Language Testing 2009 26 (4) 485—505

The influence of rater language background on writing performance assessment

Jeff S. Johnson *University of Michigan, USA*
Gad S. Lim *University of Michigan, USA**

Language performance assessments typically require human raters, introducing possible error. In international examinations of English proficiency, rater language background is an especially salient factor that needs to be considered. The existence of rater language background-related bias in writing performance assessment is the object of this study. Data for this study are ratings assigned by Michigan English Language Assessment Battery (MELAB) raters to compositions written by examinees of various language backgrounds. While most of the raters are native speakers of English, four have first languages other than English: two Spanish, one Korean, and one bilingual speaker of Filipino and Chinese (Amoy). Examinees were divided into 21 language groups. The IRT application FACETS was used to estimate and control for rater severity when calculating the amount of bias reflected by each rater's set of ratings for each language/language group. Results show that the magnitude of bias terms for all raters for all language groups was minimal, thus having little effect on examinee scores, and that there is no pattern of language-related bias in the ratings.

Keywords: MELAB, multi-faceted Rasch analysis, rater background, rater bias, second language writing assessment

I Introduction

In international tests of English proficiency, performance-based assessment of writing and speaking has come to be the norm. This type of assessment provides the advantage of directly measuring candidates' productive language skills. However, it also requires the

*The authors contributed equally to this paper. They are listed alphabetically.

Address for correspondence: Jeff S. Johnson, University of Michigan, English Language Institute, 500 East Washington Street, Ann Arbor, MI 48104–2028, USA; email: elijsj@umich.edu

DOI:10.1177/0265532209340186

use of raters – most often, human raters – and the introduction of subjectivity into scoring is a potential source of error, which could affect the reliability and validity of these exams (Dunbar, Koretz, & Hoover, 1991). As the results of these assessments are often used for high-stakes purposes (e.g. admission to university, immigration, employment), there is a need to identify the different potential sources and forms of systematic rater error, and if found to exist, to address them accordingly.

The most commonly used measure with regard to raters of performance assessments has been inter-rater reliability, though the desirability of increasing agreement by and of itself has come under question (Lumley & McNamara, 1995; Reed & Cohen, 2001; Weigle, 1998). Inter-rater reliability illumines the product of assessment but not its process. As Connor-Linton (1995) puts it, 'if we do not know what raters are doing ... then we do not know what their ratings mean' (p. 763). The different factors that raters actually consider, as well as background beliefs and predispositions that they subconsciously bring to the rating task, may affect the ratings they give and thus need to be better understood. To that end, studies have investigated such possible sources of variation as the professional background of raters (Barnwell, 1989; Brown, 1995; Shohamy, Gordon, & Kraemer, 1992), their relative severity in rating (Englehard, 1994; Lumley & McNamara, 1995), and the stability of their ratings over time (Congdon & McQueen, 2000).

For international language performance assessments in particular, which measure the language abilities of people with different first languages, the language background of raters is an especially salient factor that should be considered. To some, native speakers (NS) remain the ideal, and non-native speakers (NNS) serving as raters are considered an exceptional category, if not outright unacceptable. On the other hand, there are those who argue that there are testing situations where non-native raters are the most appropriate for evaluating examinee performance (Hill, 1996).

It needs to be pointed out, however, that the notion of a NS is not unambiguous and unproblematic (Davies, 2003), possessing as it does many possible definitions and political implications, a full discussion of which would extend into related literatures such as that of International and World Englishes (Kachru, 1992), and of second language acquisition and ultimate attainment (Birdsong, 1999; Hyltenstam & Abrahamasson, 2003). In this paper, the use of the category NS is not intended as acceptance or endorsement. Rather, it is used merely as a reflection of its ubiquity in the literature (in the

minimal sense of early childhood exposure to an L1) and, indeed, for the precise purpose of investigating the category.

By the same token, NNS do not form a unitary category either; among other things, NNS can vary widely in their L2 proficiency. They can come from cultures that have very different norms regarding communicative events, for example, politeness (Brown, 1995), which can lead to their privileging of particular rhetorical patterns (Kobayashi & Rinnert, 1996) when serving as raters. They can also come from or have lived in parts of the world with well-developed varieties of English, which can differ from a standard dialect in significant ways. It is thus conceivable that non-native raters from different language backgrounds could evaluate language performances differently.

