# Scoring Behavior of Native vs. Non-native Speaker Raters of Writing Exams

Ah-Young Kim
(Seoul National University)

Kristen di Gennaro
(Pace University)

**Kim, Ah-Young and Gennaro, di Kristen. (2012). Scoring Behavior of Native vs. Non-native Speaker Raters of Writing Exams.** ***Language Research*** **48.2, 319-342.**

In performance testing, where judges provide scores on examinees' abilities, program administrators may seek to ensure that raters are consistent in their interpretations of the scoring criteria. A Rasch analysis not only allows us to examine differences in rater consistency, but also to check for interactions between rater characteristics and scoring behavior. In this paper, we present the results of an analysis of rater severity in assessing students' writing ability. Our main focus was to examine differences between native speaker (NS) and non-native speaker (NNS) raters. Results showed that raters differed in terms of severity, with NNS raters as a group more severe than NS raters. In addition, the severity of NNS raters varied more than that of the NS raters. Bias analysis indicated some rater-examinee bias and rater-domain bias in both NS raters and NNS raters, with the majority found among the NNS raters. The paper provides implications for training NNS raters. (153 words)

**Keywords:** assessing writing ability, scoring behavior, rater severity

## I. INTRODUCTION

In many academic programs where writing ability is important, examinees are required to produce a sample of writing. Such direct tests of writing are considered a type of performance assessment since they "elicit... a performance or behavior other than simple indication of choice" and "this performance or behavior is then judged or rated, by

means of a scale, ... introduc[ing] a new type of interaction, that between the rater and the scale" (McNamara 1996, p. 121). Though many testing programs have introduced procedures for minimizing the subjectivity of scoring that is necessarily a part of human judgment, such as sophisticated scoring rubrics and repeated rater training, human raters are not "scoring machines" (Linacre 2005, p. 4), and thus will always introduce a degree of variability into the scoring procedures of direct tests. Traditionally, this variability has been considered a problem for test users, with strong agreement across raters considered more desirable (Weigle 1998). The strength of rater agreement, usually described in terms of the correlation between scores from different raters, is presented as an inter-rater reliability coefficient (Brown 2005); thus, a test could be said to have good reliability if the correlation coefficient is high.

As a result of the preference for high agreement in rater judgments, efforts are often made to reduce differences in rater behavior (Lumley & McNamara 1995). More recently, however, some researchers have drawn attention to the limitations of this definition of reliability for writing tests (Hamp-Lyons & Kroll 1997). In fact, it has been suggested that variability across raters is actually desirable, since variability is necessary for determining probabilistic scores (Weigle 1998), such as those provided in a Rasch analysis. Rather than attempt the impossible, and perhaps undesirable, feat of reducing rater variability to zero, administrators of testing programs may employ a Rasch approach for multi-faceted measurement of latent traits, such as the FACETS program developed by Linacre (2004), which allows the influence of raters to be reflected in examinees' scores such that an examinee who has been evaluated by a "harsh" rater and another examinee who has been evaluated by a "lenient" rater will have their scores adjusted to compensate for these rater variables. The same can be done by taking into account differences in item or prompt difficulty. Thus, an immediate benefit of the FACETS program is its capacity to produce examinee scores that account for the severity of raters and the difficulty of test items. An additional benefit, however, is the program's potential

to inform test administrators as to how various components in their testing procedures are functioning. Such uses highlight the potential to use FACETS as a research tool (Hamp-Lyons & Kroll 1997, Lumley & McNamara 1995). For instance, program administrators might be interested in evaluating the difficulty of test prompts or the effectiveness of test items in separating examinees into distinct levels of ability. In a performance testing situation, where judges provide scores on examinees' abilities, program administrators may want to ensure that raters are consistent in their interpretations of the scoring criteria. In addition, some researchers have noted the potential for rater differences due to such things as rater background and experience (Shohamy, Gordon & Kraemer 1992, Weigle 1998). The FACETS program not only allows us to examine differences in rater severity and consistency, but also to check for interactions between certain rater characteristics and raters' scoring behavior. It is precisely this capacity for evaluating rater behavior that we explore in this paper.

One study that highlights both the substantive and methodological capabilities of FACETS in language testing, specifically in relation to rater variability, is that by Lumley and McNamara (1995). In this study, the researchers used FACETS to see if rater behavior interacted with rating occasion; that is, they looked for systematic patterns, or bias, within raters depending on when the rating took place. Lumley and McNamara found that certain raters did, indeed, exhibit bias, defined as "significant changes in their severity" (p. 68), across different rating occasions. The researchers interpreted these results as further support for using a program such as FACETS to analyze performance data, since rater training often does not erase inter-rater variability nor account for intra-rater variability over time.

Weigle (1998) also used FACETS to examine rater behavior, but rather than analyze the interaction effects between raters and rating occasion, this study looked at potential bias as a result of a rater's background. Specifically, Weigle compared the severity and consistency of ratings from both experienced and inexperienced raters in order to see if there was an interaction between degree of rater experience and

rater behavior. Weigle found that inexperienced raters were both more severe and inconsistent in the scores they assigned to essay tests. Differences in severity of scoring remained even after training, but training did improve rater consistency, leading Weigle to conclude that rater training is more beneficial for improving intra-rater reliability than inter-rater reliability.

