticking to the Criteria”: Teachers’ Rationales When Serving as Professional Raters

## Introduction

This paper presents results from an exploratory investigation into the approaches to assessment of teachers when serving as raters on a professional writing assessment panel. In Scandinavia, as in many places throughout the world, language teachers are hired to serve as professional raters for writing assessment programmes. In Norway, for example, teachers are hired as ‘censors’ to rate Norwegian written. In addition, teachers in Scandinavian school systems actively contribute to school leaving certificates through classroom-based assessments.

Previous research has indicated substantial difficulties in establishing so-called ‘interpretive communities’ (Berge, 2002; White, 1984), which are communities in which raters

‘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’ (Skar & Jølle, 2017, p. 3). Small and large-scale studies in Norway and Sweden have indicated a generally low reliability associated with writing assessments (Berge, 2005; Skar, et al, 2017; Östlund-Stjärnegårdh, 2009). Indeed, a poor reliability of writing assessments has been common internationally as well (e.g. Weigle, 1998). In fact, this has been so common that McNamara (1996) and Eckes (2015) suggested the routine use of statistical techniques to compensate for ‘rater errors’.

A recent study (Skar & Jølle, 2017) on a long-standing writing assessment programme has provided a new understanding of this issue. The results suggested that a lack of reliability might be related to aspects beyond ‘rater effect’ (see below). In this specific setting, the findings indicated that the teachers who were hired as raters approached the rating tasks with varying motifs and rationales. While they were aware of which procedures to follow, some raters deviated from them when they deemed appropriate. This paper presents the results of an explorative follow-up study into teachers’ approaches to assessment work within the context of a professional rating panel. The purpose was to explore potential additional factors that may contribute to rater variation of writing assessments.

## Rater Effect, Assessment Literacy, Assessment Identity and Conceptions

Broadly speaking, rater reliability has been investigated either as a dimension of writing assessment programmes—investigated as rater effects—or as a dimension of classroom-based writing assessments, which is considered an aspect of assessment literacy. In the first case, participants may or may not be teachers. When they are, this does not tend to be the focus of the investigation, though there are exceptions (Jølle, 2014; Parr & Brown, 2015, p. 140). In the second case, reliability is considered and investigated as a part of a teacher’s overall assessment work (e.g. Weigle, 2007). The following sections outline previous findings related to each of these dimensions.

From a psychometric point of view, a rater effect can be defined as ‘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). Such effects threaten the validity of the interpretation and the use of scores because they introduce “construct-irrelevant variance” (Messick, 1989). Throughout the history of writing assessments, rater effects that cause low rater reliability have remained an issue (Eckes, 2015; e.g. Edgeworth, 1890; McNamara, 1996). In some educational contexts, such as the American context, it has been a major concern (Huot & Neal, 2006), sometimes causing the abandonment of direct writing assessments in favour of objectively scored indirect measures (Crusan, 2014).

In a comprehensive review, Myford and Wolfe (2003) listed the effects associated with raters, which include leniency/severity error, halo effects (i.e. failure to discriminate between assessment dimensions), a central tendency effect (i.e. using the mid-part of the scale) and a restriction of range effect (i.e. only using some part of the scale). Eckes (2012) added to the list after investigating rater cognition and found that raters’ conceptions of important elements (e.g. spelling) affect ratings. For example, “criteria perceived as highly important were more closely associated with severe ratings’ and vice versa” (Eckes, 2012, p. 270). Matre and Solheim (2016) found a connection between scale use and meta-language; as raters developed meta-language, they also developed a flexible, pragmatic approach to the rating scales. Within writing assessment programmes, these effects are more or less routinely investigated (e.g. Skar, 2017). By and large, from a psychometric perspective, the goal is to reduce the impact of individual raters on assessments; however, it seems as if rater training cannot necessarily be used to remedy issues related to rater effects (e.g. Baird, Greatorex, & Bell, 2004; Knoch, 2011; Tengberg et al., 2017; Weigle, 1998), although there are exceptions, demonstrating the possibility to achieve reasonable reliability (Brown, Glasswell, & Harland, 2004).

