Co-equal Participation and Accuracy Perceptions in Communal Writing Assessment

by Vivian Lindhardsen, Teachers College, Columbia University

The present study examines the extent of raters' co-equal engagement and accuracy perceptions in a communal writing assessment (CWA) context, where raters collaborate to reach final scores on student scripts. Results from recorded discussions between experienced CWA raters when they deliberated to reach a final score supplemented with their retrospective reports show that, although some raters were more verbose than their co-raters, they displayed signs of co-equal engagement and reached what they perceived to be the most accurate scores possible for the student scripts. This study supports a hermeneutic approach to examining validity in writing assessment.

Keywords: CWA, score accuracy, hermeneutic, co-equal, collaborative assessment, score negotiation.

Introduction

In writing assessment, as in any performance-based assessment, we aim for our raters to reach a valid score, one that most accurately reflects the writing ability of the test taker. In traditional psychometric testing, this has typically involved implementing rigorous reliability-boosting standardization procedures that include detailed scoring rubrics, as well as rater training programs that train raters to rate consistently and uniformly. However, despite rigorous standardization procedures, various studies have shown that variance persists (e.g., Ecke, 2008; Kobayashi, 1992; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000, 2003; Vaughan, 1991). Even when normed, raters respond differently to written scripts, and when they are presented with a rubric, they often find it difficult to seamlessly match a concrete piece of writing to a set of abstract descriptors (Connor-Linton, 1995, Lumley, 2002, 2005). Charney (1984) argued that such standardization makes reader responses in writing assessment unnatural because it does not allow engaged and thoughtful responses to student writing. Ironically, performance-based writing assessment was introduced to make writing tasks more authentic and allow for independent and complex reader responses, but a scoring procedure that stresses conformity potentially works against this as it attempts to control and reduce the complexity and authenticity of raters' responses to authentic writing. Huot (1996) calls this dilemma "a stalemate for writing assessment" (p. 552), and Elbow (1991) describes it as a "tension between validity and reliability" (p. xiii) because there is a potential conflict between the more objectivist paradigm of psychometrics and the more social constructivist paradigm of composition theory (Elliot, 2005; Lynne, 2004).

Communal writing assessment (CWA) has been put forth by a number of writing assessment scholars to address the shortfalls of psychometric assessment procedures in writing assessment (e.g., Broad, 2003; Broad & Boyd, 2005; Condon, 2013; Huot, 2002; Lindhardsen, 2018; Matre & Solheim, 2016; Nixon & McClay, 2007). It is a practice that involves at least two raters collaborating to reach a final assessment. It typically consists of two phases: an independent phase where the raters assess student scripts independently and a communal phase where they meet with another (or several other) rater(s) to deliberate and reach a joint score. In most CWA cases, there is little or no discussion if the raters come to the communal session with similar assessments (although deliberation is always possible if the raters feel the need to have their initial assessments validated).

The collaboration takes the form of discussions where the raters make their individual interpretations and judgments explicit so that their assessments, their reader/rater responses, can be scrutinized and validated by a fellow rater. The assessment procedure allows for engaged responses to student scripts while at the same time making sure that ratings do not become idiosyncratic because the raters' assessment are being validated in the process. After deliberating these assessments, the raters might not agree on the final decision, but at least arguments have been heard and discussed. Raters may change their assessments during the process, ideally not because they are coerced, but because a sounder argument for a more valid assessment has been put forth.

This practice of collaborating to reach a mutually agreed assessment differs markedly from that taken by psychometrically-driven written assessment procedures that stress, "Once live rating is under way, it is important to ensure that scoring is independent – that is, that raters do not see and therefore cannot be influenced by scores given by other raters" (Weigle, 2002, p. 131). Obtaining two or more independent ratings, rather than a single, communally-based rating, of course, potentially increases reliability estimates, which, from a psychometric standpoint, is the goal. However, in CWA, there is more value in one score that is the result of a deliberation between raters than having two independent raters end up with the same score. In CWA, human variance is not seen as a measurement error. In fact, variance is not only seen as an inherent trait of human behavior; it is viewed as potentially beneficial because it can bring about a synergy that can generate mutually illuminating assessments (Moss & Schutz, 2001). An added benefit to CWA is that, by engaging in mutual validation of their assessments, raters can continuously sharpen their assessment skills and develop as raters.

It must be stressed that validity and reliability are not necessarily mutually exclusive in CWA. Consistency is still valued, but the primary goal is to get at a valid score, one that respects human engagement and where subjective assessments are illuminated and open to scrutiny. Also, the psychometric aspect of standardization is not ignored in CWA as raters often work with a rubric (typically a rubric designed beforehand by experienced raters and teachers), but CWA recognizes that deliberation is important because the

complex task of assessing a piece of writing is more than just matching a script to a rubric.