In an earlier study, where Rasch procedures were used to examine the relationship between raters and essay type (personal experience expository writing or narration), Hake (1986) found that raters tended to "misgrade" pure narrative essays more than expository essays that incorporated narration, leading her to conclude that pure narrations were less likely to be "objectively graded" (p. 160). In other words, there appeared to be an interaction between raters and the type of essay they were rating. Based on the limited information provided in this study, it is difficult to determine if the misfit Hake found was due more to rater behavior or to prompt differences, but they do indicate an interaction effect between raters and prompts. Despite some limitations of this study, Hake makes the important point that as long as "we concentrate on trying to make raters alike rather than attempting to identify and account for their differences" (1986, p. 162), we ignore interesting research questions about rater variables, such as degree of tolerance for "flaws" in writing exams and differences in raters from different disciplines or even specializations within the same discipline.

In a recent study of rater differences, Chalhoub-Deville and Wigglesworth (2005) hypothesized that different groups of native speakers (i.e., from Australia, Canada, the UK, and the US) would rate examinees on the Test of Spoken English differently. While this hypothesis was not borne out in their findings, this study raises the point that future research should examine such rater differences in more detail, especially given the preponderance of English instruction and examinee evaluation in different countries, not only by different types of native speakers but also by non-native speakers.

Therefore, the current study investigated the differences in rater behavior in relation to whether raters were native speakers (NS) or non-

native speakers (NNS) of English by using Rasch analysis. There is some pre-existing research on NS versus NNS teachers' judgment in assessing speaking performance (Kim 2009, Zhang & Elder 2010). Kim (2009) found NS teachers were more detailed than NNS teachers in rating pronunciation, specific grammar use, and the accuracy of transferred information. Comparatively less research has been conducted in comparing NS and NNS rater behavior in writing assessment. In one of the few studies, Johnson and Lim (2009) used Rasch analysis and found no significant difference in rating behavior between NS and NNS rater groups. In detail, 7,400 writing samples were rated by 17 raters, four of which were NNSs. Rasch analysis findings indicated that raters' language background did not make much difference in their rating performance. However, Johnson and Lim's study examined the behavior of only four NNSs, a rather small number of NNS raters. Thus, more NNS raters' rating behavior was examined in this study.

Specifically, this study examined the rater behavior of ESL teachers on scoring written essays of new incoming students to an adult ESL program. The ESL program was managed by the Teaching English to Speakers of Other Languages (TESOL) graduate program at a university, located in New York City. The ESL courses were taught by TESOL graduate students as a partial requirement for their degree. Due to the characteristics of the TESOL program, around half of the ESL teachers were highly proficient non-native speakers of English, while the other half were native speakers. Thus, it deemed necessary to explore the differences between the two groups of raters, if any.

This study addressed the following research questions:

1. How do raters of the ESL writing exam differ in severity of scoring?
2. Do NNS raters and NS raters interact differently with examinees?
3. Do NNS raters and NS raters interact differently with certain categories of the rating scale?

## II. Method

### 2.1. Participants

#### 2.1.1. Examinees

For the purpose of this study, writing tests from 100 randomly selected incoming adult ESL students in the Spring 2005 semester were examined. The majority of the ESL students were immigrants, but a large number were international students, executives, or spouses of the international population residing in New York City and New Jersey. Participants' writing proficiency was expected to vary considerably, from low- to high-proficiency, as they had diverse educational and linguistic backgrounds. About a third of the students were Asian and more than half of the examinees were female. Examinee age varied from early twenties to mid-fifties.

#### 2.1.2. Raters

There were 17 raters in the study, all of whom were ESL teachers in the Spring 2005 semester. They were teaching in the ESL course as a partial requirement for their master's degree in TESOL. Most of the raters were novice teachers with minimal prior experience in teaching or rating essays. Out of the 17 raters, eight raters were NSs of English and the other nine were NNSs of English, with high proficiency in English. NNSs had received a Test of English as a Foreign Language (TOEFL) score above 102, indicating high-proficiency, to enroll into the TESOL master's program. The NNS raters were from Brazil (n=1), Korea (n=3), Poland (n=1), Taiwan (n=3), and Mexico (n=1). The majority of the NNSs had East Asian backgrounds (n=6). The age of the raters ranged from mid-twenties to early thirties and the majority of the raters were female (n=13).

### 2.2. Materials

#### 2.2.1. Writing Prompt

The writing section consisted of one extended production task. The writing section was part of a longer placement exam, consisting of

five sections: listening, speaking, grammar, reading, and writing. The writing test, which included one writing prompt, constituted one-fifth of an examinee's total placement score. Examinees were given 25 minutes to write an essay response to a prompt, requiring students to compare and contrast the concepts of cooperation and competition in different educational settings (see Appendix A).

2.2.2. Rating Scale

To score the essays, raters used an analytic rubric with the following four domains: content (the extent to which the response fulfills the writing task and addresses the topic), organization (the extent to which the response is coherent and cohesive), grammar (accuracy, complexity, appropriateness of language), and vocabulary (range and appropriateness of lexis) (see Appendix B). The scale for each domain ranged between 1 to 6 points with 6 being the highest possible score. Thus, the highest score an examinee could receive for the writing section was 24 points.

2.3. Procedures for Data Collection

The five parts of the ESL placement test were administered over the course of two consecutive days about two weeks before the beginning of the semester. The writing section was administered on the first day of the exam, after the listening, grammar and reading subtests. Examinees were required to write their essays by hand and in pencil in 25 minutes. Those who finished early were allowed to leave early.