Given these findings from contexts in which ratings are the chief concern, it is not surprising that reviews of classroom-based assessments, or CBAs (e.g. Brookhart et al., 2016; Harlen, 2004), have revealed multiple threats to the reliability and validity of teacher assessments. In response to such findings, and because teachers generally receive poor training related to assessments and testing, researchers have called for enhancing teachers’ “assessment literacy” (Stiggins, 1991) or “educational assessment knowledge and skills” (Brookhart, 2011, p. 7) for decades. Recently, Looney, Cumming, van Der Kleij, and Harris (2017) suggested that there is a need to expand the notion of assessment literacy to that of assessment identity, which also includes teachers’ beliefs, feelings, confidence and perceptions of their own roles as raters. Others have suggested the need for teachers to develop a rater identity to successfully fulfil the assessment-related aspects of teaching and instruction, such as conducting systematical formative assessments (Xu & Brown, 2016). Indeed, Xu and Brown (2016, p. 158) referred to forming a rater identity as the “ultimate goal” of teacher assessment literacy in practice. Assuming a rater identity refers to the incorporation of assessment into teaching activities: “if a teacher [...] regards him- or herself as a teacher, teaching practice may cease when teaching content is delivered. [...] If the teacher is conscious of also being an assessor, he/she ought to [...] implement assessment activities to measure whether what was taught has been learnt [...]” (Xu & Brown, 2016, p. 158).

For many teachers, there may be tension between instructing and assessing. As Rea-Dickins (2004, p. 253) noted in a small research review, teachers face “significant dilemmas in their assessment practices: sometimes torn between their role as facilitator and monitor of language development and that of assessor and judge of language performance as achievement”. Brown and colleagues (e.g. Brown, 2004; Brown, Lake, & Matters, 2011) have investigated teachers’ beliefs regarding common conceptions of assessments that can be related to this tension, such as assessments as a means for teaching and learning or as a means for accountability (Brown et al., 2011, p. 211). Results suggest that teachers seem to generally agree that assessments may improve teaching and learning, and some teachers also contend that assessments can be used as a tool to hold students accountable for their learning (Brown, 2004; Brown et al., 2011). It is plausible that there is a potential connection between conceptions of assessment purposes and assessment situations that teachers perceive as problematic: if one enters a high-stakes assessment situation with the belief that assessments should mainly support student learning, this conception may be at odds with the rules for high-stakes assessments. For example, scoring rules for high-stakes assessments regularly require raters to be blind to student development and to focus on which rubric best describes the object of assessment (i.e. the student text), rather than collecting evidence to determine which step to take next during instruction.

While teachers often participate in writing assessment programmes and findings from writing assessment programmes are often corroborated by CBA studies, to the authors’ knowledge, few studies have investigated professional rating panels with a focus on the members’ backgrounds as teachers. Based on the findings in the previous study of the long-standing writing assessment programme (see Introduction), the aim was to contribute to establishing a link between research that is oriented toward psychometrics and research that is oriented toward CBA. This research approach can extend the available knowledge on rater effects.

To achieve this aim, teachers who were members of a professional rating panel were surveyed using a questionnaire that contained items related to rating. The items described fictional but realistic assessment situations and were constructed to be similar to dilemmas in the sense that deviating from the panel rules would follow another rationale that presumably would be in line with other established conceptions of assessments (c.f. Brown, 2004). After obtaining the results from the questionnaires, interviews with participants with different response patterns were conducted.

This article answers the following research questions:

* Did the dilemmas discriminate between the participants and – if so – what was characteristic of dilemmas that discriminated between high and low scorers?
* How did members of the professional rating panel with different response patterns explain their choices of item options?

## Context of the Study

In 2010, the Norwegian Writing Centre (NWC) was commissioned by the Norwegian Directorate for Education and Training (NDET) to develop the National Sample-Based Writing Test (NSBWT) and the Formative Writing Assessment Package (FWAP). The NSBWT was later administered annually to a nationally representative sample of students in 5th and 8th grades. The purpose of the test was to provide a national ‘report card’ containing information about students’ writing proficiency (Skar, 2017). The results of the NSBWT were also incorporated into the FWAP, which consisted of the previous year’s NSBWT tasks, scoring scales and annotated exemplar texts. The FWAP was freely available to individual teachers to download from the internet, and teachers could compare the results in their classes with the national level.

In 2010, the NWC was also commissioned to establish a national panel of raters (subsequently the NPR). The goal was to establish an interpretive community that could reliably rate the NSBWT. Between 2010 and 2016, the NPR existed as a semi-permanent panel consisting of 60–80 teachers recruited from schools across Norway. The vast majority of teachers who were recruited to the panel were L1 teachers (Norwegian). The members of the NPR underwent extensive training, including lectures on how to use the assessment materials, and more broadly, lectures on writing assessments, formative assessments and writing instruction; however, the main training activity consisted of the assessment of texts in rater pairs (Jølle, 2014). All teachers would change partners several times to ensure maximum exposure to similar and different views of writing and writing proficiency and ways to deal with assessments. The rating sessions were often followed by consensual coordination meetings. It was assumed that these procedures would contribute to the development of an interpretive community. While individual differences remained, as a group, the teachers were able to produce reliable ratings (Skar & Jølle, 2017).

