Studies in Applied Linguistics & TESOL at Teachers College, Columbia University, Vol. 23, No. 1, pp. 37–65
Forum

Writing Task Performance and First Language Background on an ESL Placement Exam: A Many-Facets Rasch Analysis of Facet Main Effects and Differential Facet Functioning

Daniel Eskin
Teachers College, Columbia University

ABSTRACT

First Language (L1) has been assumed to play a role in Second Language ability (Bachman & Palmer, 1996). However, the interplay between them across skill, task, or scoring criteria is more complex (Hamp-Lyons & Davies, 2008). Using Many-Facets Rasch Measurement, this study investigates the main effects of examinee ability, rater severity, task difficulty, and rubric scale difficulty and functionality on the writing section of an English as a Second Language program's placement test, then compares performance among L1 Spanish and Japanese examinees to discern the presence of bias across facet. The results for examinee ability and rater severity suggest score variability not expected by the model. Regarding task difficulty and scale difficulty and functionality, it can be concluded that an argumentative essay genre was more difficult than a customer review, or that rater assessed rubric criteria for Content, Organization, and Language more harshly for the former. A bias analysis among L1 Spanish and Japanese examinees revealed that the customer review displayed a bias against Japanese examinees, particularly for Organization, while the argumentative essay displayed bias for them, particularly for Organization and Language. These results demonstrate how placement testing could inform curricula in language programs with linguistically diverse student populations.

INTRODUCTION

Over the last several decades, Second Language (L2) testing has relied on performance assessment (e.g., written essays, oral interviews) to evaluate practical command of language acquired. These methods entail more complex task design and subjective human scoring compared to multiple choice items (Bachman & Palmer, 1996; Carr, 2011). Scored performance in this form of assessment involves numerous facets of measurement, including the *raters* and *rubric criteria* (i.e., 'scales') for scoring performance, the *instruments* (i.e., tasks) for eliciting performance, and *personal characteristics* of the examinee (e.g., first language, L1, background) producing the *written or spoken sample* (Eckes, 2019; McNamara, 1996).

Given the importance of each facet, the question becomes how the effects of each can be investigated empirically. To that end, a number of psychometric models have been utilized to investigate performance assessments (e.g., Bachman et al., 1995; Lynch & McNamara, 1998). One such model, *Many-Facet Rasch Measurement* (MFRM) (e.g., Linacre, 2019), allows for probabilistic inferences to be made about examinee ability across an entire sample, and among particular examinees, based on scored performance relative to the effects of rater severity (Eckes, 2005), task difficulty (Bonk & Ockey, 2003; Myford & Wolfe, 2000), and scale difficulty (Grabowski, 2013) in the measurement process.

Using MFRM, this study will investigate the effects of examinee ability relative to rater severity, task and scale difficulty, and scale functionality on the writing section of an *English as a Second Language* (ESL) program's placement test, then will compare performance among examinees of two L1 backgrounds, Spanish and Japanese, to discern the presence of bias for or against a group across raters, tasks, and scales. An initial literature review will highlight how MFRM has been used for empirical inquiry in L2 writing assessment.

LITERATURE REVIEW

Second Language Writing Assessment and Many-Facet Rasch Measurement

The construct of writing has been characterized in many ways (Weigle, 2003; Cumming et al., 2021). As an overarching trend, Cumming et al. (2021) note that certain component abilities of writing have played prominently in past conceptualizations, notably *micro-processes* (e.g., word choice, sentence construction), *composing processes* (e.g., planning, drafting, revising), and *macro-processes* (e.g., fulfilling genre conventions, asserting a coherent perspective, expressing membership in a discourse community)" (p. 108). To measure L2 writing ability, performance tasks (e.g., argumentative essay, complaint letter, data summary) are scored using a variety of criteria, prioritizing certain components (McNamara, 1996).

In the context of writing assessment, L2 testing literature has documented the effects of rater variability (e.g. severity) through MFRM analyses (Elder et al., 2006; Weigle, 1998). Studies regarding the influence of rater-mediated scores on task difficulty (e.g., Pollitt & Hutchinson, 1987) and scale difficulty (McNamara, 1990, 1996) have also used this model. Such investigations examining the relative impact of each facet on scored performance are known as a *"Main Effects"* Analysis (Eckes, 2019). Further inquiries on the extent that a given facet *differentially* influences particular examinees (e.g., by gender, See Eckes, 2005) or particular raters (e.g., by L1 background, see Johnson & Lim, 2009) has been referred to as *"Bias Analysis"* or *"Interaction Analysis"* (Eckes, 2019).

Certain capabilities of MFRM prove useful for conducting such analyses in L2 Writing Assessment. First, MFRM distinguishes between its *Rating Scale Model* (RSM), allowing for the examination of scale functionality averaged across raters, tasks, and rubric scales in the sample, and its *Partial Credit Model* (PCM), allowing users to examine finer-grained differences among individual scales, tasks, raters, or even examinees (Choi, 2019). Second, MFRM, an extension on another model, *Item Response Theory* (IRT) (Ockey, 2021), applied to dichotomously scored items with two 'parameters', examinee ability and item difficulty (e.g., Fan & Bond, 2019), can deploy a grouping variable or, *dummy facet*, to investigate differences among subsets of a sample (e.g., based on demographic characteristics of sample) across elements of another facet (e.g.,

scales of analytic writing rubric, Di Gennaro, 2009). This analysis is referred to as *Differential Item Functioning* (DIF) in the context of dichotomously scored data (Raquel, 2019) and *Differential Facet Functioning* (DFF) in the context of performance assessment (Eckes, 2005, 2019). With these capabilities in mind, we will now consider the role of L1 Background in L2 writing performance and whether differential performance can be expected between L1 groups.

Second Language Writing Ability and First Language Background

As a "personal characteristic" of examinees, L1 Background, has been assumed to play a role in L2 ability (e.g., Bachman & Palmer, 1996). However, the extent to which L1 influences L2 ability across skill (e.g., writing), task (e.g., argumentation, summary), or scale sub-construct (e.g., content, grammar) is a more complex question. Linguistic and cultural bias has been observed in performance on proficiency exams of L2 English (Chen & Henning, 1985) and other languages (e.g., Gujord, 2022), routinely in differences across L1 and nationality (e.g., *IELTS Demographic Data 2021*, n.d.). Inevitably, such differences gloss over a litany of confounds (e.g., education, profession) influencing performance (Bigelow & Watson, 2013).

Among productive skills, differential examinee performance across L1 background is perhaps more acute in L2 speaking, where assessing sub-constructs such as pronunciation and intelligibility could potentially be influenced by examinee and rater L1 background (e.g., Shin, 2022; Yan et al., 2019). By contrast, the construct of L2 writing (e.g., Weigle, 2002), and the assessment of this ability involves fewer opportunities for rater identification of an examinee's L1 background, which could skew judgement.

As it relates to L2 Writing Assessment, Elder and Davies (1998) hypothesized a relationship between the L1 background of raters, the relative bias they would display towards other L1 groups, termed the *Language Distance Effect*. This hypothesis "theoriz(es), for example, that Japanese is at a greater distance from English than is Spanish, (and) could thus well be that on an English language examination, the amount of bias for or against Japanese could be larger compared to Spanish" (in Johnson & Lim, 2009, p. 489). However, subsequent investigations of the hypothesis using written compositions from the Michigan English Language Assessment Battery (MELAB) among examinees from a diverse sample of L1 backgrounds, and among native (Hamp-Lyons & Davies, 2008) and non-native raters (Johnson & Lim, 2010) proved inconclusive. Rather, the influence of L1 on L2 writing, or any L2 construct, may vary by sample, skill, and task (Trace et al., 2017).

Given the confounds concealed by L1 background, one might consider such groupings to be proxies for other differences, such as differing exposure to writing genres or differing strength and weakness in writing ability. Di Gennaro (2009), for instance, compared essay writing performance based on the four rubric scales across two groups of L2 speakers of English at an American university, *International Students* educated outside the United States, and *Generation 1.5 Students* educated in the US. Performance among the two groups differed significantly on two scales, *Rhetorical* and *Content control*. While the information on participant L1 across the two groups was not reported, it would seem logical to assume that, similar to Di Gennaro (2009), adult ESL learners of different L1s, raised and educated in different countries, may display differential performance across tasks and scales of a writing section an ESL placement exam.

To that end, this study first examines the main effects of examinee ability, rater severity, task difficulty, scale difficulty, and scale functionality using MFRM's PCM. The study then

focuses on the interaction between these facets among two L1 backgrounds, Spanish and Japanese, in order to discern the presence of bias in the context of an ESL placement exam.

Research questions

This study addresses the following research questions:

1. What are the main effects of examinee ability, rater severity, task difficulty, rubric scale difficulty, and scale functionality on the writing section of an ESL Placement Exam?
2. Are there systematic interaction effects with respect to Spanish and Japanese L1 background of examinees in relation to rater severity and task and scale difficulty for the writing section?
3. For facets where systematic interaction effects are present, are there also statistically significant mean differences in performance among L1 Spanish and Japanese examinees?

METHODS

Participants and Context

The Test-Takers

Two hundred and eight learners (N = 208) participated in Teachers College, Columbia University's *Community Language Program* (CLP) placement test administration. Examinees range in terms of ESL proficiency, age, L1, profession, and education (Vafaee & Yaghmaeyan, 2015). However, only examinee L1 background, and no other demographic variable, is provided in the dataset. The CLP offers courses for beginner, intermediate, and advanced learners across six levels (*ESL Integrated Skills*, n.d.). Examinees of L1 Spanish (n = 50) and L1 Japanese (n = 47) comprised nearly half of the sample. A bar plot of examinees by L1 is in Appendix A.

The Raters

Nine students (N = 9) were recruited from the Applied Linguistics and TESOL graduate program. The raters varied in their teaching and testing experience and were native (n = 3) and non-native (n = 6) English speakers. However, the L1s of raters of non-native English-speaking backgrounds are not provided in the dataset. Raters were classified as *Experienced* (Exp) (n = 5) or *Novice* (Nov) (n = 4) based on the number of semesters in which they took part in scoring.

Instruments

The Test Tasks

The CLP placement test battery consists of six sections: grammar, meaning, listening, speaking, reading, and writing. The writing section is strictly timed (45 minutes) and is

comprised of two tasks, each targeting a different writing genre. Task 1 prompts test-takers to write a customer review about a retail experience. Task 2 prompts test-takers to take a position on an argument and support this position with reasoning. The instructions suggest that they use 15 minutes of the allotted 45 minutes for the first task and 30 minutes for the second task.

The Rubric

Responses for both tasks were scored for Content control, Organizational control, and Language control on six-point scale (0-5) using an analytic rubric. Descriptors for Content control differ from bands 3 to 5 according to the target genres for each task. Descriptors from bands 0 to 2 for Content control, and for all bands (0-5) for Organizational control and Language control are the same for both tasks. In order to maintain confidentiality of the CLP placement test scoring practices, the rubrics for rating the writing tasks in the test have not been provided.

Procedures

Test Administration Procedures

The placement exam was administered at the beginning of the Spring, 2020 semester in order to place students who had signed up for courses into a given ESL course level. All sections, including writing, were administered in a computer lab.