One week after the administration of the ESL placement exam, the essays were rated by ESL teachers. All teachers for the semester were required to participate in the rating procedure. Raters were first trained through a norming session using a norming packet including samples from previous administrations. The hour-long norming was led by an experienced rater and doctoral student in the TESOL program. During the training, raters were required to read a sample essay and were encouraged to discuss why they gave certain scores to the essay. Afterwards, the trainer provided a normed score and the raters compared their scores with this normed score. This procedure was repeated several times until the raters understood the rating process well.

Immediately following the norming session, raters began rating the essays. Each essay was independently rated by two raters based on an analytic scoring rubric. After this individual scoring for each essay, the total scores from the two raters were averaged. The full duration of the rating process, excluding the norming session, was about two hours. Connectivity between the raters was maintained across the data set − a prerequisite for using a Rasch measurement model.

### 2.4. Procedures for Data Analysis

In Linacre's (2004) FACETS program, each aspect of the testing situation is considered a facet. In the present study, three facets were included: examinees, raters, and domain (writing ability). The examinee facet included 100 examinees. There were a total of 17 raters in the second facet. The domain facet was subdivided into four categories used to score writing ability: grammar, vocabulary, organization, and content. Since we expected the subcategories of writing ability to vary in difficulty, we modeled the categories of writing ability as partial credit items, allowing each category to have its own step structure. The three-facet Rasch model used in this analysis can be expressed as:

$$\log (P_{nijk}/P_{nijk-1}) = B_n - C_j - D_i - F_{ik}$$

$P_{nijk}$ = probability that examinee *n* on item *i* is rated by judge *j* with a score of *k*
$P_{nijk-1}$ = probability that examinee *n* on item *i* is rated by judge *j* with a score of *k-1*
$B_n$ = the ability of the examinee *n*
$C_j$ = the severity of rater *j*
$D_i$ = the difficulty of item *i*
$F_{ik}$ = the difficulty of achieving a score within a particular score category *k* on a particular item *i*

# III. RESULTS

## 3.1. FACETS Summary

One of the most attractive features of the FACETS program is the visual output summary in the form of a map in which all facets are clearly displayed in relation to one another and along the same linear measure. Figure 1 is the map produced for this study. We will describe each feature of this map in detail. First, it is important to understand that in order to create a standard frame of reference, only one facet is non-centered, with the mean allowed to vary, and all other facets are set with means at zero. The facet specified as non-centered is the facet of interest, or object of measurement. In this case, since we were interested in rater behavior, the object of measurement was the raters; therefore, the mean for the rater facet was non-centered, allowing the mean for raters to vary, and the examinee and item facets were each set to zero. It should be noted that setting the examinee facet to zero does not change the configuration of the scores, but does affect the logits associated with each score.

The first column on the left in Figure 1, which displays +8 at the top and -8 at the bottom, is the logit scale used to measure all facets. The second column shows the examinee results, with each asterisk representing one examinee. We can see from this column that when the mean is set to zero, scores ranged from a high score of 7.12 to a low score of -7.70, producing a spread of 15.61 logits, but that only a few examinees received scores at these two extremes. In fact, the map shows that most examinees scored in the middle of the total range. This confirmed our expectation of great variability in scores due to the differences in examinee background and experience.

The third column, and the one of interest for this paper, displays the raters' behavior in relation to one another and to the other facets in the analysis. The more severe raters, those who tended to give lower essay scores, are at the top, while the more lenient raters are at the bottom of the scale. This map indicates that of the 17 raters included in this analysis, Rater 8 was the most severe, at 2.93 logits, while Rater 9 was the most lenient, at -.79 logits, producing a spread of 3.72 logits.

Column four shows the overall difficulty of each category of each

aspect of writing ability as defined by the rubric adopted by the ESL program. According to this analysis, each category was of similar difficulty, with grammar and vocabulary slightly more difficult than organization, and content slightly easier than the other categories. Since we specified a partial credit model for each category, we can see that step structure is similar, but not identical, for each category, as represented in the final four columns.