Following the agreement with the NDET, the NSBWT was to be scored analytically based on five scales (Writer-reader-interaction, Content, Text structure, Language use and Morphology, grammar and spelling). Each scale has five levels (“Mastery levels”) that describe the requirements of each specific level (included in Appendix A). Each level is denoted “Mx”, so mastery level 1 is “M1”, mastery level 2 is “M2” and so forth until mastery level 5, which is “M5”. The scales were part of a set of ‘assessment resources’, which also included annotated exemplar texts. Moreover, all scripts were blinded, i.e. the raters did not know the genders, ages, mother tongues or schools of the students. The raters were instructed to read student texts and to compare them with rubric levels and exemplar texts to identify the best match. The raters were also explicitly told not to infer anything about the student but rather to focus on how to best describe the text at hand. An additional important feature, which was communicated to all raters, was that no student could receive raw ratings from individual raters.

## Participants and data collection

In the fall of 2016, 63 NPR members met for two days for a NSBWT rating session and workshop. All participants had been members of the NPR for at least three years and had received proper training (Skar & Jølle, 2017).

 Data collection mainly took place during the workshop and followed the principles of the explanatory mixed methods design in which the results from the first investigation informed the second investigation (Creswell & Plano Clark, 2011). First, the NPR members were surveyed using 23 dilemma-like items in a questionnaire (included in Appendix B). The participants were asked to “agree” or “disagree” with certain rater behaviours described in the items of the questionnaire. Depending on the option selected, the raters received a score of 0 when the option was in line with NPR rules (the “correct” alternative following the blueprint) or a score of 1 when the option indicated deviance from NPR rules (the “wrong” alternative deviating from the blueprint). For some items, ‘agree’ was in line with NPR rules, and for some, ‘disagree’ was in line with the rules. Thus, raters who received low scores on the questionnaire opted for alternatives that indicated an adherence to the NPR rules to a greater extent than raters who received high scores.

For example, item #2 asked NPR members to agree or disagree with the following scenario:

David notices that a student text suffers from poor text structure. On the specific scale for assessing structure, David is convinced that this text merits the lowest mark; however, it is obvious that the poor text structure is a result of ‘learning-in-progress’. David decides to acknowledge this and gives the text a higher mark.

For this item, choosing “disagree” would render a score of 0 because the NPR guideline states that raters should rate text quality only and not whether learning is in progress. Item #9 asked the members to agree or disagree with this scenario:

Lisbeth scores a text that clearly communicates the topic to the reader; however, due to what she believes is an inaccurate assignment, the Writer-reader-interaction is not entirely relevant. She scores the text “M2” (mastery level 2) even though she believes the text merits “M4” on Writer-reader-interaction.

Here, choosing “agree” would yield a score of 0 because Lisbeth must rate the text according to the guideline, and the guideline states that the Writer-reader-interaction should be relevant to the task.

Second, after the questionnaires were analysed, the four teachers with the highest score (i.e. the teachers whose answers indicated the largest deviance from NPR rules) and the four teachers with the lowest scores were invited to participate in informal follow-up interviews. The groups were presented six items that had discriminated between the two groups. The purpose of these informal interviews was twofold. The first aim was to investigate whether the participants had interpreted the content of the items in a similar fashion, which they had. The second purpose was to recruit participants to a formal interview, and two teachers from each group agreed to participate. These interviews were conducted outside of the workshop and were based on the six items, and participants were asked to explain why they agreed or disagreed. The two high-scoring participants were Hannah (14 points) and Helen (12 points), and the low-scoring participants were Laura (3 points) and Luke (4 points).