CWA operates within a hermeneutic epistemology rather than a purely psychometric one. Whereas psychometric scoring assumes that there is a "reality out there, driven by immutable laws" (Guba, 1990, p. 19), for instance in the form of a detailed rubric to which raters have to adhere consistently, a hermeneutic approach to validity is more post-positivistic in that it does not share this assumption of the existence of an absolute truth. Rather, it recognizes that interpretations and judgments are filtered through human engagement and experience and that the value of a student performance is unavoidably subjective. It is the possibility of examining the value of human assessments through deliberations that brings us closer to a valid picture of student performance. As Williamson (1993) pointed out, "Explicitness about the process of decision-making through testing is perhaps the only basis for validity in a postmodern, postpositivistic world" (p. 13). Wiggins (1993) likewise stated that "all assessment is subjective; the task is to make the judgment defensible and credible" (p. 136).

Empirical Studies on Collaborative Benefits of Communal Ratings

Although still in their infancy, empirical studies in communal assessment have shown budding hermeneutic potentials. Many earlier studies focused on the deliberations taking place during the communal rating sessions rather than on the score outcome. Thus, Condon and Hamp-Lyons (1994) reported on communal assessment sessions where raters actively exchanged their rating practices and concluded that, based on such exchanges, the resulting assessment could be considered "more accurate and more fair" (p. 284) although there were no reports on what constituted accurate assessments. Mohan and Low (1995) described in more detail the deliberations taking place during communal assessment sessions. They found lengthy rater discussions in which raters clarified their ideas, reflected on them, and searched for concrete evidence by going back to concrete passages in student scripts. Like Condon and Hamp-Lyons, Mohan and Low interpreted such interactions as potentially leading to fairer evaluations. Although his studies focused more on standard setting than on just rating, Broad (2000, 2003) also identified great potentials for communal assessments. He investigated the hermeneutic standardization practices of CWA and concluded that CWA raters took great pains to understand different and sometimes extreme perspectives and were even sometimes swayed by other standpoints. While not focused on standardized tests as such, but rather on writing pedagogy in general, Nixon and McClay (2007) conducted a case study of the interactions of three elementary school teachers and concluded that these dialogues encouraged reframing of beliefs about writing pedagogy and promoted more engaged grading practices.

The potentials of mutually illuminating assessments have been echoed in more recent studies. Jølle (2014) investigated the extent to which communal assessment raters changed their behaviors from one communal assessment to another and found that, although their behaviors did not change much from one session to the other, they did engage in several mutually illuminating meta-discussions. Matre and Solheim (2016) also examined characteristics of rater dialogues over two sessions. They found a mixture of instrumental approaches (where the raters rather mechanically used the norms as checklists and mostly commented on linguistic surface structures), and functional and flexible approaches (where raters challenged each other's contributions and treated several aspects of the student scripts at a deeper level) but noticed that some raters moved towards more functional and flexible approaches as they became more experienced. They interpreted this as progress towards sounder quality assessment skills and viewed the dialogues as highly relevant for professional development. Lindhardsen (2018) likewise found traces of mutual validation between raters during communal writing assessment sessions. With the use of think-alouds and rater discussion recordings, she systematically investigated the decision-making behaviors of raters from when they form their initial assessments during their independent rating sessions to when they deliberate and finalize their assessments in communal rating sessions. She found that the communal raters went about their task conscientiously making sure to illuminate and validate their assessments: They defined, revised, or suggested assessment strategies; they presented their overall impressions; and they gave concrete examples from the student scripts as evidence for their claims. Further, she looked at the textual features to which raters paid attention during the independent assessment sessions and the communal assessment session. She found that, as the raters moved from independent rating sessions to communal rating sessions, the attention they paid to the stated assessment criteria became more balanced; that is, the raters' comments on the textual features were more evenly distributed in the communal assessment session than in the independent assessment sessions. This led Lindhardsen to believe that raters refined their assessments during the communal assessment sessions and managed to weed out idiosyncratic assessments. The studies above point to validation taking place during CWA, and from a hermeneutic point of view, the fact that assessments are illuminated and validated leads us closer to a valid score.

Taking a more quantitative approach to determining validity of scores, Johnson, Penny, Gordon, Shumate, and Fisher (2005) looked at how closely communally-rated (i.e., CWA) scores and averaged scores approximated the scores produced by a validation committee (expert raters). Their results showed that the accuracy of the CWA scores and the averaged scores were almost the same when scored holistically. The same was the case with the use of an analytic rubric for the domains of style, conventions, and sentence formation. However, the CWA approach fared slightly better (i.e., showed no statistical significance) than the averaged-based approach with respect to the domains of content and organization. In no instances was the correlation between the averaged-based score and the scores from the validation committee higher than the correlation between the CWA scores and scores from the

validation committee. The findings that the communally-rated scores only came slightly closer to expert scores than averaged independent scores indicate that, from a psychometric point of view, CWA practices are not superior to traditional assessment practices, but it also indicates that CWA is not inferior. The benefits of CWA, then, seems that raters, by engaging in mutually illuminating discussions, validate their interpretations and judgments. From a hermeneutic point of view, this is quality assurance.