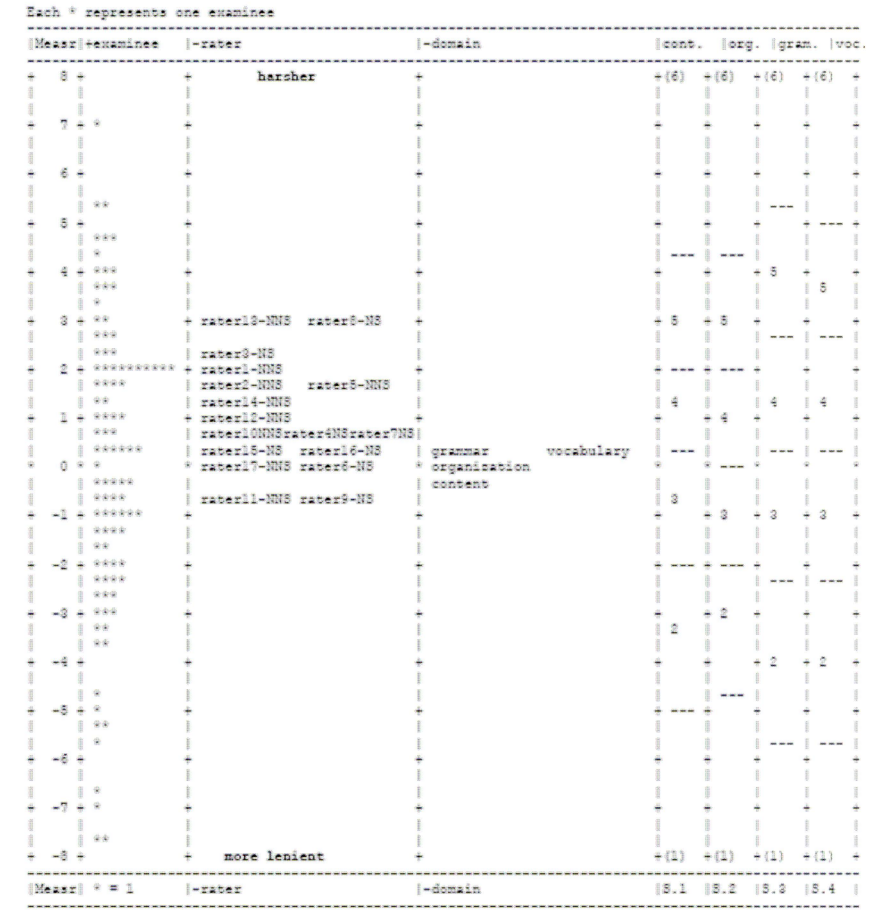


**Figure 1.** FACETS Summary: Examinee Ability, Rater Severity, Category Difficulty, and Domain Scales.

## 3.2. Rater Behavior Analysis

### 3.2.1. Descriptive Statistics of Rater Severity by Group

We will now discuss rater behavior in detail, as this is the focus of the study. First, Table 1 displays the descriptive statistics for rater severity per group: all raters, NS raters, and NNS raters. The higher the logit mean, the more severe those raters were as a group.

**Table 1.** Descriptive Statistics of Rater Severity by Group (in logits)

| | All Raters (n=17) | NS Raters (n=8) | NNS Raters (n=9) |
|---|---|---|---|
| Maximum (in logits) | 2.93 | 2.93 | 2.91 |
| Minimum | -.79 | -.79 | -.56 |
| Range | 3.72 | 3.72 | 3.47 |
| Mean Severity | 1.01 | .82 | 1.18 |
| SD | 1.10 | 1.18 | .99 |

Since the NS raters were at the extreme ends of the scale, the maximum and minimum scores are the same for both of these groups, 2.93 and -.79, respectively, creating a range of 3.72 logits. The range for NNS raters, however, is slightly smaller, since the maximum and minimum scores were less extreme, at 2.91 and -.56. The mean severity for all raters was 1.01 logits, but only .82 logits for the NS rater group. The NNS group severity mean was highest, at 1.18 logits. Thus, even though the maximum severity is lower for NNS raters, the mean severity is highest of all for this group. This indicates that the NNS group was more severe than the NS group.

### 3.2.2. Individual Rater Severity

Rater behavior can be analyzed not only in terms of raters' relative severity, but also by the consistency within each individual rater (intra-rater reliability). Table 2 presents the raters in descending order of severity measured in logits. The two groups of raters, NSs versus NNSs, are also distinguished in the table. Summary statistics for all raters, NS raters, and NNS raters are provided under the table.

The data in Table 2 indicate that Rater 8 was the most severe, with a measure of 2.93 logits and Rater 9 was the most lenient, at -.79 logits. Interestingly, NNS raters appeared to be more severe in their ratings compared to the NS raters, as there are more NNS raters at

the top of the severity scale than NS raters. In fact, of the eight most severe raters, six of them were NNS raters. This reflects the overall means of rater severity per group shown previously in Table 1, where the NNS mean severity was higher than the NS mean. It should be noted that although the NNS raters tended to be harsh graders, the NS raters were more extreme in their ratings.

**Table 2.** Rater Severity and Fit (N=17)

| Rater ID | Rater severity measure (in logits) | Standard error | Infit Mnsq |
|---|---|---|---|
| 8 (NS) | 2.93 | .24 | 1.0 |
| 13 (NNS) | 2.91 | .21 | 1.4 |
| 3 (NS) | 2.48 | .70 | .0 |
| 1 (NNS) | 1.99 | .33 | 1.3 |
| 5 (NNS) | 1.68 | .19 | .7 |
| 2 (NNS) | 1.59 | .21 | 1.0 |
| 14 (NNS) | 1.33 | .20 | 1.2 |
| 12 (NNS) | 1.09 | .22 | 1.4 |
| 7 (NS) | .64 | .30 | 1.7 |
| 10 (NNS) | .60 | .08 | .7 |
| 4 (NS) | .59 | .21 | 1.1 |
| 15 (NS) | .38 | .30 | 1.1 |
| 16 (NS) | .37 | .22 | .9 |
| 17 (NNS) | .01 | .21 | 1.7 |
| 6 (NS) | -.06 | .20 | .8 |
| 11 (NNS) | -.56 | .26 | 1.2 |
| 9 (NS) | -.79 | .20 | .8 |

*All raters:*
RMSE (Model) .28 Adj S.D. 1.04 Separation 3.72 Reliability .93
Fixed (all same) chi-square: 370.0 d.f.: 16 significance: .00