The items were created using three sources of input. First, findings were used from the previous NPR study that indicated that some raters were willing to ignore the NPR rules and assessment material, which was related to their self-perception as experts of rating before becoming a member of the NPR (Skar & Jølle, 2017). One such example was to disregard assessment material when the raters found them unreasonable, which is illustrated by item #3. Second, informal reports from raters were used. These reports were related to the specific context of the NPR. An example is item #5, which discusses rater pairs. Third, to some extent, the different concepts of assessments mentioned previously were considered (e.g. Brown, 2004). The NSBWT could be considered to serve an accountability purpose of assessments because the results would function as a norm for measuring writing proficiency in local schools, and when made public, the results would be used by policy makers and others to discuss writing instruction in Norway in general. In addition to its implications for the use of scoring rules, this purpose could be viewed as being at odds with another purpose, which is to improve teaching and learning. Items were created that would require raters to choose between using the scoring rules or disregarding them. When opting for the latter, the described actions in the item would implicitly or explicitly be related to classroom actions, such as motivating students or giving them the benefit of a doubt. These items contain specific references to students, such as item #2.

Two caveats related to the design of the survey should be mentioned. First, given that the survey targeted a population with special characteristics, it was not possible to pilot the items. Second, the survey includes a few items that are quite similar because it was assumed that some teachers would choose alternatives contrary to their beliefs but adherent to the NPR rules but that these tendencies would become clear if a sufficient number of items were included.

## Data analysis

To answer research question 1, item analysis techniques from the classical test theory (CTT) were applied. For each item, the facility value, *p*i (i.e. proportion of test takers scoring 1 on item *i*) and the discrimination index (D) were calculated. The facility value is also referred to as difficulty index and the interpretation is straightforward: the larger facility value, the larger the proportion of participants who scored 1 point. In CTT the discrimination index is calculated to find items that discriminates between test takers. Items with a high discrimination index indicate a positive relationship between the item score and total score. Items with low or negative discrimination indicate weak or negative relationship to the total score. For example, an item with negative discrimination would indicate a situation where overall low scorers receive a score, while overall high scorers does not.

The discrimination index was computed by using the formula: D = *p*HIGH - *p*LOW (van Blerkom, 2018, p. 123), where *p*HIGH is the facility value for 27% of participants with the highest total scores, and *p*LOW is the facility value for 27% of participants with the lowest total scores. The cut-off of 27 % maximizes number of participants in each group, while preserving the difference between them (Matlock-Hetzel, 1997). To adhere to the conventional rule of thumb (e.g. Bachman, 2004), items with D equal to or larger than .30 were retained in the subsequent analysis. It is important to note that unlike a conventional item analysis, these computations were not carried out to evaluate the quality of the items, per se, but to identify items that could serve the exploratory purposes of this study.

To answer research question 2, the items that were part of the interviews were sorted into the three categories. Thereafter, the authors jointly analysed how the participants positioned themselves in relation to the items in each category. For example, when one of the interviewees, Laura, explained her choice to agree with the fictitious rater in item #9 by stating that “[the fictitious rater] needs to stick to the writing prompt”, this was interpreted as a conscious and strong position vis a vis the rules of the panel. The positionings of the participants were used as a basis for a synthesis that described why the participants were motivated to make a choice overall.

## Results

### RQ1: Did the dilemmas discriminate between the participants and – if so – what was characteristic of dilemmas that discriminated between high and low scorers?

The dilemmas did discriminate between the participants. Overall, the average scale score was 7.2 (SD = 2.79), indicating that the common tendency among the raters was to opt for the alternative in which the fictitious does not act in accordance with NPR rules and guidelines (henceforth: the “wrong” alternative) for several items. The low scoring group had an average of 4.0 (SD = 0.91), and the high scoring group had an average of 10.68 (SD = 1.38). Item #9, in which the described rater correctly awarded a low score to a ‘good’ text, had the highest facility value of *p* = .79 (D = .12). Item #21, which described a rater who correctly sometimes reviewed her ratings, had the lowest facility value of *p* = .05 (D = -.11).

There were 10 items that indicated a substantial discrimination between high and low scoring participants (i.e. D ≥ .30): #1, #2, #3, #5, #8, #11, #12, #17, #19 and #23 (see Table 1). These items represent all three categories (NPR members as independent experts, specific context of the NPR, conflicting purposes with assessments), which indicates profound differences between the high and low scoring groups related to all aspects of the NPR responsibilities captured by the survey.

Using a finer lens, items #1, #3, #7, #11, #17 and #23 were all related to the first category (NPR members as independent experts). For all these items, the “wrong” alternative describes a deliberate deviation from explicit criteria or assessment resources presented to the panel. In the case of item #1 (*p* = .37; D = .58), the rater in the example chooses to include handwriting, which is not part of the criteria, in his judgement. In items #3 (*p* = .37; D = .53), #7 (*p* = .40; D = .35) and #17 (*p* = 0.27; D = .53), the described raters ignore the assessment material because they find it unreasonable. Item #23 (*p* = .32; D = .47) describes the incorrect use of the scale, and in #11 (*p* = .43; D = .36), the rater utilises holistic scoring before scoring the scripts analytically. All these items are descriptions of rater actions that place the rater above the guidelines by including new criteria or simply ignoring the established criteria. The results showed that the high scoring group preferred such alternatives relatively often.