An important precondition for communal assessment to be mutually illuminating and produce valid scores is, of course, that the communally-rated assessments be a product of co-equal rational debate, rather than a product of oppressive coercion where a more assertive voice dominates. Everybody's voice should be heard and validated. Needless to say, a rater should recede to a stronger argument but should not uncritically recede to a stronger voice. Emphasizing this precondition, Moss and Schutz (2001) assert that co-equal participation is an ideal we must strive for although they acknowledge that conversational power relations affect any dialogue in subtle ways and can perhaps never be fully reached.

A few studies have alluded to a co-equal participatory nature of communal ratings. Allen (1995) examined internet-based discussions and found that, despite initial anxieties about criticizing each other, the raters showed "an ethic of disciplined collaborative inquiry that encourages challenges and revisions to initial interpretations" (p. 68). Allen did not, however, specify what the basis was for his perception of collaborative inquiry. Likewise, raters in Moss, Schutz, and Collins's (1998) study on communal ratings of teachers' portfolios in Math reported no significant inequalities in reaching judgments although Moss et al. themselves noted some asymmetry in the rater discussions, both with respect to writing roles (one rater usually took on the role of actually writing down the final judgment) and speaking roles (although the researchers did not indicate how this asymmetry in speaking was manifested). Matre and Solheim (2016) observed that raters generally adopted symmetrical roles in rater discussions but also detected some asymmetry. Again, how this symmetry or asymmetry was realized is less clear. Johnson et al. (2005) also looked into the extent of asymmetry in communal ratings. However, rather than directly examining co-equal participation in discussions like Moss et al. did, they examined score dominance. Operating under the assumption that, if raters are equally engaged in the rating discussions, "it also seems reasonable to expect that the discussion scores would agree equally with the scores from the original raters" (Moss et al., 1998, p. 126), Johnson et al. examined to what extent the communal scores tended to agree more with the independent scores of one rater than the other. Johnson et al. (2005) did find that discussion-based (i.e., CWA) scores agreed more frequently with the original score of one of the raters, "indicating the possibility of rater dominance or deference" (p. 139). However, this trend was not statistically significant. Further, it was reported that, in instances of score dominance (when the communally rated score agreed more frequently with the original score of one rater), there was little influence on "score accuracy" (Johnson et al., 2005, p. 140), as in such instances there was a .86 correlation between the communally-rated score and the expert-criterion-related score from the validation committee.

The few studies on CWA conducted so far have pointed to budding hermeneutic assessment practices where sound assessments can be reached based on critical engagement. A couple of studies have also hinted that CWA sessions are largely characterized by non-coercive interactions. However, these studies are sparse and eclectic, and they focus on novice CWA raters. The present study seeks to shed more light on the extent to which communal writing assessments indeed demonstrate the kind of deliberate democracy that is sought after in communal assessments. More specifically, the present study seeks to uncover signs of asymmetry in experienced CWA interactions, as well as raters' own perceptions of how accurate CWA scores are. The specific research questions were as follows:

1. What is the extent of co-equal participation in CWA?
2. What is the extent of score dominance in CWA?
3. What is the relationship between co-equal participation and score dominance?
4. What are the raters' perceptions of CWA?

Methods

The data collection was conducted as part of a dissertation. It was reviewed by the institution's review board and in compliance with the institution at which the dissertation was written. The data for this study were drawn from a high school written English as a Foreign Language (EFL) exam (HHX) in Denmark, a country that for many years has employed communal assessments all through its education system (Haue, 2000). The 20 raters selected for this study were all CWA raters with several years of experience with CWA procedures and with the types of student scripts written for this exam. They were part of a national CWA rater corps employed by the Danish Ministry of Education who rate HHX's written EFL exams. This rater corps rate all the HHX written EFL exams in Denmark, and they help create the list of assessment criteria on which the student performances are assessed. The profiles of the raters participating in this study are listed in Table 1. Their names were pseudonyms, but their genders were preserved. All raters signed informed consent forms.

Table 1

Rater Profiles

Characteristics	*n*	Percent %
Gender		
Male	6	30
Female	14	70
Age		
31-40	3	15
41-50	6	30
51-60	7	35
> 60	4	20
Highest Educational Level		
MA	18	90
BA	2	10
Teaching Experience		
1-5	0	0
6-10	1	5
11-15	2	10
16-20	7	35
> 20	10	50
CWA Rating Experience		
1-5	0	0
6-10	3	15
11-15	7	35
16-20	8	40
> 20	2	10

The student scripts rated in this study were essays written in response to an integrated writing task based on two texts. The assessment criteria were length, organization, content, and use of source materials, language command, and style and format. The scripts were written by high school students for an EFL high school exit exam. They had received approximately 900 hours of instruction in English as a foreign language. Participating students signed informed consent forms.