Scoring Procedures

Raters were first trained using benchmarked sample responses through an online training system. Following the test administration, each response (i.e., 416) was assigned ratings for each scale of the analytic rubric by two raters. A requirement for MFRM analysis under default settings is sufficient "connectivity" among raters in terms of the examinee samples scored (Myford & Wolfe, 2000). For this reason, a rating plan was designed so that each rater scored two subsets of 22-23 samples, with each subset overlapping with adjacent raters in the plan. The rating plan is outlined Appendix B.

Software and Statistical Procedures

The test data were organized in Microsoft Excel, then exported to FACETS Version (3.83.1) (Linacre, 2019) to investigate the *Main Effects* of examinee ability, rater severity, task difficulty, scale difficulty, and scale functionality of this five-faceted design using the PCM to examine individual Task 1 and Task 2 scales. Additionally, a dummy facet representing L1 background was also created, but was not used to answer the first research question. FACETS provides estimates for parameters related to each facet deploying a log-linear transformation (Bond et al., 2021). This transformation allows for measures related to each facet to be placed on the same scale, known as the *Logit Scale*. The mathematical model in a five-faceted design using the PCM and a dummy facet can be expressed as follows:

$$\log\left(\frac{P_{njikx}}{P_{njikx}-1}\right) = B_n - G_n C_j - G_n D_i - G_n E_k - F_k \quad [1]$$

where:

P_{njikx} = probability of examinee, *n*, a score of category, *x*, on task, *i*, on scale, *k*, by rater, *j*
P_{njikx} – 1= probability of examinee, *n*, a score of category, *x*-1, on task, *i*, on scale, *k*, by rater, *j*
B_n = ability of examinee, *n*
C_j = severity of rater, *j*
G_n = first language background of examinee, *n* (Spanish, Japanese, Other)
D_i = difficulty of task, *i*
E_k = difficulty of scale, *k*
F_k = threshold of difficulty of being rated in category, *x*, relative to category, *x* -1

Additional indices for evaluating measures computed by the model, known as *Separation Statistics* and *Fit Statistics* (Fan & Bond, 2019; Bond, et al., 2021), are interpreted to gauge the impact of each facet (e.g., task difficulty) and the elements comprising them (e.g., each task).

To address the second research question, a *Bias Analysis* was conducted among L1 Spanish and Japanese examinees (G_n) to identify differences across raters in terms of severity (G_nC_j), tasks in terms of difficulty (G_nD_i) and scales in terms of difficulty (G_nE_k). A dummy facet variable was used, coding Spanish as "1", Japanese as "2" and other L1s as "3". Examinee data from all L1s were kept in the dataset to avoid disconnected subsets (e.g., Myford & Wolfe, 2000). However, the analysis only compared Spanish and Japanese examinees (n = 97). Lastly, analyses in SPSS (v.28) were conducted among Spanish and Japanese examinees to triangulate findings from the *Bias Analysis* for the third research question. Mean scores across task and scale were compared, and independent *t*-tests were conducted to identify significant mean differences.

RESULTS

Research Question 1

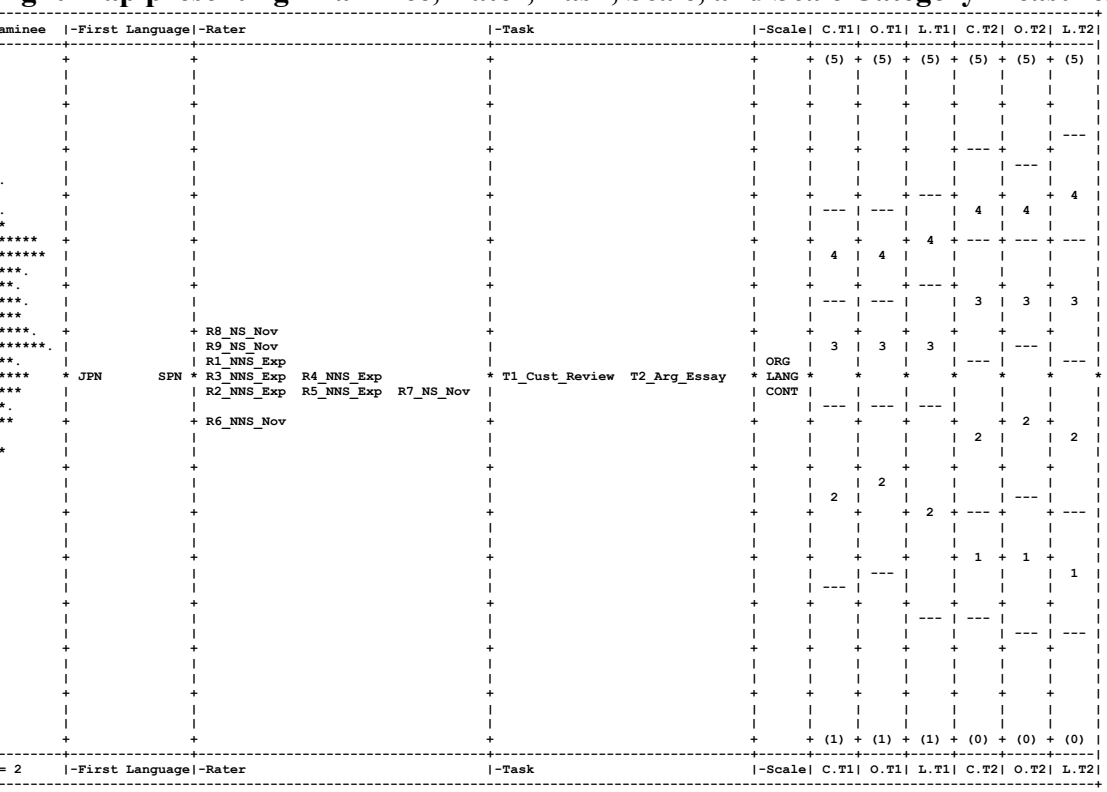
In order to demonstrate the main effects of each facet, Figure 1 provides a *Wright Map*, a graphic display of MFRM results from the FACETS output. The twelve columns represent the five-faceted design using the Partial Credit Model to represent scale functionality for each scale-within-task. The logit scale is labeled "Measr" in the first column. The logit measures for examinee ability, rater severity, task difficulty, scale difficulty, as well as scale category thresholds are positioned vertically, corresponding to examinee logit measures where a given score category for a given scale and task is most probable (Choi, 2019).

The second column shows examinees (N = 208) ordered by ability. It is positively oriented with higher-scoring examinees with positive measures at the top, and lower-scoring examinees with negative measures at the bottom. Each asterisk (*) represents two examinees and each dot (.) one examinee (See Appendix C for map with examinee L1 labels). The third column, representing two levels of the dummy facet (Spanish, Japanese), is also positively oriented to reflect differences in ability among each group of examinees on average.

The fourth column is negatively oriented, depicting differences in severity across raters (N = 9). More severe raters with positive measures are higher on the scale and more lenient raters

with negative measures are lower. The fifth and sixth columns display differences across task (k = 2) and scale (k = 3) in terms of difficulty. These facets were designed to be negatively oriented, with more difficult elements higher with positive measures and less difficult ones with negative measures. The seventh to twelfth columns represent scale functionality (k = 6) using the Rasch PCM for each rubric scale (k = 3) within each task (k = 2). The horizontal lines represent *category thresholds* where examinees at a logit would be expected to receive a given score.

FIGURE 1
Wright Map presenting Examinee, Rater, Task, Scale, and Scale Category Measures



Note: *Tasks*, T1_Cust_Review = Task 1, T2_Arg_Essay = Task 2; *Scales,* CONT = Content Control, , ORG = Organizational Control, LANG = Language Control, *Partial Credit Model for Task and Scale*, C.T1/T2, O.T1/T2, L.T1/T2, *Raters,* NS = Native, NNS = Non-native, *Experience:* Exp = Experienced, Nov = Novice

In order to characterize the main effects, Figure 1 is described in conjunction with numeric tabulations for *logit spread*, and *minimum* and *maximum* logit measures in Table 1.

TABLE 1
Summary Statistics for each Facet of Rasch Analysis

	Spread	*Min*	*Max*
Examinees *(N* = 208)	17.32	- 9.48 *(Avg: 0.50 of 5)*	+ 7.84 *(Avg: 5.00 of 5)*
Raters (N = 9)	1.76	- 0.90 *(Avg: 3.07 of 5)*	+ 0.86 *(Avg: 2.64 of 5)*
Task (k = 2)	0.28	- 0.14 *(Avg: 2.76 of 5)*	+ 0.14 *(Avg: 3.14 of 5)*
Scale (k = 3)	0.70	- 0.37 *(Avg: 3.10 of 5)*	+ 0.33 *(Avg: 2.83 of 5)*

We can characterize the facets as follows: (i) Examinees displayed a wide range of abilities but performed well overall, (ii) Raters displayed moderate differences in terms of

severity, (iii) Tasks differed incrementally in terms of difficulty but in a more pronounced manner based on *observed averages* than *task difficulty* measures, and (iv) Scales differed moderately in terms of difficulty. Lastly, by using the PCM to investigate each scale within each task, we can compare step difficulty thresholds for estimating the scale category (e.g., 1) most likely to be assigned to an examinee with a given ability measure (e.g, -5) for each task scale. Across task, the graphic display of step difficulty thresholds indicate that it is more difficult to receive a higher score category on scales for scoring Task 2, the argumentative essay (C.T2, O.T2, L.T2), than on a Task 1, the customer review (i.e., C.T1, O.T1, L.T1).

Separation Statistics and Fit Statistics

For each facet, except the dummy facet, FACETS provides statistical indicators related to (i) "the spread of element measures along the (logit scale) across a facet", called *Separation Statistics* (Eckes, 2019), and (ii) the extent to which element-level logit measures within a facet (e.g., individual examinees, raters, tasks, scales) *fit* the expectations of the model, known as *Fit Statistics* (Bond et al., 2021).

Three commonly used *Separation Statistics* include (1) the *Separation Strata* (*H*), measuring the number of "distinct classes of elements" within each facet, (2) the *Separation Reliability* (*R*), measuring the "overall precision of logit measures" within a given facet, and (3) a *Chi-Square* (χ^2) *test*, measuring the "significance of separation" among elements in a given facet, and indicating that at least two elements in the facet have statistically different measures (See Eckes, 2019). Separation statistics by facet are in Table 2. Interpretations are facet-specific.

TABLE 2
Separation Statistics for each Facet of Rasch Analysis

Facet	*Spread*	*Min*	*Max*	*Strata*	*R*	χ^2
Examinees (*N* = 208)	17.32	-9.48 *(Avg: 0.50 of 5)*	7.84 *(Avg: 5.00 of 5)*	6.70	.96	.00
Raters (*N* = 9)	1.76	-0.90 *(Avg: 3.07 of 5)*	0.86 *(Avg: 2.64 of 5)*	7.96	.97	.00
Task (*k* = 2)	0.28	-0.14 *(Avg: 2.76 of 5)*	0.14 *(Avg: 3.14 of 5)*	4.12	.94	.00
Scale (*k* = 3)	0.70	-0.37 *(Avg: 3.10 of 5)*	0.33 *(Avg: 2.83 of 5)*	8.55	.97	.00

Two commonly used *Fit Statistics* include: (1) a *Infit Mean Square* (MnSq, MS), an unstandardized value, and (2) a *Standardized Infit* (ZStd) value, both expressing the degree of fit between expected and observed data, measured using *residuals*, in overall response patterns for each element within each facet. Other fit statistics, *Outfit MS* and *Standardized Outfit*, indicate outlying cases, and are only considered for certain facets (e.g., raters) (e.g., Grabowski, 2013).