*Native raters:*
RMSE (Model) .34 Adj S.D. 1.13 Separation 3.35 Reliability .92
Fixed (all same) chi-square: 157.1 d.f.: 7 significance: .00

*Non-native raters:*
RMSE (Model) .22 Adj S.D. .97 Separation 4.40 Reliability .95
Fixed (all same) chi-square: 203.2 d.f.: 8 significance: .00

The separation index of the raters is an indication of the degree of difference among the severity of the raters. The larger the separation index, the greater the differences in severity across raters. A separation index close to 0 would indicate that the raters were rating with similar degrees of severity. For these data, the separation index for the overall group of raters was 3.72, which was quite high. This implies that the raters were very different in terms of individual severity in rating the essays. The reliability index, which ranges from 0 to 1, indicates how reliably different raters behaved; that is, how likely they were to provide different ratings. (This is almost the exact opposite of the traditional correlation coefficient used to determine how similar raters' scores are.) The reliability index value for these raters was .93, meaning that the separation index value was indeed reliable. In addition, the chi-square value of all raters of 370 (df=16) was significant at p=.00. Therefore, the null hypothesis that all raters were equally severe must be rejected. Each of these statistics confirms that these raters differed in their degree of severity.

A comparison of the range of severities of the two groups of raters reveals that the severity of the NNS raters varied more than that of the NS raters: the separation index of NNSs was higher, at 4.40 logits, than the separation index of the NSs, at 3.35 logits.

Overall, this analysis of rater severity indicates that the NS raters and the NNS raters showed group differences in the severity of their ratings: NNS raters were generally more severe than the NS raters. Moreover, the severity of the NNS raters varied more than that of the NS raters.

### 3.2.3. Rater Consistency

Rater consistency (intra-rater reliability) can be determined by analyzing the fit statistics. This statistic indicates how consistently a rater used the scoring scale across examinees. The last column in Table 2 shows the infit mean-square values for each of the raters. Values that are not within the range of two standard deviations around the mean (1.1 +/- [.4*2]) are considered misfitting values (McNamara, 1996). Thus, for these data, any mean square value below .3 or above 1.9 was considered inconsistent in rating. Rater 3, a NS, who had an infit mean-square value of 0, was determined as the rater with the most inconsistent rating. The rating scores of this rater were too predictable

and did not contain enough variation.

Infit values can also be used to measure group differences in rating. Infit mean-square values have an expected mean of 0 and a standard deviation of 1. As can be seen from Table 2, the mean and standard deviation were .9 and .4, respectively, for the NS group, and 1.2 and .3 for the NNS group. Both the mean and the standard deviation values were similar in the two groups. This implies that there was not much difference in consistency between the two groups.

### 3.3. Bias Analysis

The FACETS program also allows one to check for bias, or unexpected interactions, across various facets in the testing situation. The amount of bias is reported in logits, with its significance reported as a standard z-score (Linacre 2005). For this study, a bias analysis was performed to determine if there was any bias due to the interaction between raters and other facets. Two types of bias analysis were performed: rater-examinee bias and rater-domain bias.

#### 3.3.1. Rater-Examinee Bias

A bias analysis was performed regarding the interaction between the raters and examinees. This identifies raters whose responses appeared to form consistent patterns for certain examinees − patterns that were different from their rating pattern of other examinees, and different from other raters. For these data, there were 17 cases of unusual rater-examinee interaction out of a total 296 interactions. Table 3 presents these bias interactions between raters and examinees according to the z-score measure in the bias report.

A z-score higher than 2.0 in Table 3 implies that the examinee is being rated systematically more severely by the rater in question. A z-score lower than -2.0 indicates that the examinee in question is systematically rated more leniently by the rater. For example, the z-score between Examinee 13 and Rater 12 was lower than expected at -2.23, which indicates that Rater 12 systematically rated Examinee 13 more leniently than normal. At the same time, Rater 4 systematically rated Examinee 13 more severely than normal, which produced a high z-score value of 2.08.