The CWA procedure in this study was the same as the usual CWA procedure in the actual exams: Each rater rated a set of student scripts independently and subsequently met face-to-face with a co-rater to finalize their assessments and determine a final score for each script. The raters were given as much time as they needed for both the independent and the communal rating session. Each of the 20 raters (10 rater pairs) in this study rated 15 student scripts. They rated the scripts blindly.

To address the first research question, the extent of equal participation during the rater interactions, the CWA sessions were recorded and transcribed. They were subsequently segmented and coded, according a coding scheme used as part of a dissertation study into the nature of raters' distinct decision-making behaviors (see Lindhardsen, 2009, 2018) and was designed based on Cumming, Kantor, and Powers's (2001, 2002) coding scheme. The coding scheme was created to capture the raters' decision-making behaviors in terms of how they interpret and judge student scripts, as well as their textual focus (length, organization, content and use of source materials, language command, and style and format) and their monitoring and contextual focus. See Appendix A for the coding scheme.

Following the guidelines of Allwright and Bailey (1991) and Cumming et al. (2001, 2002), the criteria below were used for dividing the protocols into separate, comparable units:

- Pauses of five or more seconds (marked by three dots (…) in the transcriptions)
- The reading of a segment in the student script (marked by capital letters in the transcriptions)
- The start or the end of an assessment
- Each conversational turn if that turn was not an uptaker, such as "aha," "I agree," or a mere repetition of what the other person just said and with no new information added.

To determine whether the raters participated co-equally in the CWA sessions, number of words as well as number of decision-making behaviors were counted for each rater in each rater pair. Co-equal participation in conversations is a complex matter and can, of course, be measured in various ways, such as verbosity, initiating turns, interruptions, syntax structures, or choice of words (see Itakura, 2001, for various approaches to measuring conversational dominance). Well aware that a multiperspectival approach to operationalizing co-equal participation would offer a more in-depth analysis, I chose to take the quantitative measure of counting words and counting decision-making behaviors. Counting words offers an indication of verbosity—as does counting decision-making behaviors. But, as uptakers and mere repetition of words did not count as independent decision-making behaviors, the number of decision-making behaviors would also add some degree of topic control (see above on segmenting and coding the verbal data). Paired *t*-tests were conducted to compare means of the number of words and decision-making behaviors and determine statistical significance.

To address the second research question, the extent to which one rater dominated by score, scores assigned in the independent rating sessions and in the communal rating sessions were compared. Following Johnson et al.'s (2005) guidelines, a rater was said to dominate by score if their independent score was closer to the final, communally-rated score than their co-rater's score was. If, however, the rater's independent score was further away from the communal score, then they could be said to concede their score.

To get a more elaborate picture of whether the raters in the communal rating sessions were "acquiescing to the more assertive voice" (Moss, 1996, p. 26) and to see whether Johnson et al.'s (2005) assumption that score dominance is a product of lack of co-equal participation in rater conversations makes sense, the relationship between score dominance and co-equal participation was examined (Research Question 3). Each case of score dominance was examined for co-equal participation by checking whether the rater in the rater dyad who dominated by score would also be the one who dominated the conversation in terms of amount of words and amount of decision-making behaviors.

To shed further light on the extent to which CWA offers opportunities for sound and democratic assessments, the raters were asked retrospectively about their perceptions of score accuracy and about their perceptions of communal assessments in general. This last part of the questionnaire was fully open-ended with no a priori categories set. These rater reflections were obtained right after the assessment sessions had taken place. Not only would perceptions of accurate scores contribute to the overall validity of CWA, they would also indicate to what extent raters felt that the communally-rated scores were a product of careful and reasoned deliberation and not of oppressive coercion.

Results

To determine the extent to which the raters in this study engaged equally in the communal rater discussions, levels of equal participation and levels of score dominance are reported along with the relationship between co-equal participation and score dominance. Results from the retrospective questionnaire are presented to obtain a deeper understanding of the extent of co-equal participation and the raters' perception of accuracy of scores.

Operationalized as number of words and number of decision-making behaviors, Tables 2 and 3 show the extent of equal participation in terms of distribution of number of words and decision-making behaviors for each rater in each rater dyad. Table 2 summarizes the distribution of words in each rater dyad, and Table 3 displays the distribution of decision-making behaviors. Along with the distribution of words and decision-making behaviors, the *t* statistics and the *p*-values are reported for each rater pair. As can be seen from Table 2, raters differed on average 15% from each other in the rater dyads in terms of words spoken during the communal rating sessions, the difference ranging from 2% to 38%. This pattern is echoed in the distribution of decision-making behaviors (Table 3), the average difference between the raters in the rater dyads being 15.5%, ranging from 0% to 30%. The distribution of words and decision-making behaviors were aligned in that the raters who produced more words than their co-rater were also the raters that produced more decision-making behaviors. T-tests were conducted, and operating with an alpha level of .05, we can see that out of the 10 rater dyads, four showed statistical significance in the difference of words produced between the two raters, and six dyads showed no statistically significant difference. With respect to distribution of decision-making behaviors, six dyads showed a statistically significant difference, and four dyads did not. These numbers suggest that, while no rater in any pair completely dominated the conversations, there was significant variance in half the rater dyads with respect to how equally raters contributed to the rater discussions. This points to an interactional atmosphere where each rater is given a chance to voice their opinion, but it also indicates that there is not perfect co-equal participation.

