Relationship between Rater Background and Rater Performance

Research Report

Tian Song
Edward W. Wolfe
Lori Hahn
Molly Less-Petersen
Robert Sanders
Daisy Vickers

April 2014

About Pearson
Everything we do at Pearson grows out of a clear mission: to help people make progress in their lives through personalized and connected learning solutions that are accessible, affordable, and that achieve results, focusing on college-and-career readiness, digital learning, educator effectiveness, and research for innovation and efficacy. Through our experience and expertise, investment in pioneering technologies, and promotion of collaboration throughout the education landscape, we continue to set the standard for leadership in education. For more information about Pearson, visit http://www.pearson.com/.

About Pearson's Research Reports
Our network **mission** is to spark innovation and create, connect, and communicate research and development that drives more effective learning. Our **vision** is students and educators learning in new ways so they can progress faster in a digital world. Pearson's research papers share our experts' perspectives with educators, researchers, policy makers and other stakeholders.

Abstract

Although several researchers have posited relationships between rater prior experiences and rater backgrounds and their performance in operational scoring projects, very little empirical research exists to address this topic. This report summarizes how the background characteristics (demographics and professional experiences) of two samples of raters (one who scored a writing prompt and one who scored a science prompt) relate to measures of rater performance during the scoring project. Results suggest that small differences may exist between rater groups, but (a) these differences are not consistent across rater samples and (b) missing data precludes concluding whether important trends exist in these results.

Keywords: Performance assessment, Rater background, Scoring

Relationship between Rater Background and Rater Performance

In scoring projects that involve human raters, one ongoing question is whether there is a relationship between the backgrounds of raters and rater success in the scoring project. This is an important question because, should such a relationship exist, selection, placement, and training practices could be adapted to increase the likelihood of rater success for a particular scoring project. This report summarizes analyses that focus on whether raters who have different demographic and professional characteristics perform equivalently on measures of training and scoring success.

As a framework for our discussion of rater background and rating performance, we describe a model, originally presented by Pula and Huot (1993), which highlights raters' personal backgrounds, professional trainings, and work experiences as key influences on holistic scoring behaviors. Personal background may include the rater's reading and writing experiences as well as mentoring relationships relevant to the scoring task. Professional training may include formal training (K-12 and college) as well as professional development activities, such as professional conference attendance. Work experiences include both teaching and scoring experiences. Unfortunately, little research exists that seeks to determine whether these potential influences on scoring performance differentiate raters in operational scoring projects. Much of the studies that exist focus on the impact of raters' backgrounds on language test ratings (Johnson & Lim, 2009; Kachchaf & Solano-Flores, 2012; Lee, 2009; Shohamy, Gordon, & Kraemer, 1992; Weigle, Boldt, Valsecchi, Elder, & Golombek, 2003; Winke, Gass, & Myford, 2012). Only two of these studies produced results that suggest a meaningful relationship between raters' language backgrounds and the quality of the ratings that they assign. Winke, et. al. (2012) found that raters of a second language speech assessment who are familiar with the examinee's

native language tend to be more lenient than raters of the second language assessment for whom the examinee's native language is unfamiliar. Lee (2009) found that raters of a second language writing assessment for whom the target language is a second language assign scores that exhibit lower reliability and are systematically lower on grammar, sentence structure, and organization when compared to raters for whom the target language is the native language. So, generally, there is only weak evidence of any relationship between rater background and rater performance in language testing.

Only a limited amount of research exists which focuses on rater populations from contexts other than language testing in which raters engage in complex scoring decisions concerning performance assessments, constructed responses, open-ended assessment items, or essays. In the United Kingdom (UK), teachers regularly assign scores to student responses to performance items, particularly writing prompts. Meadows (2006) reviewed research conducted within that context that remained largely unpublished, and she concluded that the criteria used in the UK at that time to select raters (e.g., subject knowledge and teaching experience) were not predictive of rater performance (specifically, reliability of ratings). UK researchers subsequently published three studies that revealed no relationship between a rater's amount of teaching experience, subject knowledge, and scoring experience and the accuracy of the scores that that rater assigns (Leckie & Baird., 2011; Meadows & Billington, 2010; Royal-Dawson & Baird, 2009). So, at least in the UK, there is little evidence that raters' backgrounds can be used to predict the quality of the scores that those raters will assign. These studies focused exclusively on raters in the UK, and they largely focused on writing assessments. It is unclear whether those results generalize to scoring contexts in other national contexts and other assessment content and to what degree rater background may be related to other measures of rater success (e.g.,

achieving qualification status following rater training). This research report addresses the following research questions. *Do general rater background characteristics influence the effectiveness of rater training and the quality of holistic scores assigned by raters in statewide assessment contexts within the United States*, and *is the relationship between rater characteristics and rating quality in these contexts consistent across content domains*?

Method

To address these questions, we conducted secondary data analyses, employing multiple regression procedures, to identify relationships between rater background characteristics and measures of rating quality using data from two large-scale operational testing programs.

Data Sources

Data from two large-scale operational testing programs which employ performance assessments was utilized in this study. The first testing program focused on a norm-referenced writing assessment administered via paper-and-pencil annually to students within a single state in the United States (US). Within this program, scores are assigned to approximately 200,000 students in grades 5 through 12. Students write to one prompt per grade level each year; the prompts vary by administration but can ask for narrative stories or expository or persuasive essays. Three prompts are administered during the school year (two in spring and one in fall), and the prompts vary by administration. Students are allotted two pages on which to write their responses, and those responses are scored using a six-point holistic rubric that takes into account mechanics, ideas, organization, and voice. The second testing program focused on a criterion-referenced science assessment administered annually to approximately 70,000 students in Grades 5, 8, and 12 within a different state in the U.S. Student performance is indexed in terms of

achievement of state science standards, and the test is administered either online or in paper-and-pencil form, although the majority of tests are administered in paper-and-pencil format. Students respond to between four and six operational items. These short-response answers are generally between one-half page and one page in length, and are scored against a four-point holistic rubric that considers accuracy of the scientific content in the student's response.

In both of these scoring projects, raters were required to meet minimum requirements in order to be eligible to score operationally. For both assessments, raters must hold a minimum of a Bachelor's degree in order to be eligible for training. Additionally, the science project required all raters to complete a project-specific science placement test. Once selected, raters were assigned to projects based on geographic location, and rater training was conducted at a regional center for both projects. In addition, raters who had previously scored on one of these two projects were assigned to that same project, when possible. Raters underwent a project-specific training process using item-specific training materials, but the process was similar for the two projects. All training and qualifying materials were compiled by experienced content staff via a rigorous process directed and supervised by both a content specialist and state department of education personnel, who ultimately reviewed and provided feedback and approval of all training materials (i.e., responses, annotations, online modules, and any other project-specific information provided to raters). For the science assessment, raters received synchronous, instructor-led training in a group setting for the review of items, rubrics, using hard-copy materials for rubric, item, and anchor response review prior to an asynchronous review of training modules and assigning practice scores (blindly) to training responses (i.e., exemplar item responses that were already scored by training leaders). For the writing project, only an online asynchronous model of training was utilized. As raters completed the project- and item-specific modules, they then

moved to the task of assigning scores to training responses. On both projects at the end of each training set, raters reviewed the scores they assigned to the training responses alongside the expert scores and annotations explaining those scores.

At the conclusion of training, each rater undertook a qualifying test by assigning scores (blindly) to responses that had been previously assigned consensus scores by training leaders. For the writing project, raters who achieved 70% perfect agreement and 100% adjacent agreement on one of two qualifying sets of ten essays each were permitted to begin operational scoring. For the science project, raters who achieved an average of 80% perfect agreement on two of three qualifying sets of ten student responses each were permitted to begin operational scoring. Raters who did not achieve these standards were released from their respective scoring projects.

During operational scoring, supervisory staff on both projects monitored group and individual rater performance, including validity sets (i.e., blind scoring of expert-scored responses), inter-rater reliability (e.g., percent of inter-rater agreement), and frequency distributions (e.g., percentage of scores falling into each score category). Scoring leaders also provided guidance and feedback to raters in the form of informal comments and discussions that arose from back reading (i.e., student responses that are rescored by scoring leaders). Raters who did not meet project-specific quality metrics received additional feedback in the form of messages, additional back reading from supervisory staff, and calibration responses. Raters who failed to improve and remained below quality thresholds were not allowed to continue to score operational responses and were released from their scoring projects.

Analyses

Measures. Rater background information for both scoring projects was collected from raters using a questionnaire. We selected and recoded six background variables for this study. Some of the available variables (e.g., undergraduate major, the content area you spent most of time teaching) were not included because they contain a large amount of missing data. Table 1 identifies the rater background variables upon which we focused, and these variables served as the predictors in subsequent analyses.

Table 1. Rater Background Variables

Variable	Levels
Ethnicity	Caucasian Other
Gender	Female Male
Age	< 26 26 – 39 40 – 54 > 54
Highest Degree	Bachelor Master Doctor
Teaching Experience	Yes No
Teaching Certificate	Yes No

In addition, scores on operational papers, validity papers, and qualification papers were also collected. Table 2 summarizes the rating performance variables, which served as the

outcome measures in this study. Inter-rater reliability and validity correlation range from −1.00 to +1.00 and indicate the strength of the linear relationship between the scores assigned by two raters and the scores assigned by a rater and the expert scores, respectively, with near-zero values indicating no relationship and values near +1.00 indicating a perfect relationship. Mean inter-rater and validity absolute differences range from 0 to one less than the number of rating scale categories, with values near 0 indicating perfect agreement and values greater than 0 indicating increasing levels of disagreement. Interrater and validity agreement is reported as a percentage of agreement, with values near 1.00 indicating perfect agreement and values near 0.00 indicating no agreement. RPI is a scaled measure that combines reflects several aspects of rater performance utilized by Pearson's Performance Scoring Center.

Table 2. Rater Performance Variables

Variable	Levels
Inter-rater reliability	Correlation between the scores assigned by each rater and scores assigned by a randomly selected second rater
Mean inter-rater absolute difference	Mean absolute difference between the scores assigned by each rater and scores assigned by a randomly selected second rater
Inter-rater agreement	Percent of exact agreement on the scores assigned by each rater and scores assigned by a randomly selected second rater
Validity correlation	Correlation between the scores assigned by each rater and scores assigned by experts on validity papers
Mean validity absolute difference	Mean absolute difference between the scores assigned by each rater and scores assigned by experts on validity papers
Validity agreement	Percent of exact agreement on the scores assigned by each rater and scores assigned by experts on validity papers
RPI	"Rater Performance Index"—A scaled measure that combines several (subjective and quantitative) indicators of rating quality, including speed, scoring leader evaluation, and rater agreement.

Procedures. For each of the seven outcome variables, descriptive statistics of the outcome variable by each level of the predictor variables were computed.

Results

Writing

Rater Characteristics. Table 3 displays the characteristics of the pool of raters in the writing sample. The modal rater was Caucasian, female, over 40 years of age, had earned a bachelor's degree, had teaching experience, but did not hold a teaching certificate.

Table 3. Rater Characteristics for Writing

Variable	Group			
Ethnicity	**Caucasian**	**Other**		
	77%	23%		
Gender	**Female**	**Male**		
	78%	22%		
Age	**< 26**	**26 – 39**	**40 – 54**	**> 54**
	9%	29%	31%	31%
Highest Degree	**Bachelor**	**Master**	**Doctor**	
	57%	38%	5%	
Teaching Experience	**Yes**	**No**		
	70%	30%		
Teaching Certificate	**Yes**	**No**		
	41%	59%		

Note: N = 315 raters. Missing cases = 83 for ethnicity, 22 for gender, 0 for age, 27 for education, 0 for teaching experience, and 56 for teaching certificate.

Group Comparisons. Table 4 compares the performance of ethnicity groups across the seven rater quality indices. Non-Caucasian raters exhibited slightly better rating quality in terms of both inter-rater measures (i.e., higher inter-rater correlation, inter-rater percent of perfect agreement, and lower mean absolute score difference) and validity measures (i.e., higher validity correlation and percent perfect agreement and lower mean validity absolute score difference), as is the overall rater performance index (RPI). As is the case with all comparisons that follow, we conducted multiple regression analyses to determine whether any of these differences are statistically significant. They are not. However, the sample sizes for the two data sets are fairly small for the multiple regression analysis due to missing data on the rater demographic variables (i.e., N = 47 for writing and N = 12 for science), so the power for detecting statistically significant differences is low. Hence, in the remainder of our discussion, we simply provide descriptive comparisons between the groups.

Table 4. Rater Performance by Ethnicity for Writing

Statistic	Caucasian	Other
Inter-rater Correlation	.70	.76
Mean Inter-rater Absolute Score Difference	0.45	0.42
Inter-rater % Perfect Agreement	57%	63%
Validity Correlation	.94	.95
Mean Validity Absolute Score Difference	0.24	0.22
Validity % Perfect Agreement	74%	80%
RPI	1.27	1.36

Table 5 compares the performance of gender groups across the seven rater quality indices. Generally, there are extremely small differences between females and males with respect

to rating quality based on both inter-rater and validity indices. Given that similarity, it is especially interesting to note the relatively large between-group difference on the RPI, again favoring males.

Table 5. Rater Performance by Gender for Writing

Statistic	Female	Male
Inter-rater Correlation	.71	.73
Mean Inter-rater Absolute Score Difference	0.46	0.42
Inter-rater % Perfect Agreement	58%	60%
Validity Correlation	.94	.95
Mean Validity Absolute Score Difference	0.25	0.19
Validity % Perfect Agreement	75%	75%
RPI	1.21	1.43

Table 6 compares the rating quality indices across age groups. Generally, the age groups performed differently across the rating quality indices. For example, although raters over the age of 54 exhibited the lowest average inter-rater and validity correlations, they also produced a high inter-rater percent of perfect agreement. However, it seems that raters in the age 26 to 39 group performed best on most of these rating quality indices while those younger than 26 and older than 54 performed worst. It is also interesting to note that the youngest raters received the highest average RPI.

Table 6. Rater Performance by Age for Writing

Statistic	< 26	26 – 39	40 – 54	> 54
Inter-rater Correlation	.74	.75	.72	.66
Mean Inter-rater Absolute Score Difference	0.48	0.40	0.48	0.47
Inter-rater % Perfect Agreement	51%	62%	56%	58%
Validity Correlation	.94	.95	.94	.93
Mean Validity Absolute Score Difference	0.26	0.21	0.25	0.24
Validity % Perfect Agreement	64%	80%	74%	73%
RPI	1.41	1.39	1.17	1.26

Table 7 compares the rating quality indices across groups defined by the highest degree earned. Overall, raters who had received a doctoral degree produced the highest quality ratings in terms of both inter-rater indices and validity indices. The same is true for the RPI. These differences are quite large in most cases.

Table 7. Rater Performance by Highest Degree for Writing

Statistic	Bachelor	Master	Doctor
Inter-rater Correlation	.71	.72	.78
Mean Inter-rater Absolute Score Difference	0.46	0.46	0.33
Inter-rater % Perfect Agreement	58%	58%	69%
Validity Correlation	.94	.94	.95
Mean Validity Absolute Score Difference	0.24	0.23	0.17
Validity % Perfect Agreement	77%	77%	83%
RPI	1.29	1.20	1.60

Table 8 compares the rating quality indices according to whether raters had teaching experience or not. Generally, those without teaching experience exhibited slightly higher performance on inter-rater, validity indices, and the RPI.

Table 8. Rater Performance by Teaching Experience for Writing

Statistic	Yes	No
Inter-rater Correlation	.71	.74
Mean Inter-rater Absolute Score Difference	0.46	0.41
Inter-rater % Perfect Agreement	57%	61%
Validity Correlation	.94	.95
Mean Validity Absolute Score Difference	0.24	0.20
Validity % Perfect Agreement	75%	76%
RPI	1.19	1.39

Table 9 displays the rating quality indices by teaching certificate status. Generally, raters who had not earned a teaching certificate performed better than did raters who had earned a certificate in terms of inter-rater and validity indices and the RPI. These differences are relatively small on most measures, although the difference in RPI values is a bit larger.

Table 9. Rater Performance by Teaching Certificate for Writing

Statistic	Yes	No
Inter-rater Correlation	.69	.73
Mean Inter-rater Absolute Score Difference	0.47	0.44
Inter-rater % Perfect Agreement	58%	58%
Validity Correlation	.93	.95
Mean Validity Absolute Score Difference	0.27	0.21
Validity % Perfect Agreement	74%	79%
RPI	1.18	1.44

Science

Rater Characteristics. Table 10 displays the characteristics of the pool of raters in the science sample. The modal rater was Caucasian, male, over 54 years of age, had earned a bachelor's degree, had no teaching experience, and did not hold a teaching certificate.

Table 10. Rater Characteristics for Science

Variable	Group		
Ethnicity	**Caucasian**	**Other**	
	77%	23%	
Gender	**Female**	**Male**	
	49%	51%	
Age	**< 40**	**40 – 54**	**> 54**
	27%	25%	48%
Highest Degree	**Bachelor**	**Master**	
	68%	32%	
Teaching Experience	**Yes**	**No**	
	42%	58%	
Teaching Certificate	**Yes**	**No**	
	16%	84%	

Note: N = 79 raters. Missing cases = 15 for ethnicity, 2 for gender, 0 for age, 48 for education, 0 for teaching experience, and 48 for teaching certificate.

Group Comparisons. Table 11 compares the performance of ethnicity groups across the seven rater quality indices. The results are mixed, and any differences between the two groups are small. Overall, there is no evidence in these data of ethnicity differences with respect to rating quality.

Table 11. Rater Performance by Ethnicity for Science

Statistic	Caucasian	Other
Inter-rater Correlation	.81	.79
Mean Inter-rater Absolute Score Difference	0.26	0.28
Inter-rater % Perfect Agreement	74%	73%
Validity Correlation	.97	.98
Mean Validity Absolute Score Difference	0.05	0.04
Validity % Perfect Agreement	94%	94%
RPI	1.03	0.92

Table 12 compares the performance of gender groups across the seven rater quality indices. Generally, there are small differences between females and males with respect to rating quality based on both inter-rater and validity indices as well as the RPI. Any difference tends to indicate a slightly higher quality of the scores assigned by females.

Table 12. Rater Performance by Gender for Science

Statistic	Female	Male
Inter-rater Correlation	.82	.77
Mean Inter-rater Absolute Score Difference	0.26	0.28
Inter-rater % Perfect Agreement	74%	74%
Validity Correlation	.97	.97
Mean Validity Absolute Score Difference	0.05	0.05
Validity % Perfect Agreement	95%	94%
RPI	1.01	1.00

Table 13 displays the rating quality indices across age groups. Generally, the age groups performed very similarly across the rating quality indices, and any differences are not consistent across inter-rater, validity, and RPI indices. That is, there are no large or reliable trends in terms of the relationship between rater age and the quality of assigned ratings.

Table 13. Rater Performance by Age for Science

Statistic	< 40	40 – 54	> 54
Inter-rater Correlation	.83	.80	.80
Mean Inter-rater Absolute Score Difference	0.28	0.27	0.27
Inter-rater % Perfect Agreement	74%	74%	73%
Validity Correlation	.98	.99	.97
Mean Validity Absolute Score Difference	0.05	0.02	0.05
Validity % Perfect Agreement	95%	95%	94%
RPI	1.00	1.01	1.02

Table 14 compares the rating quality indices across groups defined by the highest degree earned. Overall, raters who had received a master degree produced slightly higher quality ratings in terms of inter-rater and validity indices as well as the RPI. However, these differences are very small.

Table 14. Rater Performance by Highest Degree for Science

Statistic	Bachelor	Master
Inter-rater Correlation	.79	.81
Mean Inter-rater Absolute Score Difference	0.27	0.27
Inter-rater % Perfect Agreement	74%	74%
Validity Correlation	.97	.98
Mean Validity Absolute Score Difference	0.06	0.03
Validity % Perfect Agreement	94%	97%
RPI	1.00	1.05

Table 15 compares the rating quality indices according to whether raters had teaching experience or not. Generally, those without teaching experience exhibited a trend of slightly higher performance on the rating quality measures.

Table 15. Rater Performance by Teaching Experience for Science

Statistic	Yes	No
Inter-rater Correlation	.81	.79
Mean Inter-rater Absolute Score Difference	0.27	0.26
Inter-rater % Perfect Agreement	74%	74%
Validity Correlation	.98	.96
Mean Validity Absolute Score Difference	0.07	0.05
Validity % Perfect Agreement	94%	95%
RPI	1.06	0.96

Table 16 displays the rating quality indices by teaching certificate status. Generally,

raters who had not earned a teaching certificate tended to exhibit higher levels of rating quality than did raters who had earned a certificate in terms of inter-rater and validity indices and the RPI. These differences are relatively small on most measures.

Table 16. Rater Performance by Teaching Certificate for Science

Statistic	Yes	No
Inter-rater Correlation	.88	.80
Mean Inter-rater Absolute Score Difference	0.21	0.27
Inter-rater % Perfect Agreement	74%	79%
Validity Correlation	.97	.98
Mean Validity Absolute Score Difference	0.08	0.05
Validity % Perfect Agreement	92%	95%
RPI	1.04	1.03

Discussion & Conclusions

Our study reports the results of rating performance as a function of rater characteristics for raters in two distinct scoring projects—one focusing on writing and the other focusing on science. Overall, we found few large differences between rater groups (Research Question 1), and what differences we did find seemed to be specific to the rating project (Research Question 2). In addition, when we observed a difference, in general, that difference manifested itself across the multiple rating quality indicators in a comparable manner. While several potentially important differences were observed between groups of raters in the writing sample, none of the differences between groups of science raters were large enough to warrant comment. In writing, the largest group differences were observed relating to rater age and highest awarded degree. For example, in both of these groups, we observed difference in inter-rater correlations and percent

perfect agreement of .09 and 11%, respectively. With respect to age, the 26 to 39-years-old group performed best while those under 26-years-old performed worst. With respect to degree, those with doctor degrees performed best. Another small difference was observed relating to ethnicity, with non-Caucasians performing better in terms of the inter-rater correlation (by about .06) and inter-rater and validity agreement percent (about 6% for both). In the writing sample, about 60% of the non-Caucasian raters were Afro-American and about 26% of them were Hispanic, with a small percentage being Native American or Asian. Further, in writing, we observed no large differences between gender and teaching experience or certification status groups. These results are comparable to those in studies of rater background conducted in the UK. That is, we found no evidence that the backgrounds of highly-trained professional raters (raters who are selected based on level of education and content knowledge) is predictive of the quality of the ratings that those raters assign. However, we should note that the small sample sizes in this study precluded conducting conclusive statistical tests of any of these comparisons, so we can offer no actionable recommendations based on these results.

This last statement emphasizes the caution that should be exercised based on the limitations of this study and suggests important points that operational scoring programs should consider when attempting to evaluate rater performance. In our data files, a sizable proportion of the raters had missing data for background variables. For example, 63% of the raters in the writing sample had missing data on at least one of the chosen demographic variables, and 83% of the raters in the science sample had missing data on at least one of the chosen demographic variables. Further, we chose to exclude additional demographic variables due to even more extensive missing data on those characteristics (e.g., college major, scoring experience, etc.). We

emphasize here that analyses such as these would require meticulous data collection strategies that avoid extensive missing data such as this.

Another issue to consider in operational scoring programs is the manner in which data such as these are collected and encoded. There are two important related considerations. First, the questions and their coding should be developed with potential intended uses in mind. For example, collecting open-ended variables (e.g., fill-in-the-blank) makes subsequent use of those variables more difficult. On the other hand, while having a restricted set of options makes subsequent analysis easier, it restricts the detail that can be addressed after-the-fact. Hence, those developing the questions and media through which rater background characteristics are captured must carefully consider the intended uses of such data and the manner in which data collection choices will influence the potential difficulty of using the collected information. Second, data base development and structuring needs to consider the fact that information from multiple sources may need to be subsequently merged. That is, it is important to have universal identifiers for the raters, and rater performance and information must be easily separated according to the specific rating project upon which subsequent analyses are to be performed. For example, if rater performance is collected in a separate database than is rater demographic, a common rater identifier is needed to merge those data. Similarly, if rater performance or demographics is collected separately for multiple scoring projects, then there needs to be a way to differentiate the rater on those multiple projects.

References

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*, 485–505.

Kachchaf, R., & Solano-Flores, G. (2012). Rater language bakcground as a source of measurement error in the testing of English language learners. *Applied Measurement in Education*, *25*, 162–177.

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

Lee, H. K. (2009). Native and nonnative rater behavior in grading Korean students' English essays. *Asia Pacific Education Review*, *10*, 387–397.

Meadows, M. (2006). *Can we predict who will be a reliable marker?* Manchester.

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Manchester.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237–265). Cresskill, NJ: Hampton Press.

Royal-Dawson, L., & Baird, J. A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice*, *28*, 2–8.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*, 27–33.

Weigle, S. C., Boldt, H., Valsecchi, M. I., Elder, C., & Golombek, P. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, *37*, 345–354.

Winke, P., Gass, S., & Myford, C. M. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252.