Of the 17 instances of examinee-rater interaction, 14 were produced

**Table 3.** Rater-Examinee Bias (arranged according to fit)

| Observed score | Expected score | Observed count | Obs-exp ave | Bias measure | Model S.E. | Z-score | Infit Mnsq | Outfit Mnsq | Sq | Nu | Exam | Measure | Nu | Rater | | Measure |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 14.9 | 4 | 1.03 | -2.13 | .75 | -2.82 | .3 | .3 | 83 | 71 | 71 | 3.60 | 8 | 8 | NS | 2.93 |
| 21 | 17.3 | 4 | .93 | -2.19 | .84 | -2.62 | 1.2 | 1.1 | 236 | 32 | 32 | 4.77 | 13 | 13 | NNS | 2.91 |
| 8 | 5.3 | 4 | .67 | -2.28 | .88 | -2.60 | 1.5 | 1.6 | 237 | 33 | 33 | -2.92 | 13 | 13 | NNS | 2.91 |
| 18 | 14.3 | 4 | .91 | -1.85 | .73 | -2.53 | .2 | .2 | 155 | 50 | 50 | 1.01 | 10 | 10 | NNS | .60 |
| 23 | 19.6 | 4 | .85 | -2.69 | 1.15 | -2.35 | .6 | .5 | 284 | 84 | 84 | 3.16 | 17 | 17 | NNS | .01 |
| 19 | 15.6 | 4 | .85 | -1.77 | .75 | -2.34 | .8 | .8 | 254 | 55 | 55 | 2.36 | 14 | 14 | NNS | 1.33 |
| 22 | 18.9 | 4 | .78 | -2.08 | .92 | -2.26 | .4 | .4 | 202 | 97 | 97 | 3.34 | 10 | 10 | NNS | .60 |
| 23 | 19.8 | 4 | .80 | -2.56 | 1.15 | -2.23 | .9 | .7 | 221 | 13 | 13 | 4.36 | 12 | 12 | NNS | 1.09 |
| 20 | 16.9 | 4 | .78 | -1.73 | .79 | -2.19 | .1 | .1 | 240 | 36 | 36 | 4.58 | 13 | 13 | NNS | 2.91 |
| 17 | 14.0 | 4 | .75 | -1.50 | .71 | -2.11 | .2 | .2 | 42 | 28 | 28 | 1.91 | 5 | 5 | NNS | 1.68 |
| 5 | 7.5 | 4 | -.64 | 2.32 | 1.15 | 2.03 | .8 | .7 | 114 | 9 | 9 | -3.29 | 10 | 10 | NNS | .60 |
| 18 | 20.6 | 4 | -.65 | 1.52 | .73 | 2.08 | .7 | .7 | 31 | 13 | 13 | 4.36 | 4 | 4 | NS | .59 |
| 17 | 19.7 | 4 | -.68 | 1.50 | .71 | 2.11 | .9 | .9 | 274 | 71 | 71 | 3.60 | 16 | 16 | NS | .37 |
| 9 | 12.1 | 4 | -.78 | 1.98 | .84 | 2.36 | .4 | .4 | 230 | 26 | 26 | 2.09 | 13 | 13 | NNS | 2.91 |
| 10 | 13.6 | 4 | -.89 | 2.13 | .81 | 2.64 | .5 | .6 | 53 | 40 | 40 | 1.68 | 5 | 5 | NNS | 1.68 |
| 18 | 22.1 | 4 | -1.02 | 2.62 | .73 | 3.59 | .2 | .2 | 163 | 58 | 58 | 5.48 | 10 | 10 | NNS | .60 |
| 13 | 18.6 | 4 | -1.39 | 2.86 | .75 | 3.83 | .3 | .3 | 189 | 84 | 84 | 3.16 | 10 | 10 | NNS | .60 |
| Observed score | Expected score | Observed count | Obs-exp ave | Bias measure | Model S.E. | Z-score | Infit Mnsq | Outfit Mnsq | Sq | Nu | Exam | Measure | Nu | Rater | | Measure |
| 12.6 | 12.6 | 4.0 | .00 | .01 | .89 | .00 | .6 | .6 | | | Mean (Count: 296) | | | | | |
| 5.0 | 4.8 | .0 | .36 | .94 | .37 | 1.11 | .7 | .7 | | | S. D. | | | | | |

Fixed (all = 0) chi-square: 362.2; d.f.: 296; significance: .01

by NNS raters. Nine of these were the result of systematic lenient rating, whereas five were due to more severe ratings. A further analysis revealed that the 17 interactions involved 14 examinees and that 10 of these were of East-Asian background (e.g., Chinese, Japanese, and Korean). Considering that most NNS raters also were of East-Asian backgrounds, the examinee-rater bias suggests some systematic interaction between the Asian examinees and Asian raters.

3.3.2. Rater-Domain Bias

A bias analysis was also performed regarding the interaction between the raters and the domains of the rating rubric. This identifies raters who responded consistently to certain domains differently from their own rating pattern of other domains, and differently from other raters. Thirteen significant bias instances were found out of the 68 possible interactions (17 raters * 4 domains). Table 4 is a summary of all significant biased interactions and is ordered according to the z-score measure in the bias report. The upper portion of the table shows the most lenient raters, and the lower portion includes the harshest raters.

A z-score higher than 2.0 in Table 4 implies that the domain was rated systematically more severely than normal by the rater in question. A z-score lower than -2.0 indicates the domain was systematically rated more leniently by the rater. For example, the interaction between Rater 1 and the domain of content produced a significant bias value of -2.75, which indicates that the rater displayed more lenient rating behavior than normal when rating content. Moreover, the same rater demonstrated a more severe than normal rating pattern when grading the grammar domain.

Of the 13 bias interactions found between raters and domains, the majority of cases (ten) involved NNS raters. Although the NS raters were not immune from bias, the NNS raters produced nearly three times more cases of bias. Of the ten NNS bias interactions, seven had z-score values over 2.0. That is, the NNS raters generally portrayed a more severe rating pattern than normal, which makes sense given that the NNS raters had a tendency to be overall more severe in their ratings. The high number of bias interactions among the NNS raters implies that they were not using the rating rubric as was expected.