Table 2

Number of Words in Communal Ratings (spoken per script by each rater in the 10 rater dyads)

Rater Dyad	Number of Words, *M (SD)*	Percent of Total Amount of Words	Difference in Number of Words (percent)	*t*	*p*
1. Pernille	199.13 (162.12)	51%	10 (2%)	0.42	.68
1. Jesper	188.93 (128.64)	49%			
2. Gitte	187.13 (81.33)	62%	72 (24%)	2.93	.01
2. Julie	115.07 (77.45)	38%			
3. Torben	258.20 (162.75)	47%	34 (6%)	-2.17	.05
3. Tina	292.07 (150.36)	53%			
4. Tove	195.20 (59.53)	58%	54 (16%)	2.95	.01
4. Susanne	14.53 (54.09)	42%			
5. Nina	204.27 (134.48)	38%	133 (24%)	-4.18	0.05
5. Jens	336.87 (211.05)	62%			
6. Lone	108.73 (56.43)	55%	21 (10%)	2.29	.04
6. Louise	88.00 (51.22)	45%			
7. Astrid	285 (261.7)	56%	60 (12%)	1.82	.09
7. Helle	225 (161.4)	44%			
8. Thea	163 (115.7)	40%	77 (20%)	-4.73	.01
8. Malene	240 (141.0)	60%			
9. Jette	93.00 (43.44)	53%	10 (6%)	1.19	.25
9. Ken	83.40 (29.11)	47%			
10. Hans	232.60 (91.50)	69%	130 (38%)	5.81	.01
10. Henrik	103.13 (38.62)	31%			
M (*SD*)			60.1 (42.3) 15% (10.4)		

Table 3

Number of Decision-Making Behaviors in Communal Ratings (produced per script by each rater in the 10 rater dyads)

Rater Dyad	Number of Decision-Making Behaviors, *M (SD)*	Percent of Total Amount of Decision-Making Behaviors	Difference in Number of Decision-Making Behaviors (percent)	*t*	*p*
1. Pernille	9.20 (6.21)	52%	1 (5%)	0.6	.56
1. Jesper	8.33 (5.14)	47%			
2. Gitte	11.27 (3.73)	58%	3 (16%)	4.19	.01
2. Julie	7.53 (3.38)	42%			
3. Torben	11.60 (4.84)	46%	2 (8%)	-2.17	.05
3. Tina	14.00 (4.81)	54%			
4. Tove	11.07 (4.64)	58%	3 (16%)	1.81	.09
4. Susanne	7.93 (2.84)	42%			
5. Nina	10.40 (5.97)	40%	5 (20%)	-3.29	.01
5. Jens	14.60 (6.19)	60%			
6. Lone	7.27 (3.20)	58%	2 (16%)	2.99	.01
6. Louise	5.00 (2.56)	42%			
7. Astrid	16.27 (9.57)	59%	5 (18%)	2.91	.01
7. Helle	10.73 (4.25)	41%			
8. Thea	7.40 (3.96)	37%	5 (26%)	-4.58	.01
8. Malene	11.60 (5.77)	63%			
9. Jette	6.87 (3.34)	50%	0 (0%)	0.45	.66
9. Ken	6.47 (1.41)	50%			
10. Hans	13.27 (3.20)	65%	6 (30%)	5.54	.01
10. Henrik	7.00 (2.93)	35%			
M (*SD*)			3.2 (1.9) 15.5% (8.7)		

To supplement the examination of co-participation in the communal rating sessions, I investigated the extent to which the scoring decisions were co-equal or whether one rater tended to dominate by score. As described above, score dominance refers to the distance between the independent scores of each of the two raters in a rater pair and their final, communally-rated score. The rater whose independent score was closer to the communally-rated score dominated by score, and the rater whose score was further away from the communally rated score conceded their score.

A total of 150 final, communally-rated scores were assigned (10 rater dyads rating 15 student scripts each). The scripts were rated on a 10-point scale. In 70 (47%) of these cases, the raters came to the communal rating sessions with similar independent scores, and so the independent scores became the final score. In 80 (53%) of the cases, however, the raters entered the CWA sessions with discrepant independent scores. In 15 (19%) of these discrepant score cases, the final, communal score became a balanced compromise between the independent scores; that is, both raters conceded their independent scores by one point, so the final score fell midway between the two raters' independent scores. In 65 of the discrepant score cases (43% of all cases), however, one rater conceded their score, so the final score was further away from the independent score of one rater than the independent score of the other rater. The relatively high number of concession cases is a product of the adjacency of scores: In most cases, one of the raters had to concede to the other because their original independent scores were adjacent (e.g., one rater had an 8 as the original independent score, and the other rater had a 9, and because no half scores could be assigned, one of the raters had to concede). Table 4 shows the score dominating/conceding behaviors of each of the raters.