Items #5 and #19 are related to the specific context of the NPR in that they described rater pair discussions. In item #5 (*p* = .27; D = .47), the fictitious raters inappropriately abandon the discussion of actual texts and instead discuss writing instruction. In item #19 (*p* = .57; D = .34), the described rater incorrectly dismisses the concept of rater pairs as a way to reach consensus. Again, the low and high scorers were separated by agreeing or disagreeing with actions that position the fictitious raters as individual members rather than as part of a collective group.

The final two items, #2 (*p* =.62; D = .56) and #12 (*p* = .44; D = .68), are both related to the third category (different concepts about assessments). Among the ten items separating high and low scorers, these had the largest facility values and the largest and third largest D-value, respectively. For these two items, most raters responded “incorrectly” except the group of low scorers. Both items described situations in which the fictitious raters’ actions were based on concerns for the student rather than applying the criteria that best described the text.

[TABLE 1 HERE]

### RQ2: How did NPR members with different response patterns explain their choices of item options?

Table 2 describes participants’ scores for the six items as well as the total scale. As can be observed, the high scorers, Hannah and Helen, scored 14 and 12 points, respectively (of 23). Thus, Hannah’s scale score was clearly above the average of the NPR. The low-scorers, Laura and Luke, scored 3 and 4 points, respectively, which was clearly below the NPR average. There were three items with perfect discrimination between the pairs (#2, #12 and #23).

[TABLE 2 HERE]

Items #7, #9 and #23 are related to the first category (NPR members as independent experts). In item #7, the fictitious rater wrongfully disregards the fact that the student has failed to use the correct genre. Hannah stated: “I would have done the same. That’s the pro with this analytical assessment. You can reward qualities where there are qualities”. Helen was a bit more uncertain but she stated: “Well, perhaps I would have given the text a lower score on one of the scales, but I think I would have assessed it independently from genre expectations”. Laura, who actually chose the “incorrect” alternative, did not agree with Hannah or Helen. Rather, she stated that: “No, he has to adhere to genre expectations. That has to be instructive”. In relation to de facto opting for the opposite alternative, she said: “Well, I must have thought it all wrong inside my head”. Luke’s reasoning was similar to that of Laura: “I disagreed because I stress the formal requirements”. In terms of positioning, all but Hannah’s reply indicated conscious positions vis-à-vis the rules of the panel. Hannah seemed to have had a different conception of the possibilities of analytical scoring.

In item #9, the fictitious rater rightfully awarded “M2”, although she preferred to award “M4”. This was the most difficult item to answer “correctly” (c.f. Table 1), and it divided the four raters into two groups, albeit the low scorer, Luke, on the survey had agreed with the high scorers. In the follow-up interviews, it became clear that Luke had changed his mind: “Well, the task is what it is, and you cannot take that into account afterwards”. As such, he agreed with Laura, who simply stated that “[the fictitious rater] has to stick to the task”. Hannah believed the text “should receive praise because it communicated well”. Helen, who also “wrongfully” disagreed with the fictitious rater, responded similarly: “I am concerned about what should instruct the scoring, but in practice, I think many end up, including myself, with ‘M4’”.

For the final item related to the first category, item #23, Hannah was the only one who opted for the “wrong” alternative. In this item, the fictitious rater decides to adjust the scores on the different scales for a specific text based on a comparison with other texts, even though the criteria suggests otherwise. Hannah explained her choice in the interview by emphasising that this should not be done for every scale and that the rater “has to find out where [i.e. which scale(s)] he wants to award the aspects he finds to be better”. Helen “rightfully” disagreed with the rater, but her argument for this was related more to thoroughness and effort than professional concerns: “I would never go back, compare and change my scoring because I had done it so thoroughly”. In contrast, Luke’s concern was different, and he seemed to think that a variation of text quality within each band level was acceptable: “An ‘M3’ is—each level has a width from a weak ‘M3’ to a strong ‘M3’—it might be a big difference”. Laura seemed to recognise the dilemma and ignored the fictitious rater’s decision to compare texts: “It is not possible to speculate like this. You have to assess based on what is written [i.e. descriptors] on every scale level”.