The domains of content and grammar both produced the most (n=4)

**Table 4.** Rater-Domain Bias

| Observed score | Expected score | Observed count | Obs-exp ave | Bias measure | Model S.E. | Z-score | Infit Mnsq | Outfit Mnsq | Sq | Nu | Rater | | Measure | Nu | Rater | Measure |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 16.3 | 6 | -.72 | -1.94 | .71 | -2.75 | 1.5 | 1.4 | 1 | 1 | 1 | NNS | 1.99 | 1 | Content | -.43 |
| 28 | 32.9 | 8 | -.61 | -1.15 | .49 | -2.36 | .8 | .8 | 28 | 11 | 11 | NNS | -.56 | 2 | Organization | .08 |
| 56 | 61.8 | 15 | -.38 | -.89 | .39 | -2.29 | 1.4 | 1.3 | 4 | 4 | 4 | NS | .59 | 1 | Content | -.43 |
| 40 | 45.6 | 16 | -.35 | -.93 | .41 | -2.25 | 1.9 | 1.5 | 34 | 17 | 17 | NNS | .01 | 2 | Organization | .08 |
| 17 | 20.4 | 7 | -.48 | -1.39 | .66 | -2.12 | 1.0 | 1.0 | 41 | 7 | 7 | NS | .64 | 3 | grammar | .17 |
| 50 | 45.7 | 13 | .33 | .94 | .47 | 2.01 | 1.8 | 1.9 | 46 | 12 | 12 | NNS | 1.09 | 3 | Grammar | .17 |
| 60 | 54.4 | 17 | .33 | .75 | .36 | 2.07 | .7 | .6 | 5 | 5 | 5 | NNS | 1.68 | 1 | Content | -.43 |
| 19 | 15.8 | 6 | .53 | 1.40 | .65 | 2.16 | .5 | .5 | 35 | 1 | 1 | NNS | 1.99 | 3 | Grammar | .17 |
| 35 | 30.9 | 8 | .51 | 1.18 | .53 | 2.21 | .3 | .3 | 62 | 11 | 11 | NNS | -.56 | 4 | Vocabulary | .18 |
| 51 | 45.7 | 16 | .33 | .98 | .42 | 2.32 | 1.0 | 1.0 | 51 | 17 | 17 | NNS | .01 | 3 | Grammar | .17 |
| 340 | 323.5 | 100 | .16 | .39 | .15 | 2.53 | .7 | .7 | 27 | 10 | 10 | NNS | .60 | 2 | Organization | .08 |
| 26 | 21.4 | 7 | .66 | 1.50 | .57 | 2.64 | 1.5 | 1.5 | 7 | 7 | 7 | NS | .64 | 1 | Content | -.43 |
| 50 | 43.7 | 14 | .45 | 1.15 | .42 | 2.71 | .3 | .3 | 53 | 2 | 2 | NNS | 1.59 | 4 | vocabulary | .18 |
| Observed score | Expected score | Observed count | Obs-exp ave | Bias size | Model S.E. | Z-score | Infit Mnsq | Outfit Mnsq | Sq | Nu | Rater | | Measure | Nu | Rater | measure |
| 55.0 | 55.0 | 17.4 | .00 | .00 | .50 | .00 | .9 | .9 | Mean (Count: 68) | | | | | | | |
| 69.4 | 69.2 | 21.1 | .26 | .66 | .25 | 1.38 | .5 | .5 | S. D. (populn) | | | | | | | |
| 69.9 | 69.8 | 21.3 | .27 | .67 | .26 | 1.39 | .5 | .5 | S. D. (Sample) | | | | | | | |

Fixed (all = 0) chi-square: 130.3 d.f.: 68 significance (probability): .00

bias interactions, whereas vocabulary (n=2) caused least bias. This implies that raters had most difficulty in agreeing with the descriptors for content and grammar domains. This finding indicates a need for clarification of the rating rubric and increased training in those domains during the rater norming session.

## IV. DISCUSSION & CONCLUSION

This study examined how NS and NNS raters differed in terms of severity in scoring ESL essays. Findings indicated that the two rater groups did, indeed, differ in terms of severity. In particular, we found that NNS raters, as a group, were more severe than NS raters. Similar results were found in Kim's (2009) study on rating speaking; but the current findings differed from Johnson and Lim's (2009) study, which found no difference between NS vs. NNS raters in rating writing samples. Such contrasting findings suggest a need for more research on the topic of rater rating severity between NSs and NNSs.

In terms of the bias between raters and examinees, we found 17 cases of bias, of which 14 involved NNS raters. Nine cases exhibited more lenient ratings, and five were more severe. Concerning the students in these bias interactions, the 17 interactions involved 14 different students, ten of whom were of East-Asian background. This is interesting because most of the NNS raters were also of East-Asian background, suggesting that there may be characteristics of East-Asian writers that NNS raters are able to identify, perhaps subconsciously, which influenced their rating behavior with these writers. In the future, in-depth qualitative research should be conducted to examine such details.

We found less bias between raters and domains, but again, the majority of the cases (ten out of 13) involved NNS raters. Unlike the bias between raters and examinees, where most cases showed more lenient behavior, the bias between raters and the scoring rubric tended to exhibit more severe rating behavior. These cases of bias suggest there may be a need for further training so that raters can better understand how to apply each part of the scoring rubric.

Overall, this study confirms the need for greater investigation into individual rater characteristics and how these characteristics may inter-

act with various facets in a performance assessment situation. Especially in cases where examinees may be rated by either NS or NNS raters, efforts should be made to ensure that rater differences do not benefit or disadvantage certain examinees. This can be done either through the use of a Rasch analysis, which allows examinees' scores to compensate for rater differences, or through enhanced rater training to encourage greater intra-rater consistency. Since NNS were found to be more severe on grammar and content domains they should receive more detailed training on these areas to prevent severe leniency or harshness in their ratings. It should be noted that the majority of raters in this study were novice raters; thus, rating experience was not considered as an additional characteristic. Yet, in the future, more rater characteristic such as teaching or rating experience should be included as a facet of Rasch analysis in order to provide more in-depth findings.