Table 4

Score Dominance in Communal Ratings

Rater Dyad	Rater	Number of Score Dominations	Number of Score Concessions	Total Number of Final Scores Assigned
1	Pernille	3	7	15
	Jesper	7	3	
2	Gitte	6	3	15
	Julie	3	6	
3	Torben	2	1	15
	Tina	1	2	
4	Tove	0	1	15
	Susanne	1	0	
5	Nina	2	3	15
	Jens	3	2	
6	Lone	3	1	15
	Louise	1	3	
7	Astrid	9	1	15
	Helle	1	9	
8	Thea	6	3	15
	Malene	3	6	
9	Jette	3	5	15
	Ken	5	3	
10	Hans	2	4	15
	Henrik	4	2	
	Total	65	65	150

As can be seen from Table 4, although there were no cases of one rater asserting complete score dominance over their co-rater, there were cases where one rater had more than half their share of score dominance over their co-rater. For instance, out of the 15 scripts scored, Jesper dominated by score seven times whereas his co-rater, Pernille, dominated by score only three times; out of the 15 scripts scored, Lone dominated by score three times whereas her co-rater, Louise, only dominated by score once; more pronounced is the score dominance in the Helle/Astrid rater pair: Out of the 15 scripts scored, Astrid dominated by score nine times whereas Helle dominated by score only once.

To determine whether the raters who dominated by score also dominated the conversation within the dyads, I examined each rater pair for its relationship between score dominance and co-equal participation in the conversations. Table 5 below juxtaposes for each rater dyad the distribution of score dominating/conceding cases and their number of words and decision-making behaviors.

Table 5

Juxtaposition of Score Dominance and Amount of Words and Amount of Decision-Making Behaviors Produced

Raters	Number (percent) of Score Dominance in Rater Pair	Number (percent) of Total Amount of Words in Rater Dyad	Number (percent) of Total Amount of Decision-Making Behaviors in Rater Dyad
Rater Pair 1			
Pernille – score conceding	3 (30%)	199 (51%)	9 (52%)
Jesper – score dominating	7 (70%)	189 (49%)	8 (47%)
Rater Pair 2			
Gitte – score dominating	6 (67%)	187 (62%)	11 (59%)
Julie – score conceding	3 (33%)	115 (38%)	8 (42%)
Rater Pair 3			
Torben – score dominating	2 (67%)	258 (47%)	12 (46%)
Tina – score conceding	1 (33%)	292 (53%)	14 (54%)
Rater Pair 4			
Tove – score conceding	0 (0%)	195 (58%)	11 (58%)
Susanne – score dominating	1 (100%)	141 (42 %)	8 (42%)
Rater Pair 5			
Nina – score dominating	2 (40%)	204 (38%)	10 (40%)
Jens – score conceding	3 (60%)	337 (62%)	15 (60%)
Rater Pair 6			
Lone – score dominating	3 (75%)	109 (55%)	7 (58%)
Louise – score conceding	1 (25%)	88 (45%)	5 (42%)
Rater Pair 7			
Astrid – score dominating	9 (90%)	285 (56%)	16 (59%)
Helle – score conceding	1 (10%)	225 (44%)	11 (41%)
Rater Pair 8			
Thea – score dominating	6 (67%)	163 (40%)	7 (37%)
Malene – score conceding	3 (33%)	240 (60%)	12 (63%)
Rater Pair 9			
Jette – score conceding	3 (38%)	93 (53%)	7 (50%)
Ken – score dominating	5 (62%)	83 (47%)	7 (50%)
Rater Pair 10			
Hans – score conceding	2 (33%)	233 (69%)	13 (65%)
Henrik – score dominating	4 (67%)	103 (31%)	7 (35%)

As can be seen from Table 5, the rater who dominated by score was not necessarily the one who was the more verbose rater during the conversations (i.e., produced more words and decision-making behaviors than their co-rater). In fact, out of the 10 rater pairs, only four pairs had one rater dominate by both score and conversation. In the remaining six pairs, the raters who dominated by score were the less verbose. Ironically, the rater who seemed to dominate the conversation the most (Hans produced on average 233 words and 13 decision making behaviors per script, taking up 65% to 68% of the words and decision-making behaviors compared to his co-rater) tended to concede his scores: He dominated by score two times, and his co-rater dominated by score four times, twice that amount.

To further validate the CWA scores, the raters in this study were asked to reflect on the cases of score discrepancy (80 cases) and to state retrospectively which score they found more accurately reflected student performance: their own score from the independent rating session or the final score assigned in the CWA session. These perceptions are displayed in Table 6.