Fit statistics are interpreted as *fitting* or *misfitting* based on thresholds for acceptable fit. For *Infit MS*, with a mean of 1.00, more stringent thresholds (0.7 to 1.3) (Bond et al., 2021, p. 242) or less stringent thresholds (0.5 to 1.5) (Eckes, 2019, p. 159) have been proposed. Alternatively, an empirical determination based on the statistical characteristics of the sample (Mean Infit MS +/- 2 Standards Deviations) has been used (Kondo-Brown, 2002). For *Infit ZStd*, with a mean of 0, a value between -2 and +2 is considered acceptable (Bond, et al., 2021, p. 242). When *misfitting*, an element is either *"Underfitting"* (*Infit MS* > 1.3, 1.5, + 2 SD above Mean, *ZStd* > +2), indicating an unpredictable response pattern across other facets based on the expectations of the model, or *"Overfitting"* (*Infit MS* < 0.7, 0.5,- 2 SD above Mean, *ZStd* > - 2), indicating a deterministic response pattern across facets based on the expectations of the model (Bond et al., 2021). Underfitting profiles are considered more problematic (McNamara, 1996).

Examinee Ability

For the first facet, the Separation Statistics displayed a large *Strata* (*H* = 6.78) indicating seven distinct levels of examinees, a high *Reliability (R* = .96) in terms of the overall precision of separation across Examinee ability measures (e.g., above .80, See Eckes, 2019, p. 169), and a significance test (χ^2 = .00) indicating that there is a statistically significant difference between at least two examinees in terms of ability. In the context of the CLP, the strata is perhaps most instructive. Course offerings (*ESL Integrated Skills,* n.d.) are separated into six levels, and the strata of the writing section displays seven distinct levels. Moreover, the separation of L2 learners into six or seven levels is supported by similar formulations in well-known proficiency scales (Harsch & Malone, 2020), suggesting the result is expected.

The fit statistics are more troubling. Researchers agree that as few misfitting profiles as possible is desirable (e.g., < 2%, Pollitt & Hutchinson, 1987). For this facet, even a less stringent threshold (*Infit MS* > 0.5 and < 1.5) would lead to 13% of examinees being classified as underfitting (*n* = 27), and 16% as overfitting (*n* = 34) (See Appendix D for list misfitting examinees using threshold). Using the empirical approach, (*Infit MS, M* = 0.99, *SD* = 0.64, +/-2 SD = [- 0.29, +2.27]), the variability in Infit MS values is untenably high, and still 4% of examinees (*n* = 8) are considered misfitting (See Appendix D for list using threshold). Given the volume of misfitting profiles, element-level inspection of each examinee displaying values outside thresholds for acceptable fit would not be tenable. With that said, the fit statistics among examinees indicate the presence of many *underfitting* and *overfitting* profiles regardless of threshold, and should be considered a point of concern, undermining the trustworthiness of scores yielded from the section. This points to issues in rater variability, discussed next.

Rater Severity

Next, the Separation Statistics showed a similarly large *Strata* (*H* = 7.96), indicating eight distinct levels of raters in terms of severity, a high *Reliability* (*R* = .96), indicating the overall precision of separation across rater severity measures, and a significance test (χ^2 = .00) indicating a statistically significant difference between at least two raters in terms of severity. Such separation is not desirable for rater severity, where ideally, differences in severity will be ameliorated by rater training procedures (e.g., Weigle, 1998). Despite such procedures, the raters displayed major differences in severity, even more than examinees in terms of ability.

Individual rater measures are displayed along with the Fit Statistics in Table 3 to highlight those with differences in severity and with misfitting profiles. The following thresholds for acceptable fit were used: (i) *Infit* and *Outfit ZStd* > -2 and < +2), (ii) *Infit MS, M* = 0.99, *SD*=0.20 +/-2 SD = [0.59, 1.39], (iii) *Outfit MS, M* = 1.00, *SD* = 0.18 +/-2 SD = [0.64, 1.36]).

TABLE 3
Raters with Misfitting Profiles (*n* = 3, 33% of total raters, *N* = 9)

Rater	*Observed Average*	*Fair Average*	*Measure*	*Infit MS*	*Infit ZStd*	*Outfit MS*	*Outfit ZStd*
Rater 8 NS_Nov	2.64 of 5	2.56 of 5	+ 0.86	0.78	**- 2.8**[b]	0.79	**- 2.8**[b]
Rater 9 NS_Nov	2.75 of 5	2.59 of 5	+ 0.79	0.61	**- 5.2**[b]	**0.64**[b]	**- 4.6**[b]
Rater 1 NNS_Exp	2.97 of 5	2.73 of 5	+ 0.40	1.10	+ 1.1	1.12	+ 1.3

Rater 4	NNS_Exp	3.07 of 5	2.86 of 5	+ 0.07	1.11	+ 1.3	1.09	+ 0.2
Rater 3	NNS_Exp	3.08 of 5	2.87 of 5	+ 0.04	0.99	- 0.1	1.02	+ 0.2
Rater 7	NS_Nov	3.17 of 5	3.05 of 5	- 0.39	0.93	- 0.8	0.91	- 1.0
Rater 5	NNS_Exp	2.72 of 5	3.05 of 5	- 0.39	1.23	**+ 2.4[a]**	1.21	**+ 2.1[a]**
Rater2	NNS_Exp	3.10 of 5	3.09 of 5	- 0.48	1.16	1.8	1.17	+ 1.9
Rater6	NNS_Nov	3.07 of 5	3.28 of 5	- 0.90	0.99	- 0.1	1.00	+/- 0.0

Note: NS = Native, NNS = Non-native, *Experience*, Exp = Experienced, Nov = Novice, *MS* = Unstandardized Infit or Outfit value, *ZStd* = Standardized Infit or Outfit value, ***a* = underfitting**, ***b* = overfitting**

Several patterns can be gleaned from Table 4. First, differences in severity are most pronounced among novice raters, who were more lenient (e.g., *Rater 6* = -.90) or severe (e.g., *Rater 8* = +.86, *Rater 9* = +.79) than experienced raters. The two most severe raters also showed overfitting profiles, based multiple infit or outfit measures, and Rater 5 displays a slightly underfitting profile based on the Standardized infit and outfit. Such differences in rater severity are not desirable. However, considering how the most prominent issues in discrepant severity and model fit are present among two novice raters, one might speculate that further experience rating the CLP writing section and undergoing more training sessions may ameliorate differences (e.g., Elder et al., 2005; Lumley & McNamara, 1995). Also, in the case of the experienced rater displaying a slightly underfitting profile, it is possible that this person would also benefit from further training on the rubric despite their experience with CLP rating procedures. Lastly, the self-consistency among most experienced raters, as demonstrated by the fit statistics, suggests that the CLP's rater training procedures, for the most part, have worked relatively well.

Task Difficulty

For tasks, the Separation Statistics exhibited a moderately sized *Strata* (H = 4.12), indicating four distinct levels of tasks in terms of difficulty, a high *Reliability* (R = .94) in terms of the overall precision of separation across task difficulty measures, and a significance test (χ^2 = .00), indicating that there is a statistically significant difference between at least two tasks in terms of difficulty. Based on an empirical approach for determining acceptable fit (*Infit MS, M* = 0.99, *SD* = 0.1 +/-2 SD = [0.98, 1.02]), both tasks displayed acceptable fit, but were different in terms of difficulty, as can be seen in Table 4.

TABLE 4
Tasks with Misfitting Profiles (*k* = 0, 0% of total tasks, *k* = 2)

Task		*Observed Average*	*Fair Average*	*Measure*	*Infit MS*	*Infit ZStd*
Task 1	Customer Review	3.14 of 5	3.05 of 5	+ 0.14	1.00	+/- 0.0
Task 2	Argumentative Essay	2.76 of 5	2.72 of 5	- 0.14	0.98	- 0.4

Note: *Infit MS* = Unstandardized Infit Mean square value, *Infit ZStd* = Standardized Infit on z-distribution

The results suggest that Task 2, an argumentative essay, was somewhat more difficult than Task 1, a customer review, leading to the *strata* indicating four distinct levels of difficulty. Based on the results, we can neither conclude that the tasks are interchangeable nor explicitly target distinct levels of proficiency. Intuitively, the task genres may target slightly different levels (Jeong, 2017; Pollitt & Hutchinson, 1987), which could be desirable in the CLP. However, only the test designers would be able to say whether this difference across task was intended.

Rubric Scale Difficulty

Lastly, the Separation Statistics for scale difficulty reveal the largest *Strata* compared to other facets (H = 8.55), indicating nearly nine distinct levels of scales in terms of difficulty, a high *Reliability* (R = .97) in terms of the overall precision of separation across Scale Difficulty measures, and a significance test (χ^2 = .00) indicating that there is a statistically significant difference between at least two scales in terms of difficulty. Further inspection of fit statistics using empirically-derived thresholds (*Infit MS, M* = 0.99, *SD* = 0.06, +/-2 SD = [0.87, 1.11]) reveals no misfitting scales. The results are outlined in Table 5.

TABLE 5
Scales with Misfitting Profiles (*k* = 0, 0% of total scales, *k* = 3)

Scale	*Observed Average*	*Fair Average*	*Measure*	*Infit MS*	*Infit ZStd*
Organizational control	2.83 of 5	2.76 of 5	+ 0.33	0.99	- 0.1
Language control	2.92 of 5	2.87 of 5	+ 0.03	0.92	- 1.7
Content control	3.10 of 5	3.05 of 5	- 0.37	1.05	+1.0

Note: *Infit MS* = Unstandardized Infit Mean square value, *Infit ZStd* = Standardized Infit on z-distribution

Overall, the Separation Statistics indicate significant differences in Scales in terms of difficulty, but no issues with model fit among them. A closer inspection of the statistical characteristics of each scale reveals that Organizational control was the most difficult, followed by Language control, then Content control. Past studies have found language-oriented scales to be scored more harshly than content-oriented scales, even if this is downplayed in the task design (McNamara, 1990, 1996). Additionally, the relative difficulty of language control as a scale has been found to differ across L2 proficiency level and task type (Grabowski, 2013). Admittedly, it is matter of debate whether discourse-level organizational features are more associated with language use or content (e.g., Cumming, et al., 2021; Di Gennaro, 2009). Nonetheless, the results are not necessarily unexpected in the context of the CLP since differences in scale difficulty have been found based on features of the written responses that are assessed.