For these first three items, the low-scorers Luke and Laura appeared to have focused more on the formal rules and guidelines of the NPR and to put aside personal meanings of right and wrong than Hannah and Helen, who conceptualised the assessment situations as instances of formative assessments. The latter two also seemed to focus on students’ motivation and what they perceived as benefits of neglecting criteria in search for qualities to praise.

Item #19 belongs to the second category (scoring within the specific context of the NPR). The fictitious rater wrongfully believes that consensus among raters is too ambitious an objective, and all but Laura chose the “wrong” alternative; however, when interviewed, she had changed her mind and stated that she agreed that the most important aspect of the work of the raters is to develop enough competence to argue one’s judgements. Luke also agreed and did indeed state that the ability to defend his judgements was “one of the most important skills” he had developed during his time serving on the panel. Helen stated that she agreed with the rater in the item because she believed the single most important aspect for her was to be thorough. Hannah, who also opted for the “wrong” alternative, offered a somewhat different explanation and stated that “there are so many feelings involved” that it is impossible to “decisively determine if [the writing] is right or wrong”.

This item did not sufficiently discriminate between the pairs of high and low scorers, but there was a subtle difference between Luke and Laura on the one hand and Hannah and Helen on the other. While the first two discussed the defending of judgements, i.e. preparing for a legitimate negotiation of marks, Hannah seemed to find consensus irrelevant (because a student script cannot be judged as right or wrong anyway), and Helen seemed to focus on being thorough.

Items #2 and #12 belong to the third category and discriminated perfectly between high and low scorers. In both cases, the fictitious raters wrongfully draw inferences about the writer rather than focusing on the text at hand. Laura and Luke, who answered both items “correctly”, were sceptical of making such inferences. Regarding item #2, where the rater awards a higher mark than adequate because there was learning in progress, Luke said that “I think I disagree because [...] if it is a poorly structured text this time, then he should not get a high mark on structure. Maybe next time he will have learnt [to structure text]”. Because the mark is awarded based on inferences of the writer’s learning process, Laura said that she remained sceptical of the rater’s choice: “I think one moves in a questionable direction if one begins to speculate about that”. Helen, on the other hand, said that “if it is obvious, I would have opted for ‘M2’ [the higher mark]”. Hannah positioned herself differently from the low scorers and stated that “again, it is about searching for qualities, especially with weak students”.

Regarding item #12 in which the fictitious rater wrongfully chooses to give the text a higher mark than it deserves, Hannah continued to discuss rewarding aspects that can be rewarded: “it is obviously about rewarding stuff, but also about punctuation, [...] the message becomes clearer when the punctuation is good”. Helen also opted for the “wrong” alternative but reasoned more extensively that “if you award ‘M1’ [the lowest mark], then you signal [to the student] that [she or he] has not mastered anything. If you give [the text] ‘M2’, it will be like saying ‘there is something there’. On the other hand, it is a bit awkward because of the orthography problems. [...] Well, it is about digging for gold, [and] I have developed from in a way [...] counting errors to [...] [if] the punctuation is better than expected, well, that counts more now”. In other words, Helen expressed that her current assessment practices involved searching for qualities and allowing them to weigh more than deficits. Luke and Laura did not agree with Hannah or Helen. Indeed, Luke stated that: “I really disagree. [...] The formal requirements are exactly the same, so I do not think this is a proper way to compensate for language deficits. [...] I like to say: stick to the formal requirements”. Laura did not offer a lengthy explanation but only stated: “I think she should stick to the criteria”.

As with the first three items, it seems as though there was a difference in the way the high and low scorers positioned themselves vis-a-vis the NPR rules and also vis-a-vis the NPR objectives. Luke and Laura consistently reasoned in ways that expressed a stance in which the scoring rules of the NPR were important to follow, even if the individual raters disagreed, and that the objective of the NPR was to provide (reliable) ratings to the NSBWT rather than ratings that were to be consumed by the student writer. Helen and Hannah seemed to adopt an opposing stance and used words such as “signal” when discussing the effects of rating. In other words, Hannah and Helen did not seem to view their responsibility as primarily oriented toward the objectives of the NPR and the NSBWT programme but also oriented toward students participating in the NSBWT.