## References

Brown, J. D. (2005). *Testing in language programs.* New York: McGraw Hill.

Chalhoub-Deville, M. and Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes* 24.3, 383-391.

Hake, R. (1986). How do we judge what they write? In K. L. Greenberg, H. S. Weiner, and R. A. Donovan, eds., *Writing assessment: Issues and strategies*, pp. 153-167. NY: Longman.

Hamp-Lyons, L. and Kroll, B. (1997). *TOEFL 2000-writing: Composition, community, and assessment* (TOEFL Monograph Series, MS 5). Princeton, NJ: Educational Testing.

Johnson, J. S. and Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing* 26.4, 485-505.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing* 26.2, 187-217.

Linacre, J. M. (2004). *Facets Rasch measurement computer program*. Chicago: Winsteps.com

Linacre, J. M. (2005). *A user's guide to FACETS*. Chicago: Winsteps.com.

Lumley, T. and McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing* 12.1, 54-71.

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

Myford, C. M. and Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report No. 65). Princeton, NJ: Educational Testing Service.

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

Smith, R. M. (2003). *Rasch measurement models: Interpreting WINSTEPS/ BIGSTEPS and FACETS output.* Maple Grove, MN: JAM press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15.2, 263-287.

Zhang, Y. and Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing* 28.1, 31-50.

## APPENDIX A.

# Writing Prompt

*Directions: You will have 25 minutes to write a well-organized essay on the following topic. Before you begin writing, consider carefully and plan what you will write. Make sure you proofread your essay. Write your essay below.*

Topic: Most people think that American schools encourage both cooperation and competition. What about education in your country? Which is considered more important, cooperation or competition? Use specific reasons and examples to support your answer.

**APPENDIX B.**

# Scoring Rubric for the Writing Test

| Domain | Level | Criteria |
|---|---|---|
| Content | 6 Excellent | The essay fulfills the writing task and addresses the topic richly and fully. |
| | 5 Strong | The essay fulfills the writing task and addresses the topic appropriately. |
| | 4 Good | The essay generally fulfills the writing task and addresses the topic adequately. |
| | 3 Acceptable but limited | The essay only partially fulfills the writing task and addresses part of the topic. |
| | 2 Weak | The essay shows a vague or incomplete understanding of the writing task and topic. |
| | 1 Seriously flawed | The essay suggests a lack of understanding of the writing task and topic. |
| Organization | 6 Excellent | Coherence between paragraphs, sentences, and ideas is successfully achieved through a variety of methods. |
| | 5 Strong | Coherence is adequate and achieved through the use of transitional words and phrases. |
| | 4 Good | The essay is generally well organized; there may be a limited lack of coherence and difficulty with paragraphing. |
| | 3 Acceptable but limited | Paragraphs often lack coherence; sentences are not well connected; paragraphs are not appropriately connected to each other. |
| | 2 Weak | There is little coherence within and across paragraphs; sentences are strung together somewhat haphazardly; there is little or no clear attempt to connect paragraphs. |
| | 1 Seriously flawed | There is no attempt to divide the essay into conceptual paragraphs, or the paragraphs are unrelated. |

| Domain | Level | Criteria |
|---|---|---|
| Grammar | 6 Excellent | Except for rare minor errors, the writer's use of grammar is precise and does not obscure meaning. |
| | 5 Strong | Effective and appropriate use of grammar in general; minor errors in articles, agreements, verb forms, and no incomplete sentences; meaning is never obscured. |
| | 4 Good | Competent control of grammar in general; there may be errors in article use and verb agreement and several errors in verb form; Errors affect clarity occasionally, but generally do not obscure meaning. |
| | 3 Acceptable but limited | Limited competence in use of grammar; occasional major errors or frequent minor errors in grammar occasionally hinder comprehension. |
| | 2 Weak | Frequent errors in verb formation, articles, and incomplete sentences; sentence structures make comprehension difficult. |
| | 1 Seriously flawed | Numerous problems in all areas of grammar; sentence construction is so poor that sentences are often incomprehensible. |
| Vocabulary | 6 Excellent | There is a wide range of appropriately used vocabulary; there are few problems with word choice. |
| | 5 Strong | Vocabulary is broad and is generally used appropriately. |
| | 4 Good | Vocabulary is generally adequate, but may sometimes be inappropriately used. |
| | 3 Acceptable but limited | Vocabulary shows some flexibility, but may be inaccurate or repetitive in places. |
| | 2 Weak | Vocabulary is quite basic and word choice is inaccurate. The writer may rely on repeating words and expressions from the prompt. |

| | | |
|---|---|---|
| | 1 Seriously flawed | Vocabulary is extremely restricted and repetitively used, and more sophisticated attempts at word choice are often inaccurate or inappropriate. |

Ah-Young Kim (Co-Author)
TEPS Center, Language Education Institute
Seoul National University
1 Gwanak-gu Gwanak-ro, Seoul 151-742, Korea
Email: akim@snu.ac.kr

Kristen di Gennaro (Co-Author)
Pace University
One Pace Plaza
New York, NY 10038, U.S.A.
Email: kdigennaro@pace.edu

Received : June 17, 2012
Revised version received: July 13, 2012
Accepted: July 29, 2012