Rubric Scale Functionality

Given the findings for task and rubric sub-construct difficulty, an examination of the scale functionality within each task was undertaken using the Rasch PCM. FACETS output allows us to evaluate scale functionality by examining frequency of use and step threshold difficulties for score categories numerically and graphically. Numerically, *Rasch-Andrich* ("*Most-Probable From"*) *Thresholds* provides step threshold difficulties delineating the logit measure at which an examinee is most likely to receive a given scale category (e.g., 1-to-2). Graphically, FACETS represents the probability of receiving a category score at a given examinee ability level through *Probability Curves.* The numeric and graphic output allows users to identify the extent to which the category scores increase monotonically with step threshold difficulties. Additionally, by using the PCM, differences across task and rubric scale can be observed. The step difficulty thresholds by scale within each task are in Table 6 (See Appendix E for category score frequency table).

TABLE 6
Scale Category across Tasks – Rasch-Andrich ("Most Probable From") Threshold *(k* = 6)

	Task 1 – Customer Review						Task 2 – Argumentative Essay					
Cat	C.T1	Outfit MS	O.T1	Outfit MS	L.T1	Outfit MS	C.T2	Outfit MS	O.T2	Outfit MS	L.T2	Outfit MS
0	--*	--	--*	--	--*	--	--*	0.9	--*	0.6	--*	0.7
1	--*	1.1	--*	1.7	--*	1.8	**- 5.02**	1.0	**- 5.55**	1.1	**- 5.33**	0.5
2	**- 4.57**	0.9	**- 4.24**	1.0	**- 5.32**	1.1	**- 3.06**	1.2	**- 2.52**	0.9	**- 3.01**	0.8
3	**- 0.50**	0.8	**- 0.73**	0.9	**- 0.56**	1.0	**+ 0.56**	1.1	**+ 0.66**	1.0	**+ 0.42**	0.9
4	**+ 1.93**	1.2	**+ 1.92**	0.6	**+ 2.17**	0.8	**+ 3.01**	1.1	**+ 3.11**	1.1	**+ 2.99**	0.8
5	**+ 3.13**	1.2	**+ 3.06**	1.0	**+ 3.71**	1.3	**+ 4.51**	0.9	**+ 4.30**	1.0	**+ 4.93**	1.0

Note: * = Rasch-Andrich Threshold not calculated, *Cat* = Category Score, *Scales,* C = Content, , O = Organization, L = Language, *Task and Scale*, C.T1/T2, O.T1/T2, L.T1/T2, Outfit MS = Outfit Mean Square

As we see, there is monotonic increase in the category scores and step difficulty thresholds across all three scales for the two tasks. However, differences emerge in terms of scale functionality across tasks. First, the lower scale categories (0-to-1, 1-to-2) are more frequently used for Task 2 than Task 1, leading *Rasch-Andrich ("Most Probable From")* *Thresholds* to be computed between the lowest two categories for Task 2, but not Task 1. Also, the step difficulty thresholds correspond to higher examinee ability measures for Task 2 compared to Task 1 in the middle (Category "3") and at the top of the scale (Category "5").

Finer-grained differences within scales are present. For instance, it was most difficult to receive the top score ("5") for Language control for both tasks, in line with past findings regarding the difficulty of such scales (e.g., McNamara, 1996). With that said, two overall conclusions can be drawn. First, the argumentative essay may have been more difficult for examinees compared to the genre of a customer review. Second, the raters may have more stringently applied the rubric scales categories, described in similar terms aside from Content control from Bands 3-5 to Task 2 responses compared to Task 1 responses. Some possible reasons for this difference could be that (i) the argumentative essay was more difficult than the customer review for examinees, (ii) the rubric criteria for Content, Organization, and Language were assessed more harshly by raters for the argumentative essay, or (iii) a combination of both.

Research Question 2

L1 Spanish and Japanese examinees comprised nearly half of the sample. For this reason, performance among them had a larger impact on measures compared to those of other L1s. The distribution of ability measures by sub-group is in Appendix F.

In order to identify the presence of systematic bias, or DFF, for or against Spanish or Japanese examinees among raters, tasks, or scales, the following interactions were investigated: (1) A Two-way interaction between sub-group ability and severity among raters, (2) A Two-way interaction between sub-group ability and difficulty among tasks, (3) A Two-way interaction between sub-group ability and difficulty among scales averaged across tasks, and (4) A Three-way interaction between sub-group ability and difficulty among scales within tasks.

The following indices from *Bias Analysis* results in FACETS output were considered. First, the *Bias Size* was considered, indicating the direction and magnitude of bias in logits. For

these analyses, a positive direction (+) is interpreted as a bias *for* an L1 and a negative bias direction (-) is interpreted as a bias *against* an L1 because the dummy facet is associated with examinee ability, a positively-oriented facet. The magnitude of the *bias size* indicates the extent that the interaction increased (e.g., + 0.2) or decreased (e.g., - 0.2) ability measures among examinees of a given L1 when scored by a given rater, performing a given task, or scored using a given scale (Eckes, 2005; Grabowski, 2013). Second, the *z-score*, an inferential statistic and its associated probability were considered. Researchers recommend a *z*-score of +/- 1.96 (Johnson & Lim, 2010; McNamara, 1996) or +/-2 as the threshold for determining if statistically significant bias is present in the interaction (Di Gennaro, 2009; Eckes, 2005; Elder, et al., 2005; Grabowski, 2013; Lumley & McNamara, 1995). This threshold is associated with .05 probability, indicating with 95% certainty that a significant difference exists.

L1 x Rater Severity

Bias Analysis was conducted between Spanish and Japanese examinee ability and rater severity. Full results are in Appendix G, and for rater 8, in Table 7.

TABLE 7
Interaction between Spanish and Japanese Examinee Ability and Rater 8 Severity

Rater	L1	Bias Size	*z*-score	Probability
Rater 8_NS_Nov *(Severity: +.86)*	Spanish (*n* = 22)	+ 0.38	+ 1.92	.0590
	Japanese (*n* = 26)	- 0.22	- 1.53	.1271

Note: *L1 English,* NS = Native, , *Experience*, Nov = Novice; *n* = Responses scored among each sub-group

While *bias size* varies in terms of direction and magnitude, no systematic bias (*z-score* +/- 2, <.05 level) was found. Rater 8, a novice rater with the highest severity measure (*logit* = +.86), showed the largest bias measure among the group (*Bias Size* = +.38), with near significant bias *for* L1 Spanish examinees, indicating that this rater was noticeably less severe than his overall measure when rating L1 Spanish examinees in the sample. However, the results must be interpreted cautiously since each rater did not score an equal number of responses written by Spanish and Japanese examinees, limiting any generalizable claims that can be made.

L1 x Task Difficulty

Next, a Bias Analysis was conducted between Spanish and Japanese examinees in terms of ability and task in terms of difficulty. The results of the bias analysis are in Table 8.

TABLE 8
Interaction between Spanish and Japanese Examinee Ability and Writing Task

Task	L1	Bias Size	*z*-score	Probability
T1. Customer Review	Spanish	+ 0.14	+ 1.46	.1441
	Japanese	**- 0.34**	**- 3.51**	**0005****
T2. Argumentative Essay	Spanish	- 0.14	- 1.44	.1506
	Japanese	**+ 0.34**	**+ 3.54**	**.0005****

Note: ** = probability statistically significant at a < .01 level; **Bold** = Task with the presence of bias

As seen above, two systematic sources of bias emerge. First, Task 1 displayed significant bias *against* L1 Japanese examinees (*Bias Size* = -.34), indicating that the task was significantly more difficult for this group compared to the average measures. Conversely, Task 2, displayed significant bias *for* L1 Japanese examinees (*Bias Size* = +.34), indicating the task was significantly less difficult for this group compared to average measures. Based on the analysis, there is a systematic interaction effect between task genre and performance among Japanese examinees, manifesting in significant bias *against* them on Task 1 and *for* them on Task 2.

L1 x Scale Difficulty

An analysis was also conducted among Spanish and Japanese ability and average scale difficulty across task. Results are in Table 9.

TABLE 9
Interaction Spanish and Japanese Examinee Ability and Rubric Scale Difficulty

Sub-Component	L1	Bias Size	*z*-score	Probability
Content Control	Spanish	+ 0.12	+ 1.04	.2989
	Japanese	- 0.06	- 0.48	.6328
Organizational Control	Spanish	+/- 0.00	+/- 0.00	.9989
	Japanese	+ 0.02	+ 0.15	.8847
Language Control	Spanish	- 0.13	- 1.08	.2823
	Japanese	+0.05	+ 0.38	.2823

As seen above, no systematic bias was found for either L1 and any particular scale difficulty averaged across tasks. In tandem with the findings regarding task difficulty and Japanese examinee ability, it can be concluded that differences in performance across tasks were concealed when inspecting scale difficulty averaged across tasks. However, if significant bias for Japanese examinees was present in Task 1, and against them in Task 2, but no bias was present among scales averaged across tasks, would significant bias be present for scales within tasks?

L1 x Task Difficulty x Scale Difficulty

Lastly, a Bias Analysis involving Spanish and Japanese ability and scale difficulty within each task was conducted. The results are in Table 10.

TABLE 10
Interaction between Spanish and Japanese Examinee Ability, and Task and Scale Difficulty

Scale	L1	Bias Size	*z*-score	Probability
Task1_Content control	**Spanish**	**+ 0.42**	**+ 2.65**	**.0095****
	Japanese	- 0.26	- 1.58	.1174
Task1_Organizational control	Spanish	+ 0.04	+ 0.25	.8053
	Japanese	**- 0.45**	**- 2.71**	**.0080****
Task 1_Language control	Spanish	- 0.08	- 0.46	.6451
	Japanese	- 0.31	- 1.78	.0787
Task 2_Content control	Spanish	- 0.25	- 1.58	.1174
	Japanese	+ 0.15	+ 0.92	.3578
Task 2_Organizational control	Spanish	- 0.04	- 0.25	.8068
	Japanese	**+ 0.48**	**+ 2.93**	**.0043****
Task 2_ Language control	Spanish	- 0.18	- 1.05	.2942

Japanese	**+ 0.39**	**+ 2.29**	**.0245***

Note: ** = significant at a < .01 level; * = significant at a < .05 level; **Bold** = Scale with the presence of bias

Referring to the results, four scales displayed significant bias for or against a particular L1. For Task 1, Content control displayed a significant bias *for* Spanish examinees (*Bias Size* = +.42) and Organizational control displayed a significant bias *against* Japanese examinees (*Bias Size* = -.45). For Task 2, both Organizational control and Language control showed a significant bias *for* Japanese examinees (*Bias Size* = +.48, +.39, respectively).

So what does this tell us? For Task 1, Spanish examinees performed significantly better for Content control compared to average measures, suggesting the *differential* ability to convey the content effectively in a customer review. By contrast, Japanese examinees performed significantly worse for Organizational control compared to average measures, indicating a *differential* struggle in effectively organizing their customer review responses. For Task 2, Japanese examinees performed significant better on both the Organizational control and Language control scales compared to average measures, suggesting a *differential* ability to organize their response and use language effectively in the context of an argumentative essay.

Research Question 3

To triangulate patterns of systematic bias identified from the Bias Analyses, a comparison of means and independent *t*-tests were conducted across task- and scale-level performance among Spanish and Japanese examinees. Since *Bias Sizes* are computed in MFRM in relation to average measure in the sample, means for the entire sample are in Table 11.

TABLE 11
Comparison of Means by Writing Task and Subscale (*N* =208)

Rubric Scale	Task 1		Task 2		Independent *t*-test	
	M	SD	*M*	SD	*t*(414)	*p*
Avg	3.14 out of 5	1.01	2.76 out of 5	1.04	-3.780	<.001**
CC	3.35 out of 5	1.11	2.85 out of 5	1.12	-4.573	<.001**
OC	3.02 out of 5	1.06	2.64 out of 5	1.10	-3.588	<.001**
LC	3.05 out of 5	0.98	2.78 out of 5	1.07	-2.684	.008**

Note: **Avg** = Scale Average, **CC** = Content control, **OC** = Organizational control, **LC** = Language control, ** = *t*-test statistically significant at a < .01 level

As seen above, there are significant differences ($p <. 01$) between the tasks on average (*M = -0.38*), as well as for Content control (*M = -0.50*), Organizational control (*M = -0.38*), and Language control (*M* = -0.27). Based on the results, we can assume that the expectation of the MFRM model is that Task 2 is more difficult than Task 1 overall and across sub-construct.

Within-group Comparisons across Writing Task

To illustrate within-group differences, mean comparisons and *t*-tests were conducted across tasks, and scales within tasks. The results are in Table 12.

TABLE 12
Comparison of Means across Customer Review and Argument Essay Tasks

		Task 1		Task 2		Independent *t*-test	
		M	SD	*M*	SD	*t*(98)*t*(92)	*p*
Avg	L1 Spanish	3.03 out of 5	1.01	2.56 out of 5	1.03	-2.304	.023*
	L1 Japanese	2.97**jp** out of 5	0.84	2.87**JP** out of 5	0.92	-0.550	.583
CC	L1 Spanish	3.29**SP** out of 5	1.11	2.68 out of 5	1.15	-2.699	.008**
	L1 Japanese	3.15 out of 5	0.91	2.95 out of 5	1.03	-0.998	.321
OC	L1 Spanish	2.90 out of 5	1.09	2.44 out of 5	1.10	-2.100	.038*
	L1 Japanese	2.84**jp** out of 5	0.91	2.77**JP** out of 5	1.02	-0.351	.726
LC	L1 Spanish	2.90 out of 5	0.94	2.55 out of 5	1.03	-1.775	.079
	L1 Japanese	2.93 out of 5	0.83	2.88**JP** out of 5	0.87	-0.285	.776

Note: **Avg** = Scale Average, **CC** = Content control, **OC** = Organizational control, **LC** = Language control, **SP** = Mean flagged for significant bias for Spanish examinees, **jp** = Mean flagged for significant bias against Japanese examinees, **JP** = Mean flagged for significant bias for Japanese examinees, * = *t*-test statistically significant at a < .05 level, ** = *t*-test statistically significant at a < .01 level

Across tasks, comparisons can be made between Spanish and Japanese examinees. On the one hand, performance among Spanish examinees on Task 1 and Task 2 follows a similar pattern as the overall sample, with significantly lower scores on Task 2 overall (M = - 0.47, $p < .05$), for Content control (M = -0.61 , $p < .01$) and Organizational control (M = -0.46, $p < .05$), and noticeably, though not significantly, lower for Language control (M = -0.35). On both tasks, Spanish examinees scored below the average for the entire sample (e.g. *Task 1 Avg* = 3.03 v. 3.14, *Task 2 Avg* = 2.56 v. 2.76), but follows the same trend. On the other hand, performance among Japanese examinees deviated from this pattern. While their scores decreased from Task 1 to Task 2, the differences were not significant overall.

To further illustrate how performance among Japanese examinees deviated from the overall pattern is that, for Task 1, they scored below the average (*Avg* = 2.97 v. 3.14), but for Task 2, above the average (*Avg* = 2.87 v. 2.76). This finding suggests that the results from the MFRM bias analysis for Japanese examinees may have been influenced by deviations from patterns in the overall sample, rather than just in comparison to Spanish examinees.

Between-Group Comparisons across Spanish and Japanese Examinees

Further inspection of mean differences and *t*-tests between groups for tasks overall and scales within task was conducted, with results in Table 13.

TABLE 13
Comparison of Means across Spanish and Japanese Examinees

		L1 Spanish (n = 50)		**L1 Japanese** (n = 47)		**Independent *t*-test**	
		M	SD	*M*	SD	*t*(95)	*p*
Avg	Task 1	3.03 out of 5	1.01	2.97jp out of 5	0.84	0.317	.752
	Task 2	2.56 out of 5	1.03	2.87JP out of 5	0.92	-1.560	.122
CC	Task 1	3.29SP out of 5	1.11	3.15 out of 5	0.91	0.677	.500
	Task 2	2.68 out of 5	1.15	2.95 out of 5	1.03	-1.250	.277

OC	Task 1	2.90 out of 5	1.09	2.84^{jp} out of 5	0.91	0.294	.770
	Task 2	2.44 out of 5	1.10	2.77^{JP} out of 5	1.02	-1.529	.129
LC	Task 1	2.90 out of 5	0.94	2.93 out of 5	0.83	-0.166	.868
	Task 2	2.55 out of 5	1.03	2.88^{JP} out of 5	0.87	-1.699	.093

Note: **Avg** = Scale Average, **CC** = Content control, **OC** = Organizational control, **LC** = Language control, **SP** = Mean flagged for significant bias for Spanish examinees, **jp** = Mean flagged for significant bias against Japanese examinees, **JP** = Mean flagged for significant bias for Japanese examinees

For Task 1, no significant mean differences between Spanish and Japanese examinee performance were identified. For Task 2, Japanese examinees performed better than Spanish examinees on average (*M = -0.31*) and across scales, though no mean differences were significant using independent *t*-tests. The results suggest that Japanese examinees performed better on this task overall, for Organizational control and Language control, specifically.

Triangulation of Results

By comparing mean differences between Spanish and Japanese examinees, and among the entire sample, the manner by which MFRM identified systematic bias comes into focus. For the entire sample, Task 2 was more difficult than Task 1, so deviations from that pattern were flagged as displaying bias. Spanish examinees, who performed significantly worse on Task 2 compared to Task 1, followed that pattern. However, Japanese examinees did not, scoring below the average for Task 1 and above the average for Task 2, leading MFRM to flag score deviations as *bias against them* for Task 1 and *for them* for Task 2.

To illustrate, the Content control scale for Task 1 was flagged for displaying significant bias for Spanish examinees (*Bias Size* = +.42). On this scale, Spanish examinees performed better than Japanese examinees (3.29 v. 3.15), but was lower than the overall sample (3.35), which could be a factor in this scale mean being flagged for displaying significant bias.

For Japanese examinees, performance on Task 1 was flagged for displaying significant bias against them (*Bias Size* = -.34), specifically for Organizational control (*Bias Size* = -.45), while Task 2 was flagged for displaying significant bias for them (*Bias Size* = +.34), specifically for Organizational control and Language control (*Bias Size* = +.48 and .39).

In terms of Task 1, performance among Japanese examinees was similar to that of Spanish examinees (2.97 v. 3.03), though below the overall sample (3.14). Similar results were found for Organizational control compared to Spanish examinees (2.84 v. 2.90) and the entire sample (3.03). As for Task 2, Japanese examinees performed slightly worse than for Task 1, but above mean scores among Spanish examinees (2.87 v. 2.56) and the entire sample (2.76). The same can be said for Organizational control and Language control, in which Japanese examinees scored above Spanish examinees (OC = 2.77 v. 2.44, LC = 2.88 v. 2.55) and above the mean scores for the entire sample (OC = 2.64, LC = 2.78). The results suggest that the degree to which Task 2 scores decreased plays a role in significant bias for Japanese examinees being identified.

DISCUSSION

The purpose of the study was two-fold: (1) to examine the main effects of examinee ability, rater severity, task difficulty, scale difficulty, and scale functionality using MFRM's PCM on an ESL placement exam, and (2) to investigate the interaction between these facets

among two sub-groups, L1 Spanish and Japanese examinees, for the presence of bias. Three research questions were posed, the first relating to the main effects of each facet among the entire sample and the second and third pertaining to the Bias Analysis among the sub-groups.

For the first research question, the effects of each facet of measurement can be summarized as follows. Examinee ability measures varied widely and were characterized by many examinee scores not fitting the expectations of the MFRM model (Bond, et al., 2021; Pollitt & Hutchinson, 1987). On its face, the presence of a large group of misfitting examinees is problematic. However, since scores from the writing section of the CLP placement test are not used to make placement decisions, one might ask whether similar issues would present themselves across all five sections of the exam.

The effects of rater severity were similarly troubling, varying widely among the nine raters, though most prominently among novice raters, and less so among most experienced raters. Additionally, three raters were found to be misfitting, two of which were novice raters found to be the most severe and overfitting. This finding is consistent with past research in L2 Writing assessment (e.g., Elder, et al., 2005, Weigle, 1998), which have found that novice raters display mores issues with discrepant severity and misfitting profiles. For this reason, it is worth considering whether the CLP's training procedures for scoring the writing section are effective at ameliorating such issues over time. The issue of *overfitting raters*, in particular, is associated with using a restricted range of scores in comparison to other raters, known as a *halo effect*, (Grabowski, 2013). So, it would stand to reason that subsequent rater training could focus on how each level of the rubric scale (0-5) is appropriately applied to examinee responses.

The main effects of task difficulty, rubric component difficulty, and scale functionality using the Rasch PCM should be considered in tandem. For tasks, the genre of an argumentative essay was found to be somewhat more difficult compared to that of a customer review. For scales overall, Organizational control was found to be the most difficult, followed by Language control, and then Content control. The scale functionality of the three scales within each task using the PCM further draws into focus differences in task genre difficulty, with Content, Organization, and Language functioning in a more difficult manner for the argumentative essay, with the Language scale being the most difficult to receive the maximum score "5" for both tasks. These results indicate a *method effect* of task difficulty (Bachman & Palmer, 1996), manifested in differing analytic rubric scale difficulty and functionality across tasks. Indeed, past research has found in L2 writing (e.g., Jeong, 2017; Pollitt & Hutchinson, 1987) that writing tasks requiring familiarity with varied genres may differ in difficulty among examinees. Similar findings have been observed in relation to difficulty levels of writing sub-constructs, such as Rhetorical control (Di Gennaro, 2009) or Language control (McNamara, 1990, 1996).

For the CLP, the question is whether the task genre or scoring responses via Content, Organization, and Language led to differences in performance. Indeed, an academically-oriented argumentative essay, requiring an opinion supported by reasons and examples seems more difficult than an informal customer review, requiring a description of only what they liked and disliked. However, since task performance is determined by ratings for Content, Organization, and Language, the effect of task and scale difficulty is challenging to disentangle (See Cummings, et al., 2021, p. 132, On Task Stimuli & Rating Criteria).

The second research question investigated the presence of systematic bias, or *DFF*, for or against L1 Spanish and Japanese examinees across raters, tasks, scales, and scales-within-tasks. Relatedly, the third question examined mean differences in performance across the sub-groups in order to illustrate how the results from the Bias Analysis in MFRM were computed. Overall, the

Bias Analysis found no significant bias for or against Spanish or Japanese examinees among individual raters or individual scales averaged across task. However, among tasks, the customer review was found to display a significant bias *against* Japanese examinees, specifically when scored for Organizational control, and Content control displayed a significant bias for Spanish examinees. Conversely, the argumentative essay displayed a significant bias *for* Japanese examinees, and specifically when scored for Organizational control and Language control. This suggests a differential difficulty of task genre and scale sub-construct, potentially associated with L1 (Elder & Davies, 1998), but more likely, with exposure to different genres (Jeong, 2017).

Upon further inspection of mean differences, performance among Spanish examinees was consistent with overall trends in the sample (i.e., scoring significantly lower for Task 2 than Task 1), leading to fewer mean scores being flagged as displaying significant bias for or against them. Japanese examinees deviated from this trend (i.e., scoring only slightly lower for Task 2 than Task 1, and well-above other groups for Task 2), leading MFRM to identify a significant bias *against them* for the customer review, and *for them* for the argumentative essay, and suggesting that Bias Analysis results were influenced by the model's expectation that Task 2 was significantly more difficult than Task 1 (Bond et al., 2021).

In sum, we can not attribute differences in performance to L1. Indeed, cultural and linguistic bias has been noted on L2 English proficiency exams (e.g., Chen & Henning, 1985). However, since L1 background has a been found to serve as a proxy belying a range of other demographic variables, one might ask how L2 proficiency and experience (Grabowski, 2013) or L2 learner's educational background (Di Gennaro, 2009) influenced the results, as those of differing L1s may have different reasons, educationally, professionally, and personally, for attending the CLP.

CONCLUSION

The current study used MFRM to investigate, first, the main effects of the test facets on test scores from the writing section of an ESL program's placement exam. The results regarding the facets of examinee ability and rater severity suggest issues in score variability not expected by the model. Based on the results for task difficulty, scale difficulty, and scale functionality, it can be concluded that either (i) the task genre of an argumentative essay was more difficult than the genre of a customer review, (ii) the scales for Content, Organization, and Language were assessed more harshly by raters based on the rubric descriptors, or (iii) a combination of both.

A bias analysis further examined whether systematic bias was present between these facets among L1 Spanish examinees and Japanese examinees. The results revealed that the task genre of a customer review displayed a significant bias against L1 Japanese examinees, particularly for Organization, while the task genre of an argumentative essay displayed significant bias for them, particularly for Organizational control and Language control. A review of the descriptive statistics further illustrated this pattern in performance.

For each analysis, notable limitations exist. First, for the main effects of each facet, characteristics unique to the sample (e.g., misfitting examinees, novice misfitting raters) may undermine conclusions that can be drawn from the results. Additionally, while differences in task difficulty and scale functionality were found, the study was unable to attribute such differences to specific causes (e.g., genre-specific instructions in task stimulus, descriptors across scale categories in the rubric, how raters were trained to score responses) (e.g., Cumming et al., 2021).

As for the Bias analysis, the sub-groups of L1 Spanish and Japanese examinees were conveniently sampled from the CLP Test Administration, and no further information was available regarding their education, profession, or reason for attending the CLP. For this reason, the extent that differences in writing performance across examinees of L1 Spanish and Japanese background was a *proxy* for other demographic variables, such as professional or educational experience (e.g., Di Gennaro, 2009), could not be investigated empirically using the data available from this CLP placement test administration.

Similarly, another limitation related to the Bias Analysis in this study was that the CLP rating plan for scoring the writing section in this administration only classified raters as *native* or *non-native English speakers* and did not disclose the L1 background of raters whose native language was not English. As a result, it was not possible to evaluate empirically the *Language Distance Effect* (e.g., Elder & Davies, 1998; Hamp-Lyons & Davies, 2008; Johnson & Lim, 2010) since testing this hypothesis would require a comparison of bias size according to rater L1 background and examinee L1 background. Moreover, even if rater L1 background were to have been provided, the number of responses that each rater scored that were written by L1 Spanish and L1 Japanese examinees varied widely (see Appendix B). Such differences in the number of scored responses written by L1 Spanish and Japanese examinees would have severely limited the conclusions that could be drawn about the *Language Distance Effect* hypothesis in this study.

Despite the limitations, the current study identified differences in performance across more and less academic writing genres, and across sub-constructs for measuring writing ability among all examinees, and *differentially* among Spanish and Japanese examinees. Pedagogically, in the CLP, this information could inform curricular design regarding the genres that that student are exposed to at certain course levels (e.g., Yelp Reviews) and how instructional material could emphasize features of the genre that correspond to target sub-constructs (e.g., Topical, Rhetorical, Linguistic features). Likewise, the placement exam could be used to identify sub-groups (e.g., by L1) that systematically perform worse on certain tasks and sub-constructs, so that these patterns could be inform teaching decisions in courses in which those sub-groups are enrolled. Methodologically, follow-up studies could consider other demographic variables beyond L1 to investigate the extent the impact that linguistic, educational, and professional background have on performance across writing genre and sub-construct.

ACKNOWLEDGEMENTS

Special thanks to Dr. Kirby Lim for her continual support and guidance throughout the time that I've studied under her, and to Pieter Lauwaert and Tania Cristina Alfonso Quitian, who used the similar datasets from past CLP placement tests, for their willingness to collaboratively make sense of their statistical findings.

REFERENCES

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language testing, 12*(2), 238-257. https://doi-org.ezproxy.cul.columbia.edu/10.1177/026553229501200206

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bigelow, M., & Watson, J. (2013). Literacy, and orality in L2 learning. *The Routledge handbook of second language acquisition*, 461-475.

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9780429030499

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press. https://doi-org.ezproxy.cul.columbia.edu/10.1191/0265532203lt245

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*(2), 155-163. https://doi-org.ezproxy.cul.columbia.edu/10.1177/026553228500200204

Cumming, A., Cho, Y., Burstein, J., Everson, P., & Kantor, R. (2021). Assessing academic writing. In *Assessing academic English for higher education admissions* (pp. 107-151). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9781351142403

Choi, I. (2019). Application of the rating scale model and the partial credit model in language assessment research. In V. Aryadoust, & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I* (pp. 132-152). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9781315187815

Di Gennaro, K. (2009). Investigating differences in the writing performance of international and Generation 1.5 students. *Language Testing*, *26*(4), 533-559. https://doi-org.ezproxy.cul.columbia.edu/10.1177/0265532209340190

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust, & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume I* (pp. 153-175). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9781315187815

Elder, C., & Davies, A. (1998). Performance on ESL examinations: Is there a language distance effect? *Language and Education*, *12*(1), 1-17. https://doi.org/10.1080/09500789808666736

Elder, C., Knoch, U., Barkhuizen, G., & Von Randow, J. (2005). Individual feedback to enhance rater training: Does it work?. *Language Assessment Quarterly*, *2*(3), 175-196. https://doi.org/10.1207/s15434311laq0203_1

ESL Integrated Skills (n.d.). Community Language Program. Retrieved October 11, 2022 from https://www.tc.columbia.edu/communitylanguage/clp-courses/esl-integrated-skills-course/

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In *Quantitative Data Analysis for Language Assessment Volume I* (pp. 83-102). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9781315187815

Grabowski, K. (2013). Investigating the construct validity of a role-play test designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. J. Ross, & G.

Kasper (Eds.), *Assessing second language pragmatics* (pp. 149-171). Palgrave Macmillan. https://doi.org/10.1057/9781137003522_6

Gujord, A.-K. H. (2022). Who succeeds and who fails? Exploring the role of background variables in explaining the outcomes of L2 language tests. *Language Testing*, *40(2*) https://doi.org/10.1177/0265532222110011

Hamp-Lyons, L. I. Z., & Davies, A. (2008). The Englishes of English tests: Bias revisited. *World Englishes*, *27*(1), 26-39. https://doi-org.ezproxy.cul.columbia.edu/10.1111/j.1467-971X.2008.00534.x

Harsch, C., & Malone, M. E. (2020). Language proficiency frameworks and scales. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 33-44). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9781351034784

IELTS Demographic Data 2021 (n.d.). IELTS. Retrieved December 4, 2022 from https://www.ielts.org/for-researchers/test-statistics/demographic-data

Jeong, H. (2017). Narrative and expository genre effects on students, raters, and performance criteria. *Assessing writing*, *31*, 113-125. https://doi.org/10.1016/j.asw.2016.08.006

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505. https://doi-org.ezproxy.cul.columbia.edu/10.1177/0265532209340186

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31. https://doi-org.ezproxy.cul.columbia.edu/10.1191/0265532202lt218oa

Linacre, J.M. (2019). *FACETS: Rasch measurement computer program* (Version 3.83.1). Mesa Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language testing*, *12*(1), 54-71. https://doi-org.ezproxy.cul.columbia.edu/10.1177/026553229501200104

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180. https://doi-org.ezproxy.cul.columbia.edu/10.1177/026553229801500202

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the Test of Spoken English assessment system. *ETS Research Report Series*, *2000*(1), i-51. https://doi.org/10.1002/j.2333-8504.2000.tb01829.x

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52-76. https://doi-org.ezproxy.cul.columbia.edu/10.1177/026553229000700105

McNamara, T. F. (1996). *Measuring second language performance*. Longman Publishing Group.

Ockey, G. J. (2021). Item response theory and many-facet Rasch measurement. In G. Fulcher, & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 462-476). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9781003220756

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language testing*, *4*(1), 72-92. https://doi-org.ezproxy.cul.columbia.edu/10.1177/026553228700400107

Raquel, M. (2019). The Rasch measurement approach to differential item functioning (DIF) analysis in language assessment research. In V. Aryadoust, & M. Raquel (Eds.), *Quantitative*

Data Analysis for Language Assessment Volume I (pp. 103-131). Routledge. https://doi-org.ezproxy.cul.columbia.edu/10.4324/9781315187815

Shin, J. Y. (2022). Investigating and optimizing score dependability of a local ITA speaking test across language groups: A generalizability theory approach. *Language Testing*, *39*(2), 313-337. https://doi-org.ezproxy.cul.columbia.edu/10.1177/02655322211052680

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing*, *34*(2), 151-174. https://doi-org.ezproxy.cul.columbia.edu/10.1177/02655322156235

Vafaee, P., & Yaghmaeyan, B. (2015). Providing evidence for the generalizability of a speaking placement test scores. *Iranian Journal of Language Testing*, *5*(2), 78-95.

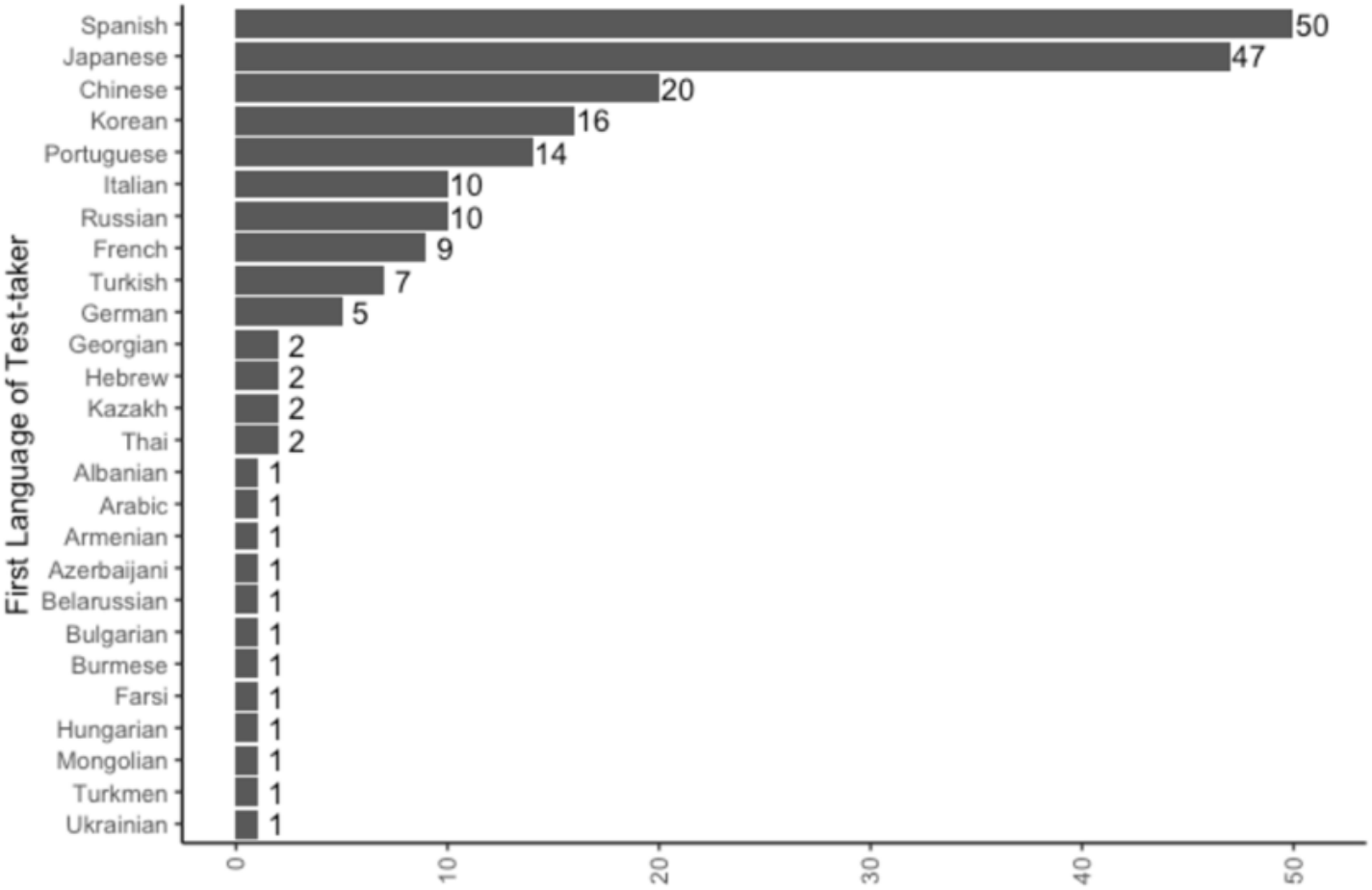
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language testing*, *15*(2), 263-287. https://doi-org.ezproxy.cul.columbia.edu/10.1177/026553229801500205

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

Yan, X., Cheng, L., & Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examinee L1 background on scores of an ITA speaking test. *Language Testing*, *36*(2), 207-234. https://doi-org.ezproxy.cul.columbia.edu/10.1177/02655322187757

APPENDIX A

Bar Plot of First language Background of Examinees (*N* = 208)



Notes:

1. L1 Turkmen speaker also reported Russian as a native language.
2. L1 Mongolian speaker also reported Chinese as a native language.
3. L1 Belarussian speaker also reported Russian a native language.

APPENDIX B

Rating Plan and Samples Scored for each writing task (*k* = 2) (*N* = 208)

	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**
Subset 1	1-23	24-46	47-69	70-92	93-115	116-138	139-161	162-184	185-208
Subset 2	24-46	47-69	70-92	93-115	116-138	139-161	162-184	185-208	1-23
Total	**46**	**44**	**45**	**44**	**44**	**44**	**44**	**45**	**46**
Spanish	11	13	14	10	8	10	10	11	13
Japanese	11	8	9	9	11	9	7	13	17

Note: *Rater Number*, R#, *Subset:* Samples numbers *(1-23) Spanish:* Total samples scored by rater written by L1 Spanish examinees, *Japanese:* Total Samples scored by rater written by L1 Japanese examinees

APPENDIX C

Wright Map Column for Examinee Ability with L1 Background labels

Measr	+Examinee														
7	072_GER														
	085_SPA														
6															
	206_KOR														
	187_KAZ														
5															
	001_SPA	015_SPA	111_JAP												
	007_JAP	023_CHI	073_MON_CHI	149_CHI	154_POR										
4	048_SPA	075_SPA	140_JAP	168_TUR											
	068_FRE	078_JAP	094_JAP	166_CHI	191_JAP										
	018_JAP	026_ITA	039_FRE	046_CHI	106_ITA	131_ARA									
3	003_JAP	010_CHI	019_JAP	027_RUS	028_POR	070_TUR	074_FRE	118_JAP	145_POR	148_THA	150_BUL	151_RUS	164_ITA	174_KOR	
	009_CHI	047_TUR	083_KOR	086_SPA	104_CHI	109_ARM	110_UKR	112_TUR	115_SPA	133_GER	146_SPA	160_POR	181_JAP	195_SPA	201_SPA
:	: 208_SPA														
	044_JAP	057_JAP	060_SPA	066_JAP	082_SPA	093_ITA	159_ITA	165_ITA	171_POR	183_RUS	193_JAP				
2	008_SPA	029_SPA	033_POR	036_FRE	065_SPA	071_JAP	079_KOR	134_GER	176_RUS						
	005_JAP	011_JAP	034_JAP	035_SPA	081_CHI	153_AZE	173_SPA	184_POR	186_JAP	190_SPA	192_JAP				
	025_SPA	038_KOR	043_ITA	087_CHI	090_KOR	092_RUS	096_CHI	132_CHI	172_SPA	198_RUS					
1	004_GER	013_SPA	020_JAP	021_JAP	032_GER	054_FRE	055_KOR	129_SPA	137_BEL_RUS	179_HEB	197_JAP	200_SPA	202_TUR		
	031_SPA	061_SPA	076_HEB	084_SPA	088_CHI	091_SPA	101_SPA	120_SPA	123_ITA	130_CHI	135_CHI	152_JAP	157_KOR	163_POR	175_JAP
:	: 177_POR	203_SPA													
	030_KOR	049_RUS	077_CHI	141_SPA	147_SPA	162_KOR	178_GEO	194_JAP	204_JAP						
* 0	* 040_FRE	052_JAP	053_JAP	058_JAP	062_POR	097_CHI	107_JAP	108_JAP	114_CHI	124_TUR	143_FRE	161_POR			
	016_KOR	017_HUN	037_SPA	041_TUR	056_KAZ	089_JAP	099_FRE	102_JAP	126_SPA	180_CHI					
	006_SPA	012_POR	042_ALB	080_JAP	117_JAP	121_JAP	158_SPA								
−1	002_POR	059_SPA	063_POR	142_ITA	144_CHI	155_JAP	182_JAP	205_JAP							
	051_SPA	128_KOR	167_JAP												
	024_JAP	069_SPA	113_SPA	116_TUR_RUS	196_ITA	207_SPA									
−2	045_BUR	064_SPA	127_RUS	185_JAP											
	125_JAP														
−3	138_JAP	199_FRE													
	100_SPA														
	022_SPA	170_SPA													
−4	169_SPA														
	014_KOR	067_GEO	098_CHI												
	119_KOR	156_SPA													
−5	103_KOR	188_RUS	189_THA												
−6	050_KOR														
−7															
	105_RUS														
−8	095_FAR	122_JAP	136_SPA	139_SPA											
Measr	+Examinee														

Notes: L1 of Examinee denoted by first three letters in Language (e.g, SPA = Spanish)

APPENDIX D

Examinees with Misfitting Profiles

Using Infit MS Thresholds of 0.5 to 1.5 (Eckes, 2019)

Examinees with Underfitting Profiles (n = 27, 13% of total examinees, N = 208)

Examinee	*Observed Average*	*Fair Average*	*Measure*	*Infit MnSq*	*Infit ZStd*
049_Russian	2.75 of 5	2.66 of 5	+ 0.32	**4.10**[a]	**4.4**[a]
098_Chinese	1.33 of 5	1.27 of 5	- 4.30	**3.82**[a]	**4.6**[a]
129_Spanish	3.17 of 5	2.89 of 5	+ 0.92	**3.69**[a]	**4.0**[a]

135_Chinese	3.08 of 5	2.82 of 5	+ 0.73	**3.51**[a]	**3.8**[a]
093_Italian	3.67 of 5	3.59 of 5	+ 2.45	**3.32**[a]	**4.1**[a]
045_Burmese	1.92 of 5	1.90 of 5	- 2.15	**3.00**[a]	**2.9**[a]
002_Portuguese	2.08 of 5	2.26 of 5	- 0.90	**2.27**[a]	**2.1**[a]
088_Chinese	2.83 of 5	2.85 of 5	- 0.81	**2.27**[a]	**2.3**[a]
031_Spanish	2.83 of 5	2.81 of 5	- 0.71	**2.25**[a]	**2.1**[a]
070_Turkish	3.83 of 5	3.86 of 5	+ 2.99	**2.22**[a]	**2.7**[a]
125_Japanese	1.92 of 5	1.74 of 5	- 2.75	**2.06**[a]	1.9
161_Portuguese	2.83 of 5	2.59 of 5	+ 0.11	**2.04**[a]	2.0
157_Korean	3.00 of 5	2.74 of 5	+0.53	**2.02**[a]	2.0
045_ Mongolian*	4.50 of 5	4.81 of 5	+ 4.47	**1.99**[a]	2.0
176_Russian	3.25 of 5	3.35 of 5	+ 1.96	**1.96**[a]	2.0
043_Italian	3.08 of 5	3.06 of 5	+ 1.32	**1.93**[a]	1.9
051_Spanish	2.17 of 5	2.10 of 5	- 1.42	**1.88**[a]	1.7
037_Spanish	2.42 of 5	2.40 of 5	- 0.45	**1.82**[a]	1.7
059_Spanish	2.25 of 5	2.19 of 5	- 1.14	**1.68**[a]	1.4
137_Belarussian**	3.17 of 5	2.89 of 5	+ 0.92	**1.67**[a]	1.4
143_French	2.83 of 5	2.59 of 5	+ 0.11	**1.63**[a]	1.3
069_Spanish	2.08 of 5	2.02 of 5	- 1.71	**1.61**[a]	1.2
181_Japanese	3.58 of 5	3.70 of 5	+ 2.67	**1.58**[a]	1.4
127_Russian	2.17 of 5	1.99 of 5	- 1.84	**1.53**[a]	1.1
112_Turkish	3.75 of 5	3.67 of 5	+ 2.62	**1.52**[a]	1.3
068_French	4.33 of 5	4.04 of 5	+ 3.35	**1.52**[a]	1.3
174_Korean	3.67 of 5	3.79 of 5	+ 2.84	**1.51**[a]	1.3

Note: *Examinee* = Examinee Label, *Observed Avg* = Average Rating on 6-point scale (0-5), *Measure* = Examinee Ability Logit Estimate, *Infit Mnsq* = Unstandardized Infit Mean square value, *Infit ZStd* = Standardized Infit on z-distribution value, ***a* = underfitting**, ***b* = overfitting** (Bond, et al. 2021, p. 242), * = Examinee 045 reported Mongolian and Chinese as L1, ** = Examinee 137 reported Belarussian and Russian as L1

Examinees with Overfitting Profiles (*n* = 34, 16% of total examinees, *N* = 208)

Examinee	*Observed Average*	*Fair Average*	*Measure*	*Infit MnSq*	*Infit ZStd*
152_Japanese	3.00 of 5	2.74 of 5	+ 0.53	**0.14**[b]	**- 3.2**[b]
147_Spanish	2.92 of 5	2.67 of 5	+ 0.32	**0.15**[b]	**- 3.1**[b]
167_Japanese	2.08 of 5	2.15 of 5	- 1.28	**0.25**[b]	**- 2.2**[b]
156_Spanish	1.42 of 5	1.20 of 5	- 4.52	**0.23**[b]	**- 2.8**[b]
113_Spanish	2.08 of 5	2.04 of 5	- 1.65	**0.28**[b]	**- 2.0**[b]
119_Korean	1.42 of 5	1.20 of 5	- 4.52	**0.31**[b]	**- 2.3**[b]
006_Spanish	2.17 of 5	2.35 of 5	- 0.61	**0.33**[b]	- 1.8
191_Japanese	3.83 of 5	4.23 of 5	+ 3.76	**0.34**[b]	**- 2.4**[b]
145_Portuguese	4.17 of 5	3.86 of 5	+ 2.98	**0.35**[b]	**- 2.3**[b]
202_Turkish	2.58 of 5	2.89 of 5	+ 0.91	**0.35**[b]	**- 2.1**[b]
060_Spanish	3.67 of 5	3.56 of 5	+ 2.39	**0.36**[b]	**- 2.2**[b]
022_Spanish	1.33 of 5	1.50 of 5	- 3.55	**0.36**[b]	**- 2.2**[b]
207_Spanish	1.83 of 5	2.05 of 5	- 1.60	**0.37**[b]	-1.6

048_Spanish	4.42 of 5	4.33 of 5	+ 3.98	**0.38[b]**	- 1.9
035_Spanish	3.17 of 5	3.14 of 5	+ 1.51	**0.39[b]**	- 1.8
086_Spanish	3.75 of 5	3.78 of 5	+ 2.83	**0.39[b]**	**- 2.1[b]**
158_Spanish	2.50 of 5	2.29 of 5	- 0.79	**0.40[b]**	- 1.8
194_Japanese	2.42 of 5	2.71 of 5	+ 0.43	**0.40[b]**	- 1.8
041_Turkish	2.50 of 5	2.48 of 5	- 0.20	**0.42[b]**	- 1.7
141_Spanish	2.92 of 5	2.67 of 5	+ 0.32	**0.42[b]**	- 1.6
124_Turkish	2.75 of 5	2.51 of 5	- 0.10	**0.42[b]**	- 1.6
082_Spanish	3.50 of 5	3.53 of 5	+ 2.33	**0.43[b]**	- 1.8
201_Spanish	3.33 of 5	3.74 of 5	+ 2.75	**0.43[b]**	- 1.6
016_Korean	2.25 of 5	2.44 of 5	- 0.33	**0.43[b]**	- 1.5
182_Japanese	2.17 of 5	2.23 of 5	- 0.99	**0.43[b]**	- 1.5
208_Spanish	3.25 of 5	3.65 of 5	+ 2.57	**0.44[b]**	- 1.5
203_Spanish	2.50 of 5	2.80 of 5	+ 0.68	**0.47[b]**	- 1.5
014_Korean	1.05 of 5	1.25 of 5	- 4.36	**0.45[b]**	- 2.0
011_Japanese	3.00 of 5	3.26 of 5	+ 1.76	**0.47[b]**	- 1.4
100_Spanish	1.58 of 5	1.53 of 5	- 3.48	**0.47[b]**	- 1.4
186_Japanese	2.83 of 5	3.17 of 5	+ 1.58	**0.47[b]**	- 1.4
172_Spanish	2.92 of 5	3.00 of 5	+ 1.18	**0.48[b]**	- 1.4
164_Italian	3.67 of 5	3.79 of 5	+ 2.84	**0.49[b]**	- 1.5
142_Italian	2.42 of 5	2.22 of 5	- 1.03	**0.49[b]**	-1.4

Note: *Examinee* = Examinee Label, *Observed Avg* = Average Rating on 6-point scale (0-5), *Measure* = Examinee Ability Logit Estimate, *Infit Mnsq* = Unstandardized Infit Mean square value, *Infit ZStd* = Standardized Infit on z-distribution value, ***a* = underfitting**, ***b* = overfitting** (Bond, et al. 2021, p. 242)

Using Empirical Threshold for Acceptable Fit (e.g., Kondo-Brown, 2002)

Examinee Fit Summary Stats: (*Mean Infit MS* = 0.99, *SD* = 0.64, +/-2 SD = [- 0.29, +2.27])

Table 4
Examinees with Most Underfitting Profiles (*n* = 8, 4% of total examinees, *N* = 208)

Examinee	*Observed Average*	*Fair Average*	*Measure*	*Infit MS*	*Infit ZStd*
049 Russian	2.75 of 5	2.66 of 5	+ 0.32	**4.10[a]**	**4.4[a]**
098 Chinese	1.33 of 5	1.27 of 5	- 4.30	**3.82[a]**	**4.6[a]**
129 Spanish	3.17 of 5	2.89 of 5	+ 0.92	**3.69[a]**	**4.0[a]**
135 Chinese	3.08 of 5	2.82 of 5	+ 0.73	**3.51[a]**	**3.8[a]**
093 Italian	3.67 of 5	3.59 of 5	+ 2.45	**3.32[a]**	**4.1[a]**
045 Burmese	1.92 of 5	1.90 of 5	- 2.15	**3.00[a]**	**2.9[a]**
002 Portuguese	2.08 of 5	2.26 of 5	- 0.90	**2.27[a]**	**2.1[a]**
088 Chinese	2.83 of 5	2.85 of 5	- 0.81	**2.27[a]**	**2.3[a]**

Note: *Examinee* = Examinee Label, *Observed Avg* = Average Rating on 6-point scale (0-5), *Measure* = Examinee Ability Logit Estimate, *Infit MS* = Unstandardized Infit Mean square value, *Infit ZStd* = Standardized Infit on z-distribution value, ***a* = underfitting** (Kondo-Brown, 2002, p. 14).

APPENDIX E

Scale Category Use across Task and Scale *(Percentage and Count by Scale)*

Table 15
Scale Category across Scale and Task – Percentage Category used *(k* = 6)

	Task 1 – A Customer Review			Task 2 – An Argumentative Essay		
Scale Category	C.T1	O.T1	L.T1	C.T2	O.T2	L.T2
0	**0%** (4)	**0%** (4)	**0%** (4)	**2%** (18)	**2%** (18)	**2%** (17)
1	**3%** (18)	**7%** (31)	**3%** (18)	**6%** (23)	**9%** (37)	**6%** (25)
2	**19%** (76)	**24%** (99)	**26%** (104)	**28%** (112)	**33%** (133)	**28%** (156)
3	**33%** (134)	**38%** (153)	**39%** (160)	**36%** (148)	**35%** (142)	**38%** (156)
4	**24%** (97)	**19%** (78)	**21%** (87)	**19%** (78)	**15%** (59)	**19%** (79)
5	**21%** (85)	**12%** (49)	**10%** (41)	**9%** (37)	**6%** (27)	**5%** (22)

APPENDIX F

Examinee Ability Measures across Spanish and Japanese L1 Leaners

First Language	- 7.5 to - 2.1	- 2.0 to + 0.0	+ 0.1 to + 2.0	+ 2.1 to + 9.5
Spanish (n = 50)	8 *(16%)*	9 *(18%)*	20 *(40%)*	13 *(26%)*
Japanese (n = 47)	3 *(6%)*	12 *(26%)*	17 *(36%)*	15 *(32%)*

APPENDIX G

Bias Analysis – L1 and Raters

Two-Way Interaction between L1 Spanish and Japanese Examinee Ability and Rater Severity

Rater	L1	Bias Size	*z*-score	Probability
R1_NNS_Exp *(Severity: +.40)*	Spanish (*n* = 22)	+ 0.04	+ 0.18	.8539
	Japanese (*n* = 22)	+ 0.02	+ 0.13	.8956
R2_NNS_Exp *(Severity: -.48)*	Spanish (*n* = 26)	- 0.14	- 0.78	.4364
	Japanese (*n* = 16)	+ 0.29	+ 1.33	.1908
R3_NNS_Exp *(Severity: -.04)*	Spanish (*n* = 28)	+ 0.14	+ 0.77	.4436
	Japanese (*n* = 18)	- 0.20	- 0.94	.3515
R4_NNS_Exp *(Severity: +.07)*	Spanish (*n* = 20)	- 0.17	- 0.80	.4242
	Japanese (*n* =18)	+ 0.18	+ 0.79	.4303
R5_NNS_Exp *(Severity: -.39)*	Spanish (*n* = 16)	+ 0.21	+ 0.83	.4108
	Japanese (*n* = 22)	- 0.09	- 0.43	.6656
R6_NNS_Nov *(Severity: -.90)*	Spanish (*n* = 20)	- 0.20	- 0.89	.3762
	Japanese (*n* = 18)	- 0.05	- 0.19	.8496
R7_NS_Nov *(Severity: -.39)*	Spanish (*n* = 20)	- 0.22	- 0.98	.3316

	Japanese (*n* = 14)	+ 0.11	+ 0.44	.6610
R8_NS_Nov *(Severity: +.86)*	Spanish (*n* = 22)	+ 0.38	+ 1.92	.0590
	Japanese (*n* = 26)	- 0.22	- 1.53	.1271
R9_NS_Nov *(Severity: +.79)*	Spanish (*n* = 26)	- 0.05	- 0.25	.8001
	Japanese (*n* = 34)	- 0.15	+ 0.16	.3574

Note: *Rater*, R#, *L1 English,* NS = Native, NNS = Non-native, *Experience*, Exp = Experienced, Nov = Novice; *n* = The number of samples scored that were written by L1 Spanish or Japanese examinees

Daniel Eskin is a doctoral student in Applied Linguistics at Teachers College, Columbia University. His research interests focus on areas of second language assessment including assessing second language pragmatics, scenario-based language assessment, and learning-oriented assessment. Correspondence should be sent to his email at dae2129@tc.columbia.edu.