## Discussion

Undoubtedly, the main finding of the present study is that the survey captured differences in response patterns among the participating raters. The use of items in which fictitious raters deliberately choose to disregard NPR guidelines and rules could have resulted in most teachers answering in the same way; after all, there is no reason to suspect that NPR members are not aware of the NPR guidelines and rules. Rater variation is important considering the main objective of rater training for the NPR, namely to establish an interpretive community. While previous research has demonstrated that such a community is threatened because raters differ in severity (Skar & Jølle, 2017), these results indicate an even more profound threat because NPR members may apply different rationales when rating texts.

The results furthermore showed a discrimination between high and low scoring participants for all three categories (NPR members as independent experts, specific context of the NPR, conflicting purposes with assessments). This indicates that the differences between the groups cannot be reduced to a matter of more or less adherence to local procedures. Rather, the differences can plausibly be explained by differences in rater perspectives. The results indicate that the high and low scoring groups have different views regarding the purpose and the consequences of the NPR ratings. While the latter group tended to choose alternatives that were in line with the official guidelines, the former chose alternatives that either supported the belief that the individual (expert) raters can choose whether to follow the guidelines or not or student-centered alternatives in which the described actions could encourage or motivate students or could give them the benefit of the doubt.

For research question 2, the interviews indicated that the raters could explain their choices of item options in different ways. While the low scorers Laura and Luke consistently related their choices to the rules of the NPR, Hannah and Helen related their choices to what they perceived as the best outcome for the individual student. In some cases, the discrimination between Luke and Laura on the one hand and Helen and Hannah on the other was not perfect, either because Luke or Laura opted for the “wrong” alternative or because Helen opted for the “correct” alternative (cf. Table 2); however, when Luke and Laura were “wrong” in the survey, they used the interviews to “correct” their answers (e.g. Laura for #7 and Luke for #9 but not #19). In contrast, when Helen opted for the “correct” alternative on item #23, her reasoning in the interview did not seem to follow any NPR rule. Instead, her view of herself as an expert rater just happened to coincide with the rationale inherent in the NSBWT construct. As such, the interviews only strengthened the results from the first analysis when items that discriminated between high and low scorers were identified.

The results of the present study cannot easily be explained by the results of previous investigations into rater variation. For example, when faced with dilemma #9, Hannah “wrongfully” disagreed with the fictitious rater, and she reflected on her role as a rater in this way: “There is a difference in a way in assessing texts as part of the NPR contrary to being a teacher (...). It might be that you can be harsher sitting on the panel, but I believe we should search for qualities where there are qualities”. As such, Hannah expressed a rationale for assessments that is not in line with the NPR but that is in line with her role as a teacher. This is not related to raters’ conceptions of important elements (Eckes, 2012), the use of meta language (Matre & Solheim, 2016) or rater effects, such as the central tendency and halo effects (Myford & Wolfe, 2003). The results may not even be explained by the notion of assessor identity (Xu & Brown, 2016). The NPR members who participated in the interviews by no means indicated that they did not include assessments in their teaching activities. On the contrary, and as Hannah expressed, assessments can be viewed as a part of regular teaching activities.

To understand the results, it is proposed that the answers from the interview can be used to tentatively place the NPR members along a continuum (Figure 1). At one end of the continuum is the position as a “professional rater”, i.e. a NPR member who strictly adheres to the guidelines and rules of the rating panel. Luke and Laura positioned themselves toward this end by stressing the rules and guidelines, such as when Laura refused to speculate about the students’ learning and when Luke did not want to “compensate for language deficits”. Neither Luke nor Laura expressed concerns for the students, the students’ motivation or learning trajectories. A position close to the professional rater-end of the continuum would explain this. The NPR training activities and guidelines and rules stressed that the most important task of the panel was to reliably rate student texts by matching text qualities to assessment criteria. As such, Luke and Laura reasoned in a way that was in line with the objectives of the NPR.

At the other end of the continuum is the position ‘independent teacher expert’, i.e. a NPR member who consistently believes it is the individual teacher's prerogative to choose which guidelines and rules are relevant (hence ‘independent’) and who infers a great deal about the student from the text. These inferences are used during the assessment, such as when choosing between different mastery levels. Hannah and Helen positioned themselves toward this end of the continuum, such as when they stated that they would have no problems disregarding the fact that a student had chosen the wrong genre. This indicates that Hannah and Helen agree that there may be personal, hidden or informal criteria that the NPR member can use when s/he feels it is appropriate, which reflects the independence at this end of the continuum. Another example is when Hannah claimed to be “digging for gold”, which could be related both to independence (prerogative to neglect deficits in the student text) and to a teacher position in which student motivation is important. Thus, the continuum may help us to understand that the NPR members may all have rater identities, but differentiate in stances toward assessment work, which may affect their judgements of student text quality.