Table 6

Raters' Perceptions of Score Accuracy

Number of discrepant scores cases (Adjacent scores, such as 8 and 9, were also considered discrepant).	80
Number score discrepancy cases in which raters believed their independent score to be more accurate	4 (5%)
Number of discrepancy cases in which raters believed the final, communally-rated score to be more accurate	76 (95%)

In a striking 95% of the cases in which there was a discrepancy in the raters' independent scores and the scores in the communal ratings, the raters viewed the scores finalized in the communal sessions to be more accurate than the scores they assigned on their own in the independent rating sessions. This indicates a strong faith in CWA as a valid scoring method. Despite the raters' faith in CWA's ability to produce accurate scores, we must caution this against the fact that all raters in this study were seasoned CWA raters and part of the national rater corps, and thus their overwhelmingly positive attitude might be colored by a subconscious desire to validate their jobs as CWA raters.

In the retrospective questionnaire, the raters were also asked what they perceived to be the advantages or disadvantages of co-rating procedures in general. All 20 raters answered the questionnaire. Their comments (in Danish, but translated into English in this article) were all positive and fell into the following broad a posteriori categories as shown in Table 7. Raw numbers of raters who made comments are included along with a percentage of raters who made those comments. See Appendix B for the full range of comments.

Table 7

Raters' General Perceptions of CWA

Perceptions of CWA (20 raters)	Frequency (percent) of raters who gave such a comment
They offer the opportunity to reach the most accurate score possible.	17 (85%)
They offer raters an opportunity to refine their assessment strategies in general.	4 (20%)
They ensure raters assess by the same standards.	5 (25%)

Most of the raters (85%) pointed to CWA's ability to offer the opportunity to reach the most accurate score possible. Of these, many commented that "four eyes are better than two" (Gitte, Tina, Jens, Jette, Ken). More specifically, some raters mentioned that the reason CWA opens an opportunity to reach the most accurate score possible is that the discussions weed out idiosyncratic assessments: "One's own idiosyncrasies or preferred aspects don't carry too much weight" (Jesper), and they prevent mistaken judgments from counting towards final scores: "It does happen that one has to revise one's assessment because one has overlooked things, was distracted" (Susanne). One rater also believed that the communal ratings made sure too much focus was not put on language errors:

> Another thing is the importance attached to the content and language of the scripts. You can focus so much on grammatical errors that you actually forget that the student has attempted to construct complex sentences about some idea, so there will be a lot of extra errors. (Gitte)

Four raters commented on the opportunities CWA offers for refining and reassessing general assessment strategies, indicating a rater development potential inherent in such assessment practices. As one rater put it: "We all need to test our judgments against others" (Susanne).

Five raters mentioned CWA makes sure raters assess by the same standards. One could argue that conscious matching of the scripts to the rating scale would result in assessments by the same standards. However, as one rater pointed out, such a matching exercise is close to impossible: "It is important because raters must judge by the same standards. There are so many things you cannot make rules for. You cannot just assign a score like that. That is why conversation is important" (Pernille).

As with the perceptions of the accuracy of assigned scores, the raters' comments on the CWA practices in general may have been

overwhelmingly positive because the raters are experienced CWA raters. Their positive attitude to CWA could be a way of validating their jobs as raters. Despite the probability that raters' perceptions of CWA might be colored by their loyalty to their profession as CWA raters, CWA was perceived to hold strong advantages. Not only was CWA perceived to produce more accurate scores because idiosyncratic prejudices can be illuminated, they were also perceived to hold a rater development potential in that raters receive a chance to validate their assessments with and against other professionals.

Discussion and Conclusion

This validation study of CWA as a writing assessment procedure supports earlier results, indicating that CWA represents fertile ground for reaching hermeneutically valid scores of writing ability. Along with studies that show raters engaging critically in rater discussions to make sure their interpretations and judgments are vetted and validated (Broad, 2000, 2003; Condon & Hamp-Lyons, 1994; Jølle, 2014; Lindhardsen, 2009, 2018; Mohan & Low, 1995; Nixon & McClay, 2007), this study suggests raters also engage in discussions where each rater's voice is heard, and authentic and complex responses can be put forth to refine assessments. About half of the rater dyads in this study exhibited statistical significance in the number of words and decision-making behaviors produced, indicating there was not complete co-equal participation. This was not, however, reflected in score dominance as the more verbose raters (in terms of words and decision-making behaviors) did not tend to be ones dominating by score. Thus, the study did not support Johnson et al.'s (2005) assumption that if raters dominate by score, they also dominate the discussion. It must be stressed, however, that focusing only on verbosity, operationalized as number of words and decision-making behaviors, as this study does, offers a mere glimpse of the extent of equal engagement in raters' discussions. A more nuanced investigation into the intricacies of power dynamics in discussions, for instance by focusing on discourse features such as initiating turns, interruptions, number of hedges, types of suprasegmentals, would provide a more nuanced picture of the extent of equal engagement.