1 Literature review

Studies comparing NS and NNS who rate oral and written language performance differ in their findings. Some indicate that NS are harsher in their evaluations than are NNS (Barnwell, 1989), whereas others find that the opposite is true, that NNS raters are more severe. For example, Fayer & Krasinski (1987) made recordings of Puerto Rican learners of English and played these speech samples for two groups of judges: university students who were native English speakers and Puerto Rican Spanish speakers taking high intermediate-level English classes at the university. They found that while the two groups assigned learners comparable scores, the non-native English speakers were more likely to report being 'annoyed' by particular speech features, especially pronunciation errors and hesitations. In a study of the writing of Chinese learners of English, with language teachers serving as raters, Shi (2001) found that Chinese NNS of English identified more negative features of learners' writing while raters who were NS of English made significantly more positive comments. When the teachers scored the learners, however, an inversion took place: the NS raters who noted more positive features gave lower marks to the learners than did the NNS raters who dwelled on the negative aspects of the writing. Shi attributes this phenomenon to the raters' taking on 'a double role of a strict native speaker and a lenient EFL teacher' (p. 312).

All three studies share a common limitation, however, which is that the raters were not trained on the rating scale. Barnwell had explicitly set out to compare 'naïve' raters vis-à-vis trained ACTFL raters, Fayer and Krasinski used university students as raters, and

the raters in Shi's study appeared to confuse the different roles they were playing. Another limitation is that all three studies provide only group-level comparisons of NS and NNS raters, and relied on small samples of language performance. This makes it difficult to ascertain the appropriateness of scores for individual performances. Raters might indeed focus on different aspects of language performance, but who is to say which raters gave ratings closer to the 'true score'? Conclusions regarding the presence or absence of bias – that is, construct-irrelevant variance affecting scores test takers receive – cannot be drawn.

Other studies have used item response theory-related techniques to address the question of bias while also accounting for other factors. In one such study, Brown (1995) provided NS and high-proficiency NNS, all of whom were language teachers or tour guides, one day of training in the rating of an oral language test. Using the multiple-facet extension of the Rasch model, she found that NNS are harsher than they should be with regards to politeness and pronunciation. She also found that NNS' scoring is more likely to overfit; that is, there is insufficient variability in the ratings they assign. On the other hand, NS tend to be more diverse in their use of rating scales, as well as more diverse in their relative severity. It was theorized that these findings are perhaps the result of NNS hewing more closely to the rating scale as opposed to NS taking a more intuitive approach in rating. Another conjecture was that NS are better at making finer distinctions of ability in their own language (Shi, 2001). In any case, these findings provide additional evidence that factors other than those enumerated in the rating scale affect raters' ratings (Hamp-Lyons, 1989; Vaughan, 1991; Lumley, 2006).

Most of these studies on rater language background have been on speaking, and it appears that quite a few of the findings apply only to that skill area. The features that Fayer and Krasinski's (1987) raters found annoying – pronunciation and hesitation – are features of language performance that are absent in writing. Similarly, politeness, which Brown's (1995) raters treated differentially, is more salient in oral communication than it is in written communication. For that matter, while the literature is replete with references to native *speakers*, there is hardly any mention of native *writers*. Writing is a skill that both NS and NNS come to and learn later in life than speaking, suggesting at the very least the possibility that NS and NNS might not be so different from each other when it comes to writing, and

perhaps in the evaluation of the same. Thus, it is worth asking if there exists a rater language background effect for writing.

Additionally, most studies on rater language background effects thus far have compared NS raters with NNS raters from only one other language. This invites questions about how generalizable any findings are. Elder and Davies (1998) have raised the possibility of a language distance effect on language examination performance. It is theorized, for example, that Japanese is at a greater distance from English than is Spanish. It could thus well be that on an English language examination, the amount of bias for or against Japanese could be larger compared to Spanish. This question can only be answered if multiple L1s are accounted for at the same time.

An attempt to account for both rater language background and language distance effect was made by Hamp-Lyons and Davies (2008), who looked at compositions written for the Michigan English Language Assessment Battery (MELAB), an international examination of English language proficiency. Their sample consisted of 60 compositions written by native speakers of Arabic, Bahasa Indonesia/Malay, Chinese, Japanese, Tamil, and Yoruba. In addition to the official MELAB ratings, given by native English speakers, these compositions were also rated by raters who shared the examinees' L1 and by raters who did not share their L1 – this to see whether there is language background-related bias in the exam and among raters with differing L1 backgrounds. Their study, however, had a number of intervening variables – trained and untrained raters with differing levels of reliability, the use of two different rating scales, and a data set of limited size, among other things – which limited the conclusions that could be drawn regarding language background-related bias.

2 Research questions

While the official MELAB scores in the Hamp-Lyons and Davies study were given by native English speakers, the pool of trained composition raters in the MELAB program actually includes NNS from a few different language backgrounds. It is possible then to avoid a number of confounding factors, such as the use of multiple rating scales and the use of trained and untrained raters, in exploring the questions of interest. This present study uses a much larger sample of compositions written by MELAB examinees, covering an approximate three-year period, and scored by official raters from

different native language backgrounds, to investigate the following research questions:

1) Does the language background of a rater have an effect on the rating of performance assessments in the MELAB writing test?
2) If rater language-background bias does exist, is there an identifiable pattern of interaction between the rater and examinee first languages, that is, a language distance effect?

II Method

1 The test

The MELAB is an advanced-level English proficiency test for adults who use English as a second or foreign language, and who use the scores for university and college admission and for various professional and research purposes. The test includes sections assessing each of the four language skill areas. (For test reviews, see Chalhoub-Deville, 2003; Purpura, 2005; Weigle, 2000.) In the writing section, examinees are given 30 minutes to compose a handwritten composition on one of two prompts. These prompts, which test takers do not see in advance, can call for narrative, expository, or argumentative modes of writing.

Each composition is scored using a holistic, 10-point scale by at least two raters,[1] who are not told the L1 of examinees. If the two ratings are separated by more than one scale-point, a third rater adjudicates. The final score is the average of the ratings that are either equal or different by one scale-point (English Language Institute, 2005). Examinees are allowed to request a rescore if they feel that the score they received is inaccurate; thus, there are potentially up to six ratings for each composition.

2 The examinees

The sample for the present study consists of all examinees who took the MELAB between October 2003 and July 2006, and who claimed a first language or language group spoken by at least 1% of the total MELAB population. (Language group refers to languages that have multiple dialects, e.g., Amoy, Cantonese, Hakka, and Mandarin are

[1] Details of the scale and rubrics can be found at http://www.lsa.umich.edu/eli/testing/melab

Table 1 Examinees' L1s

Language (Group)	n
Chinese	1584
Filipino	1239
Farsi	564
Arabic	496
Korean	425
Russian	370
Spanish	352
Punjabi	336
Urdu	334
English	314
Hindi	207
Malayalam	206
Romanian	195
Japanese	132
Vietnamese	99
Somali	97
Bengali	95
Gujarati	92
German	88
Portuguese	88
Tamil	87

all coded under 'Chinese'.) The above parameters yielded 7400 examinees, representing 21 different language and language groups (Table 1), and for which there were a total of 15,635 ratings.

In some sense, the data used in this paper can be considered an instance of convenience sampling (McMillan & Schumacher, 2001), employing data that just happen to be available, the generalizations from which can be limited if not misleading. But the sample is only a convenience sample if the population is defined as all, similar advanced-level English proficiency tests, and few studies would claim to make such generalizations given that different tests have known differences in test methods, constructs, and contents. On the other hand, the population can be more narrowly defined as all those who take the MELAB. By such a definition, the present sample can very nearly be considered a population sample: after the selection criteria had been followed, the resulting sample constituted 90% of all MELAB tests in the given time period. Thus, while the method of sampling was less than ideal, the present study is still based on a large number of operational test data.

Two more things need to be said about the sample. First, there is a sizable number of English NS who took the MELAB, mostly

for professional certification, for whom the test was not actually designed. As these NS potentially belong to a different population, whose presence in the dataset could affect the accuracy of outcomes, analyses including and excluding English NS were conducted and then compared. Second, some examinees take the MELAB multiple times. An interval of six weeks is required before a person can take the examination again, and a person can only take the MELAB a maximum of four times in any given 12-month period. In the period of time between tests, especially where the interval is longer, it is conceivable that examinees' ability levels could have changed. With this in mind, it would be inappropriate to treat multiple sets of ratings from one examinee as a single case, as it would mean trying to estimate a single ability level, where multiple ability levels might actually be represented. If the opposite were true, that examinees' abilities did not change with time, estimates would not be affected; FACETS would simply provide separate but identical estimates for said examinees. Thus, for the purposes of this study, repeat examinees' sets of ratings were treated as separate cases.

3 The raters

The examinees' compositions were scored by 19 MELAB writing test raters, all of whom were regular employees of the English Language Institute, University of Michigan. All raters went through a standardized training and certification program, and rater performance after certification is continually monitored. Two of the raters had a limited number of ratings ($n < 30$) and were excluded from this study, leaving 17 raters for the analysis. Of the remaining raters, four were men and 13 were women, and their experience rating MELAB writing ranged from five months to over 21 years, with an average of just under five years.

While a majority of raters are NS of American English, four raters had L1s other than English. Two are native Spanish speakers originally from South American countries (raters R03 and R05), one rater is from the Philippines, whose home languages growing up were Chinese (Amoy) and Filipino (Tagalog) (R06), and one is a NS of Korean born in Korea, but who grew up mostly in the United States (R14). Raters R05 and R06 are male, while R03 and R14 are female. All four have native or native-like proficiency in English. In addition, R05 and R06 have experience teaching English at various levels. The number of non-native raters in this study is small, and how representative they are of speakers from their respective first-language

communities is unclear. As such, extreme caution should be exercised in generalizing the findings of this study to other NNS raters.

The distribution of ratings for this study is normal, with a skew of 0.31 and kurtosis of 0.18. All scale points were used, but the lowest rating is underused; where the other nine scale points have FACETS outfit mean squares ranging from 0.8 to 1.2, the lowest scale point has a value of 4.6. For the time period covered by this study, the reported agreement rate – that is, identical or adjacent ratings – was 97.59%, and mean inter-rater Pearson product–moment correlations ranged from 0.81 to 0.88 (Johnson, 2004, 2005, 2006, 2007).

4 Analysis

The IRT program FACETS (Linacre, 2006) was used to model the relationship among the three facets – ratings, raters, and examinee language group. Ratings, which on the 10-point MELAB scale range from 53 to 97 (weighted to match other sections of the test), were converted into a 0 to 9 scale to make the results more easily interpretable. In the results presented in this paper, one point represents the difference between one scale point and the next higher or lower scale point. The model for the study can be expressed in the following way:

$$\log (Pergx / Pergx - 1) = Ae - Sr - Ng - Dx$$

where:

Ae = ability of examinee e

Sr = severity of rater r

Ng = examinee native language group g

Dx = difficulty of category x relative to category x − 1

Pergx = probability of examinee e from language background g receiving a rating of x when rated by rater r

FACETS supplies estimates for each of the three facets on a common logit scale, as well as a bias/interaction report showing observed and expected rating means for each language group. The difference between the two means represent the rater bias, if any, regarding particular language groups. Also presented in the results are descriptive statistics showing the spread of each facet, which can reveal differences in rater severity, as well as fit statistics, which provide an indication of rater consistency.

The data were run through FACETS two times: one time including examinees who are NS of English in the analysis, and another time

excluding this group of examinees. This was done to see if the presence of the English NS group affected the accuracy and fit of the model.

III Results and discussion

1 Comparing models

The results of the analysis provide some support to the notion that the NS of English belong to a different population than NNS examinees (Table 2). They have a higher mean rating than all other language groups, and are separated from the next highest language group by almost seven-tenths of a band. The practical implication

Table 2 Estimates with and without English NS

	With English NS		Difference in estimates (With English – Without English)			
	Observed score	Fair score	Fair score	Measure (Logit)	Infit Mn Sq	Outfit Mn Sq
English	6.6	5.30				
German	5.9	5.16	0.05	0.10	0.0	0.0
Hindi	5.5	4.99	0.03	0.15	0.0	0.0
Urdu	5.4	4.99	0.04	0.15	0.0	0.0
Bengali	5.2	4.92	0.03	0.15	0.0	0.0
Romanian	5.2	4.90	0.03	0.16	–0.1	0.0
Gujarati	5.0	4.86	0.03	0.16	–0.2	–0.2
Tamil	5.2	4.86	0.04	0.13	0.2	0.2
Portuguese	5.1	4.84	0.03	0.15	–0.1	0.0
Spanish	5.0	4.84	0.04	0.11	0.0	0.0
Russian	5.0	4.79	0.04	0.13	0.0	0.0
Punjabi	4.8	4.74	0.05	0.10	–0.3	–0.2
Filipino	4.8	4.73	0.04	0.15	0.3	0.2
Farsi	4.7	4.67	0.06	0.08	–0.1	0.0
Vietnamese	4.7	4.63	0.07	0.05	0.0	0.0
Chinese	4.5	4.49	0.05	0.08	0.0	0.0
Malayalam	4.5	4.46	0.04	0.14	0.0	0.0
Arabic	4.4	4.45	0.05	0.10	0.0	0.0
Japanese	4.3	4.41	0.02	0.22	–0.1	0.0
Korean	4.1	4.36	0.06	0.07	0.0	0.0
Somali	4.1	4.33	0.04	0.09	0.0	0.0
Mean			0.07	0.00	0.0	0.0

Separation (with English): 10.00, reliability: 0.99
Separation (without English): 8.76, reliability: 0.99
Fixed (all same) chi-square (with English): 3424.3, df = 20, $p < .00$
Fixed (all same) chi-square (without English): 2209.4, df = 19, $p < .00$

of including these English NS, it appears, is that in the attempt to fit a line through the data, it indicates a much lower fair rating estimate for English. Conversely, it yields marginally higher fair rating estimates and logit measures for the other languages than would be the case if English were not included (columns 4 and 5). An alternate explanation would be that the fair rating estimates are indeed accurate, and that MELAB raters have just been systematically too lenient – by 1.3 bands – when rating compositions written by NS of English. Knowledge and intuition about the data would indicate that the latter explanation is not very likely.

While the estimates might vary to some extent when NS of English are included and when they are excluded, in practical terms, the differences are negligible. Fit changes for a small number of language groups, but remains the same for most of them. The signal to noise ratio between measured variance among languages and measurement error is robust in both cases, as shown by the separation index. (The only potentially important difference between the two models is with regard to one NNS rater (R05), whose infit measure falls on different sides of acceptability in the two models. Reference to this particular result will be made in an appropriate section of this paper.) On the whole, because the differences between the two models are minor, and including English NS provides information for one more language group, the analyses proceeded using that model. However, it should be kept in mind that results relating to English NS examinees need to be interpreted with caution, and that adjustments must be made in the appropriate directions.

2 Rater severity and fit

To provide an overview of rating behavior, the measurement report for raters is given in Table 3, ordered by severity. Rater severity is measured in logits, centered around zero, where positive numbers indicate harshness and negative numbers indicate leniency. The report shows that most raters clustered closely around the mean, indicating that no rater was especially severe or lenient. R11 was the only rater whose severity was more than two standard deviations from the mean. The four non-native raters were all quite moderate in their rating, all falling within the middle of the group and within one standard deviation of the mean.

The infit and outfit mean square residuals provide an indication of rater consistency. The former is weighted towards expected responses, while the latter statistic is unweighted and is more sensitive to extreme ratings. Raters whose fit statistics are much higher than

Table 3 Rater severity and fit

Rater	Ratings	Severity (logit)	Model error	Infit MnSq	Outfit MnSq	Non-English L1
R11	152	1.62	0.17	1.1	1.1	
R15	1115	0.75	0.06	0.7	0.6	
R08	1341	0.64	0.06	0.8	0.7	
R14	1506	0.52	0.05	1.1	1.1	Korean
R06	2627	0.32	0.04	0.8	0.7	Filipino, Chinese
R07	1749	0.16	0.05	1.0	0.9	
R13	811	0.13	0.07	1.3	1.2	
R02	88	0.12	0.22	1.0	1.0	
R01	3612	0.12	0.03	1.0	0.9	
R16	587	–0.03	0.09	1.1	1.0	
R09	599	–0.18	0.09	0.8	0.8	
R05	120	–0.27	0.18	1.5	1.5	Spanish
R03	47	–0.60	0.31	1.0	0.9	Spanish
R04	665	–0.63	0.08	1.1	1.0	
R12	131	–0.65	0.18	1.0	0.9	
R17	126	–0.70	0.19	0.9	0.8	
R10	235	–1.32	0.14	0.9	1.0	
Mean	912.4	0.00	0.12	1.0	0.9	
SD	1003.2	0.67	0.08	0.2	0.2	

Separation: 4.62; Reliability: 0.96
Fixed (all same) chi-square: 562.5, df = 16, $p < .00$

the expected value of 1.0 rate inconsistently and unpredictably (i.e., ratings exhibit too much variation), while those with values far below 1.0 are too consistent and do not distinguish between different performances (i.e., ratings exhibit too little variation). There is no fixed cutoff or rule for which infit values are too high or too low. Linacre (2002) suggests mean square values between 0.5 and 1.5 are practically useful, while McNamara (1996) proposes values within two standard deviations from the mean as a guideline, which in this case would yield a somewhat more stringent infit acceptable value range of 0.6 to 1.4, and 0.5 to 1.3 for outfit.

From Table 3, it appears that rater R05 is the only one who might not fit the model, rating somewhat inconsistently with a 1.5 for both infit and outfit. These fit values, while outside the acceptable range by McNamara's standard, are considered acceptable by Linacre's guidelines. This is in fact the rater who, in the model that excludes English NS, has mean square residuals (1.2 for both infit and outfit) that fall safely within both standards. These differing fit statistics for R05 suggest that this rater is right at the bounds of acceptable fit, perhaps

confused by or just having problems rating compositions by English NS in particular. Overall, the fit statistics suggest that raters have been consistent in their ratings, and are using all parts of the scale.

3 Language background effect

Bias terms, the difference between expected and observed ratings, were measured for each rater for each first language for which data were present. This resulted in 313 measured bias terms, 47 of which were statistically significant ($|z\text{-score}| > 1.96$). Some of these significant terms, however, involve small n-sizes – six involve less than five cases, and a total of 15 involve less than 10. Thus, it is uncertain whether these terms represent real bias towards the language group indicated or only towards the particular individuals encountered by the rater. It stands to reason that a rater should have read a certain number of compositions from a particular language group before the observed bias, whether for or against, can meaningfully be called bias towards that language group. For the purposes of this paper, that minimum number was arbitrarily and conservatively set at five compositions, and only bias terms involving five or more ratings are reported.

It should also be kept in mind that not all statistically significant outcomes are substantive ones. The magnitude of the bias becomes substantive when the difference between observed and expected mean ratings is large enough to affect results. In this case, one point being the difference between one scale point and another, bias would have to be greater than 0.5 for a majority of ratings to be affected in either direction. Where the values are less than 0.5, bias would be present, but examinees in the end are still more likely to receive a rating the model indicates they should get.

Tables 4 to 7 present significant bias terms for each of the four raters with first languages other than English, as well as bias terms for their native languages, whether or not these terms meet the minimum n or are statistically significant. On the whole, these raters do not show significant bias for or against compositions written by

Table 4 Significant and native language bias terms for R03 (L1 = Spanish; n = 47)

Language group	n	Mean difference	Z-score	Infit MnSq
Spanish	4	0.08	–0.35	0.8

Table 5 Significant and native language bias terms for R05 (L1 = Spanish; n = 120)

Language group	n	Mean difference	Z-score	Infit MnSq
English	14	–0.67	4.82	1.7
Filipino	7	–0.42	2.32	1.7
Urdu	6	0.45	–2.32	1.4
Spanish	9	0.09	–0.53	0.6

Table 6 Significant and native language bias terms for R06 (L1 = Chinese, Filipino; n = 2627)

Language group	n	Mean difference	Z-score	Infit MnSq
Romanian	78	–0.15	2.63	1.0
Russian	136	–0.09	2.17	0.7
Chinese	552	0.01	–0.61	0.8
Filipino	398	0.01	–0.53	0.7

Table 7 Significant and native language bias terms for R14 (L1 = Korean; n = 1506)

Language group	n	Mean difference	Z-score	Infit MnSq
Chinese	326	0.06	–2.44	1.2
Farsi	100	0.15	–3.01	1.3
Filipino	249	–0.06	2.14	1.1
German	23	–0.51	4.96	2.1
Portuguese	16	0.25	–2.12	0.9
Korean	105	0.10	–2.20	1.0

people who share their L1s, other than R14, who showed a slight significant bias for compositions written by Korean L1 examinees (observed – expected mean = 0.10, Table 7). In addition, R14 also showed some bias for compositions written by Chinese, Farsi, and Portuguese speakers, and some bias against Filipino speakers. While R14 apparently has substantial bias against German speakers, the infit mean square of 2.1 is more than two standard deviations from the mean, indicating that this bias is not consistently observed or measured, and should thus not be considered as an example of bias. R05, whose L1 is Spanish, showed substantial bias against speakers of English (Table 5). As English L1 examinees' abilities are already underestimated by the model, the real value of this bias term will

be larger than the −0.67 indicated. Keeping in mind that R05's fit statistics moved from borderline acceptable to clearly acceptable after English NS were removed from the model, it appears that R05's problem lies mainly in rating compositions written by NS of English. R05's bias for speakers of Urdu and against Filipino speakers was moderate, and R05 showed no significant bias for or against any other language group. R06, the Chinese- and Filipino-L1 rater (Table 6), showed small but significant bias against speakers of Romanian (−0.15) and Russian (−0.09). Rater R03, whose L1 is Spanish, read a very small number of compositions (n = 47, Table 3), which in part accounts for the absence of significant bias terms (Table 4).

Substantial bias terms for all raters, and bias terms larger than 0.25, are shown in Table 8. Three bias terms are equal to or larger than 0.5, though the one involving bias for English (R04), which was measured at 0.54, is known to be an overestimate. As such, there are two substantial bias terms, for which none has a mean difference between observed and expected ratings larger than one point. (The largest difference is 0.68, belonging to rater R17, in favor of Urdu speakers.)

Table 8 Bias terms, by language group

For			Against	
≥0.50	≥0.25		≥0.25	≥0.50
		Arabic	R04	
	R04	Bengali		
		Chinese		
R04		English		R05*
		Farsi		
		Filipino	R05*	
	R01	German		
		Gujarati		
		Hindi		
	R10	Japanese		
		Korean		
		Malayalam		
	R14*	Portuguese		
		Punjabi		
	R11	Romanian		
		Russian		
	R08	Somali	R09	
		Spanish		
		Tamil		
R17	R05*	Urdu		
		Vietnamese	R09	

*NNS rater

Of the bias terms that are 0.5 or higher, one belongs to an English L1 rater and one to a non-English L1 rater. No rater in the study had more than one substantial bias term. Given the small number of significant bias terms, it would be difficult to interpret these findings as forming patterns that indicate differences in rating behavior between native and non-native English-speaking raters in this study.

It can also be seen from Table 8 that there are more bias terms in favor of examinees than against examinees; that is, errors are more likely to be Type I, where examinees receive higher scores than they deserve, and less likely to be Type II, where examinees undeservedly get lower scores.

IV Conclusion

This analysis indicates that there is no discernible pattern of language background-related bias in the ratings for this large set of MELAB writing tests. There were few significant bias terms, and the magnitudes of these were mostly insubstantial. Keeping in mind the scoring procedures in which multiple raters read each composition and aberrant ratings are not included in the computation of examinees' final scores, the analysis provides evidence that ratings in this performance assessment of writing are on the whole accurate, reliable, and fair.

The main question this study addresses is the influence of rater language background on ratings of writing performance assessments. From this study, it appears that it is possible for a small number of native-like NNS of English from three different language backgrounds to be trained to be just as effective in rating writing performance as are their NS counterparts. Other than one non-native rater showing slight bias in favor of examinees who share her L1, these raters have rated examinees, whether sharing their L1 or not, without substantive language-related bias.

Admittedly, as has been previously noted, the number of non-native raters in this study is small, and they come from only a few L1s, so it is unknown whether the findings would hold for raters from other L1 backgrounds, or even for others from the same L1 backgrounds. While it has been shown that these NNS raters can rate accurately in this context, it is not necessarily the case that other NNS raters can in other contexts. In addition, the non-native raters in this study were all highly proficient in English, and it is worth asking if

NNS at a lower level of proficiency can perform in the same way. The raters in Fayer and Krasinski's (1987) study, for example, were at the high-intermediate level of proficiency. Thus, is there a minimum level of language proficiency required for non-native raters to become indistinguishable from NS raters? And is amount of training a factor in determining this minimum level and in the preparation of raters at different proficiency levels? Further, if NS and NNS raters can become indistinguishable from one another, native status would need not be a category used for determining who can serve as raters. In which case, what are those qualities and characteristics that make for good raters? These questions all have implications for rater recruitment, training, and deployment, and are directions for future research to take.

The previous finding that non-native raters are more likely to overfit (Brown, 1995) was not sustained in the current study. Non-native raters were found all across the fit statistics distribution. Also, contrary to previous studies, non-native raters as a group were not found to be more or less harsh that NS raters. The difference in findings between this and previous studies can perhaps be attributed, as has been hypothesized in this paper, to the different objects of their study: performance assessments of writing and speaking, respectively. The features that native and non-native raters noted differently in previous studies – pronunciation, hesitation, and politeness – either do not apply or are not as salient in writing. Speaking, and the rating of it, is also more linear in time; responses are needed almost instantaneously for conversation to proceed smoothly, leaving little time for planning. On the other hand, when writing, examinees have more time to plan, edit, and revise their performance. In the same way, raters can also review writing performance more easily – re-reading parts of the composition if they feel the need to – before deciding on what ratings to give. In speaking, unless the performances are recorded, the opportunity for such review is not available.

It has been hypothesized that NNS raters could rate performance assessments differently because they could come from places with well-developed varieties of English, which might cause them to overlook or accept features that are unacceptable in a standard dialect. (In the current study, only the Chinese- and Filipino-L1 rater comes from a country with a clearly developed variety of English.) Another reason given, which applies more generally to all raters, is that the influence of examinee L1s that are far different from the raters' own

L1s might affect their judgment of said examinees' performances. These L1-related reasons, if actually present, could potentially show up only in discrete and very particular ways, for instance, lexical choice. Most parts of the examinees' responses would still exhibit features consistent with a Standard English. Thus, it could well be that the L1-related features actually *did* influence raters' judgments, except that Standard features were more pervasive and overwhelming, reducing the effect of L1-related features on the final score. The MELAB uses holistic scoring, where all factors and considerations are in some way averaged and subsumed under a single final score. Could it be that if analytic scoring were used, where different components are identified and separately rated, and where greater nuance is required, that rater language-background effects might be measured and detected in the components where variations and differences reside? This is also a question for future research to address.

In any event, NS and NNS raters rating consistently and identically is one goal, but this is subsidiary to the more important consideration of both groups rating appropriately. This is something that IRT analysis alone cannot answer. It is recommended that verbal protocol analysis studies such as those done with NS raters (Vaughan, 1991; Cumming, Kantor, & Powers, 2001; Lumley, 2006) be extended to include NNS raters, and to investigate the way they approach the task of rating.

The second research question, on a possible language distance effect, depended on significant findings to the first question. If there was no meaningful bias in the first place, then there can be no differing magnitudes of bias of which to speak. The present study found few substantive bias terms, which would make it appear that there is no discernible language distance effect. It must be remembered, however, that the study included only a limited number of non-native raters, and more raters at different hypothesized distances from English would be needed to conclusively address that question.

In closing, this study provides evidence that rater language background can be minimized and made a non-factor in the scoring of writing performance assessments. Certain native and non-native raters can be trained to use a rubric so that the rating behavior of one group cannot be distinguished from the other. Which people can and cannot be trained for the task remains an open question. Where research on rater effects is concerned, the study highlights the need to distinguish between the rating of oral and written language performance. Evidence suggests that findings which apply to one do

not necessarily extend or apply to the other. More careful research into each, separately, can together shed more light on and add to the validity of language performance assessments.

Acknowledgements

The authors thank India Plough and Diane Larsen-Freeman, who read earlier drafts of this paper, and the anonymous Language Testing reviewers, whose comments have improved this article. A version of this paper was presented at the 2007 American Association for Applied Linguistics annual conference in Costa Mesa, California.

V References

Barnwell, D. (1989). 'Naïve' native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*, 152–163.

Birdsong, D. (Ed.) (1999). *Second language acquisition and the critical period hypothesis*. Mahwah, NJ: Lawrence Erlbaum.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*, 1–15.

Chalhoub-Deville, M. (2003). Fundamentals of ESL admissions tests: MELAB, IELTS, and TOEFL. In D. Douglas (Ed.), *English language testing in US colleges and universities*. 2nd ed. (pp. 11–35). Washington, DC: NAFSA.

Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, 163–178.

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

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. TOEFL Monograph Series, MS-22. Princeton, NJ: Educational Testing Service.

Davies, A. (2003). *The native speaker: Myth and reality.* Clevedon, UK: Multilingual Matters.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289–303.

Elder, C. & Davies, A. (1998). Performance on ESL examinations: Is there a language distance effect? *Language and Education*, *12*, 1–17.

Englehard, G. (1994). Examining rater errors in the assessment of written compositions with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

English Language Institute, University of Michigan (2005). *Michigan English language assessment battery: Technical manual 2003*. Ann Arbor, MI: English Language Institute, University of Michigan.

Fayer, J. M. & Krasinski, E. (1987). Native and nonnative judgements of intelligibility and irritation. *Language Learning*, *37*, 313–326.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 229–244). Tubingen: Gunter Narr.

Hamp-Lyons, L. & Davies, A. (2008). The Englishes of English tests: Bias revisited. *World Englishes*, *27*, 26–39.

Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing*, *5*, 29–50.

Hyltenstam, K. & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 539–588). Oxford: Blackwell.

Johnson, J. S. (2004). *2003 MELAB data analysis: Descriptive statistics and reliability estimates*. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2005). *MELAB 2004: Descriptive statistics and reliability estimates*. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2006). *MELAB 2005: Descriptive statistics and reliability estimates*. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2007). *MELAB 2006: Descriptive statistics and reliability estimates*. Ann Arbor, MI: English Language Institute, University of Michigan.

Kachru, B. B. (Ed.) (1992). *The other tongue: English across cultures*. 2nd ed. Urbana, IL: University of Illinois Press.

Kobayashi, H. & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning*, *46*, 397–437.

Linacre, J. M. (2002). What do infit, outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2006). *Facets Rasch measurement computer program*. Chicago: Winsteps.com.

Lumley, T. (2006). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

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

McMillan, J. H. & Schumacher, S. (2001). *Research in education: A conceptual introduction*. 5th ed. New York: Longman.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Purpura, J. E. (2005). Michigan English language assessment battery (MELAB). In S. Stoynoff & C. A. Chapelle (Eds.), *ESOL tests and testing* (pp. 87–91). Alexandria, VA: TESOL.

Reed, D. J. & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin, (Eds.), *Experimenting with uncertainty: Language testing essays in honor of Alan Davies* (pp. 82–96). Cambridge: Cambridge University Press.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303–325.

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.

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* (pp. 111–125). Norwood, NJ: Ablex.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287.

Weigle, S. C. (2000). Test review: The Michigan English language assessment battery (MELAB). *Language Testing*, *17*, 449–455.