[FIGURE 1 HERE]

The differences between NPR members revealed by this investigation may have implications for utilising teachers on professional rating panels. Some teachers may find it difficult to shift from a teacher position to a rater position, which conventional rater training may not be able to resolve.

## Conclusion

While the most recent NPR study (Skar & Jølle, 2017) showed that the NPR raters displayed common rater effects (such as different severity levels), the results of this study indicated that it may be necessary to investigate rater actions in terms of the continuum presented. As shown by the dilemma-survey, the raters opted for alternatives related to different stances. If the results of this study are generalisable, the facilitators of panels such as the NPR should provide adequate training to both raters who view themselves as part of a collective group that rates texts reliably and raters who view themselves as independent teacher experts with the prerogative to choose actions that they believe will benefit the students most.

There are some limitations to this study that should be addressed in future studies of the same phenomenon. First, it would be interesting to explore potential associations between rater stance as expressed on a dilemma scale and actual ratings. Second, if one or several scales were to be used again, it would be necessary to optimise the psychometric qualities, such as by dismissing items with a low or negative discrimination. Third, it would be of great interest to validate the tentative findings by using techniques such as think aloud protocols, which could reveal the rationale behind actual ratings. Fourth, future investigations should include additional participants from different types of rating panels. If the findings then appear to be robust, rater positioning may be an important element of rater training.

Unfortunately, the research results have added to the catalogue of explanations of rater variation. Should these preliminary findings be representative of other writing assessment settings as well, professionals responsible for rater training programmes must take them into account.

## Literature

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Baird, J., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 331–348. https://doi.org/10.1080/0969594042000304627

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. (2005). Studie 3: Skriveprøvens pålitelighet. In K. L. Berge, L. S. Evensen, F. Hertzberg, & W. Vagle (Eds.), *Ungdommers skrivekompetanse. Bind I: Norsksensuren som kvalitetsvurdering* (pp. 101–113). Oslo: Universitetsforlaget.

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

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., … Welsh, M. E. (2016). A Century of Grading Research. *Review of Educational Research*, *86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Brown, G. T. L. (2004). Teachers’ conceptions of assessment: implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301–318. https://doi.org/10.1080/0969594042000304609

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

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers’ conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, *27*(1), 210–220. https://doi.org/10.1016/j.tate.2010.08.003

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Crusan, D. (2014). Assessing Writing. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 201–215). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla067

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

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

Edgeworth, F. Y. (1890). The Element of Chance in Competitive Examinations. *Journal of the Royal Statistical Society*, *53*(3), 460–475, 644–663.

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Huot, B., & Neal, M. (2006). Writing Assessment. A Techno-History. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 417–432). New York.

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

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior — a longitudinal study. *Language Testing*, *28*(2), 179–200. https://doi.org/10.1177/0265532210384252

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. R. (2017). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assessment in Education: Principles, Policy & Practice*, 1–26. https://doi.org/10.1080/0969594X.2016.1268090

Matlock-Hetzel, S. (1997). Basic Concepts in Item and Test Analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, January, 1997. Retrieved from http://ericae.net/ft/tamu/Espy.htm

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: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.

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.

Parr, J., & Brown, G. T. L. (2015). Learning about writing: A consideration of the recently revised asTTle: Writing. *Curriculum Matters*, *11*, 134–154. https://doi.org/10.18296/cm.0008

Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing*, *21*(3), 249–258. https://doi.org/10.1191/0265532204lt283ed

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

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

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

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

Stiggins, R. J. (1991). Assessment Literacy. *Phi Delta Kappan*, *72*(7), 534–539.

Tengberg, M., Borgström, E., Lötmarker, L., Sandlund, E., Skar, G. B., Sundqvist, P., Walkert, M. & Wikberg, K. (2017). *Likvärdig bedömning av elevers språkförmågor*. Karlstad: Karlstads universitet.

van Blerkom, M. L. (2017). *Measurement and statistics for teachers*. New York: Routledge.

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

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, *16*(3), 194–209. https://doi.org/10.1016/j.jslw.2007.07.004

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

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010

Östlund-Stjärnegårdh, E. (2009). Bedömaröverensstämmelse - ämnet Svenska. In *Bedömaröverensstämmelse vid bedömning av nationella prov. 2009-04-14. Dnr 2008:286*. Stockholm: Skolverket. Retrieved from http://www.skolverket.se/publikationer?id=2200