Underscoring the increased validity potentials as a result of critical engagement, retrospective reports indicated that CWA raters perceived CWA to produce more accurate scores than their own individual scores. The reports also showed that raters stressed the importance of deliberation to allow for authentic and complex responses to student scripts while also ensuring that assessments are validated to avoid idiosyncratic assessments. Again, there are limitations associated with these findings, as the raters' overwhelmingly positive attitude towards CWA could be the result of an indirect attempt to validate their jobs as CWA raters.

Despite these limitations, from a hermeneutic perspective, CWA demonstrates promising potentials for reaching valid scores. With the added benefit of refining assessment strategies in general, as reported by raters in the current study as well as other studies, the benefits of CWA in terms of accuracy of scores does not just pertain to a single CWA event, but also to future CWA sessions in that the reframing that takes place during CWA discussions awards CWA raters with continuous professional assessment development that can fine-tune their assessment skills for future rating sessions.

Finally, the epistemological underpinnings of CWA align with an understanding that we come closer to a performance's true value if our different viewpoints are respected and made explicit and open to scrutiny by others. Consistency of scoring, while valued, is not the ultimate goal. Validated reasoning is. In CWA, through the unraveling and reframing of the interpretations and judgments of student scripts, new and more refined values can be generated and explored.

Author Note

Vivian Lindhardsen, Ph.D., is a lecturer in Language and Education at Teachers College, Columbia University. She received her Ph.D. in TESOL/Applied Linguistics from Copenhagen Business School, Department of Management, Society, and Communication, and her MA and BA in English Language and Literature from Copenhagen University.

References

Allen, M. S. (1995). Valuing differences: Portnet's first year. *Assessing Writing 2*(1), 67-89.

Allwright, D., & Bailey, K. (1991). *Focus on the language classroom: An introduction to classroom research for language teachers*. Cambridge: Cambridge University Press.

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2), 213-260.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan, UT: Utah State University Press.

Broad, B., & Boyd, M. (2005). Rhetorical writing assessment: The practice and theory of complementarity. *Journal of Writing Assessment, 2*(1), 7-20.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*(1), 100-108.

Condon, W., & Hamp-Lyons, L. (1994). Maintaining a portfolio-based writing assessment: Research that informs program development. In L. Black, D. A. Daiker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment* (pp. 277-285). Portsmouth, NH: Boynton/Cook.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762-765.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making, and development of a preliminary analytic framework* (TOEFL Monograph Series Report No. 22). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

Ecke, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Elbow, P. (1991). Foreword. In P. Belanoff & M. Dickson (Eds.), *Portfolios: Process and product*. Portsmouth, NH: Boynton/Cook.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang Publishing Inc.

Guba, E. G. (Ed.). (1990). *The paradigm dialogue*. Newbury Park, CA: Sage.

Haue, H. (2000). Prøver og eksamen – norm og udfordring – set i et historisk perspektiv. *Uddannelse*, 4. Copenhagen: Undervisningsministeriet.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*(4), 549-566.

Huot, B. (2002). *(Re)articulating writing assessment for teaching and learning.* Logan, UT: Utah State University Press.

Itakura, H. (2001). *Conversational dominance and gender*. Amsterdam: John Benjamin's Publishing Company.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly, 2*(2), 117-146.

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing*, *20*, 37-52.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26,* 81-112.

Lindhardsen, V. (2009). *From independent ratings to communal ratings*. Frederiksberg, Denmark: Samfundslitteratur.

Lindhardsen, V. (2018). From independent ratings to communal ratings: A study of CWA raters' decision-making behaviors. *Assessing Writing, 35*, 12-25.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing 19*(3), 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Lynne, P. (2004). *Coming to terms: Theorizing writing assessment in composition studies*. Logan, UT: Utah State University.

Matre, S., & Solheim, R. (2016). Opening dialogic spaces: Teachers' metatalk on writing assessment. *International Journal of Educational Research, 80*, 188-203.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment. Selected Papers from Studies in Language Testing 3*, 92-114. Cambridge: Cambridge University Press.

Mohan, B., & Low, M. (1995). Collaborative teacher assessment of ESL writers: Conceptual and practical issues. *TESOL Journal, Autumn*, 28-31.

Moss, P. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(1), 20-29.

Moss, P., & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal, 28*(1), 37-70.

Moss, P., Schutz, A., & Collins, K. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education, 12*(2), 139-161.

Nixon, R., & McClay, J. K. (2007). Collaborative writing assessment: Sowing seeds for transformational adult learning. *Assessing Writing, 12*, 149-166.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment. Selected papers from the 19th Language Testing Colloquium, Orlando, Florida* (pp. 129-152). Cambridge, England: Cambridge University Press.

Sakyi, A.A. (2003). *A study of the holistic scoring behaviours of experienced and novice ESL instructors* (Unpublished doctoral thesis). Department of Curriculum Teaching and Learning. The Ontario Institute for Studies in Education of the University of Toronto.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex.

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

Wiggins, G. (1993). *Assessing student performance*. San Francisco, CA: Jossey Bass.

Williamson, M. M. (1993). An introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton.