Crowd-Sourcing Human Ratings of Linguistic Production

Scott Crossley (scott.crossley@vanderbilt.edu)
Department of Special Education, 230 Appleton Place
Nashville, TN 37203-5721

Sara Cushing (stcushing@gsu.edu)
Department of Applied Linguistics/ESL, P.O. Box 4099
Atlanta, GA 30302-4099

Kristopher Kyle (kkyle2@uoregon.edu)
Department of Linguistics, Straub Hall 1290 University of Oregon Room #161
Eugene, OR 97403-1290

Scott Jarvis (scott.jarvis@utah.edu)
Department of Languages & Communication, Bldg 255 S Central Campus Dr., Rm 2300
Salt Lake City, UT 84112

Abstract

This study examines the reliability and validity of using two types of crowd-sourced judgments to collect lexical diversity scores. Scaled and pairwise comparison approaches were used to collect data from non-expert Amazon Mechanical Turk workers. The reliability of the lexical diversity ratings for the crowd-sourced raters was assessed along with those from trained raters using a variety of reliability statistics. The validity of the ratings was examined by 1) comparing crowd-sourced and trained ratings, 2) comparing crowd-sourced and trained ratings to ratings of language proficiency, and 3) by using an objective measure of lexical diversity to predict the crowd-sourced and trained ratings. The results indicate that scaled crowd-sourced ratings showed strong reliability in terms of text and rater strata and showed fewer misfitted texts than the trained raters. The scaled crowd-sourced ratings were also strongly predicted by lexical diversity features derived from the texts themselves.

Keywords: language production, expert ratings, crowd-sourced rating, natural language processing

Introduction

Using human judgments of language production to better understand and assess language acquisition, language perception, language proficiency, and language use has a long history in cognitive assessments. The most common approach to collecting human judgments of language phenomena is through expert raters (i.e., raters that have experience in rating, expertise in the subject, and have been trained; Lumley, 2005). Expert raters have been used to assess the quality of language production in terms of writing quality, speaking proficiency, and lexical production (Crossley, Salsbury, McNamara, & Jarvis, 2011; Kim, Crossley, & Kim, 2022). The scores assigned by experts provide evidence to support inferences about cognitive processing (Kim, 2015). However, variance in raters' judgments that may reflect inattention, poor training, or distractions may affect accuracy, which makes assessing rater reliability an important component to help ensure validity in language assessment (Bachman, 2004).

While expert raters are the norm in high-stakes standardized assessments, some scholars have examined the reliability of non-expert ratings (i.e., raters with no experience rating, no expertise in the subject manner, and have not been trained) in providing judgments of language processing, which is a lower-stakes endeavor. While most of these studies have examined human judgments of writing and speaking proficiency broadly (Cumming, 1990; Weigle, 1999), a few have examined specific linguistic constructs such as lexical diversity, which measures the variety of words produced (e.g., Jarvis, 2017; Vanhove et al., 2019). In some cases, non-expert raters have shown strong quantifiable reliability through the use of inferential statistics and probability values (Cumming, 1990). However, qualitative examinations of expert and non-expert raters have demonstrated differences in self-control strategies, editing approaches, and knowledge, indicating potential dissimilarities in the processes used when rendering judgments (Cumming, 1990; Isaacs & Thompson, 2013).

The purpose of this study is to examine alternative, technologically-enhanced methods of collecting non-expert ratings of language production that rely on crowd-sourcing techniques, which allow for a greater quantity of reliable ratings to be collected. Collecting a larger quantity of ratings from a crowd-sourced population may be beneficial in reducing potential sampling bias by collecting ratings from a more generalized participant pool that better reflects the demographic and knowledge base of the community. Using crowd-sourced raters could provide a better diversity of reliable ratings that are less biased leading to cognitive modeling that is more accurate. Additionally, using crowd-source approaches generally speeds up data collection, allowing for quicker feedback to be provided to test-takers, teachers, and administrators, and is more cost-efficient (Litman, Robinson, & Abberbock, 2017). This study,

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

specifically, compares a corpus of transcribed second language (L2) speech samples that were scored by normed, trained raters with expertise in language analytics (henceforth trained raters) for lexical diversity using a ten-point scale. The transcribed samples were also scored by crowd-sourced non-expert raters (henceforth crowd-sourced raters) using two approaches: A ten-point scale and a pairwise comparison approach. The research questions answered in this study are whether trained and crowd-source raters are reliable and whether the judgments they make have internal and external validity.

Study One

The first study assesses the reliability of trained and crowd-sourced raters using Cohen's Kappa Many-Facet Rasch Measurement (MFRM) analyses.

Corpus

Human transcribed speaking samples were selected from a subset of the National Institute of Information and Communications Technology (NICT) Japanese Learner English (JLE; Izumi, Uchimoto, Isahara, 2004) Corpus (n = 250). Each sample was about 200 words taken from an interview conducted between a test taker and a test administrator. Each test-taker was assigned a speaking proficiency level score based on the American Council on the Teaching of Foreign Languages (ACTFL) standard speaking test guidelines given by the test administrator at the time of the interview. We extracted speech sample transcriptions from the JLE that were related to a picture description task, in which the interviewee was asked to construct a story based on information depicted in a picture or a set of pictures.

Lexical Diversity Ratings: Trained Raters

To collect trained scaled ratings, we trained two undergraduate seniors in a Department of Linguistics to assess lexical diversity in texts. We did not consider our raters to be experts because they did not have experience rating, although they did have expertise. Raters were told they would rate the transcribed speech samples on a scale of 1-10 with 1 being a text with low lexical diversity and 10 being a text with high lexical diversity (see Jarvis, 2017). Lexical diversity was defined for the raters as "The variety of word use that can be found in a person's speech or writing." The raters were also given three sample texts that were judged to contain low, average, and high lexical diversity. Raters were trained on 100 unrelated transcribed speech samples. Once acceptable inter-rater reliability was reached (a weighted Kappa > .60), the raters scored the entire set of samples independently. They next collaboratively adjudicated any samples where their score difference was greater than one.

Lexical Diversity Ratings: Crowd-Sourced Raters

To collect crowd-sourced scaled ratings of lexical diversity, we used the Amazon crowd-sourcing platform Mechanical Turk (M-Turk) to recruit crowd-sourced workers from the United States. We collected ratings in two different data collections. The first collected scaled ratings similar to the trained raters. The second used a pairwise comparison approach. For the scaled ratings, crowd-sourced raters recruited through M-Turk were given the same definition of lexical diversity as the trained raters. To ensure they were untrained, the crowd-sourced raters were not given sample texts or any other type of data that could lead to training. The raters would then see a single text sample and be asked to assign it a score using a 1 to 10 scale for lexical diversity. In total, 308 raters completed the task and provided 5,776 judgments of lexical diversity across the 250 texts. For the pairwise ratings, the raters saw ten pairs of texts and were asked to choose which text in each pair had greater lexical diversity. L2 speaking samples were randomly selected to be either Text 1 or Text 2. We recruited 372 crowd-sourced raters who provided 3751 ratings for the 250 texts.

Reliability Metrics

Kappa values were computed for trained raters to examine reliability in the lexical diversity ratings. We also conducted a series of MFRM analyses for the trained raters and the two sets of crowd-sourced raters (scaled and pairwise) using Facets Version 3.83 (Linacre, 2021) to examine reliability in the lexical diversity ratings. MFRM analyses compute the probability of receiving a particular score on a rating scale as a function of both the ability of the candidate (i.e., the lexical production of the interviewee in the JLE corpus) and of the severity of the rater (i.e., rating more severely than other raters). The MRFM analyses can be used to examine how reliably raters analyze separate texts into different levels of lexical diversity, the severity of the raters, the consistency of raters, and the percentage of texts measured accurately. For the crowd-sourced scaled and pairwise raters, the MRFM also examined what percentage of raters can be considered "good faith" raters—i.e., raters whose patterns of scoring are neither too noisy (because they were simply scoring randomly) nor too muted (in the case of scale ratings, often because raters used a very restricted score range).

Four different MRFM analyses were run. Two analyses were run using the two trained raters; one using a rating scale model, which assumes that both raters are using a common rating scale, and the other using a partial credit model, which allows a separate rating scale to be modeled for each rater. Next, an analysis was run using the crowd-sourced scaled data using a rating scale model (that is, with a single rating scale shared by all raters). For the crowd-sourced pairwise ratings, we used the Facets software to conduct a pairwise comparison using the Bradley-Terry-Luch (BTL) model for paired comparisons. Since raters are not judging against a scale (an absolute judgement), but simply comparing two texts (a relative judgment), rater severity is not part of the measurement model, and raters are all assumed to have a severity of 0. However, rater fit statistics can be used to determine whether rater behavior is aberrant, relative to the overall variability in the data.

Results

For the trained raters, the initial Kappa value for inter-rater reliability after independent scoring was $K = .690$, indicating moderate agreement values. The MRFM analysis for text reliability showed the texts were divided into 3.99 levels of lexical diversity with a reliability for .88 (akin to a Kappa value) for the rating scale model and 3.95 levels of lexical diversity with a reliability of .88 for the partial credit model. The crowd-sourced scaled raters were divided into 2.91 levels (reliability = .79) while the crowd-sourced pairwise ratings were divided into 2.28 levels (reliability = .68). Severity for the trained raters using the rating scale model was high (10.79) with strong reliability (.98). The partial credit model reported lower severity (1.98) and lower reliability (.60). Severity for the crowdsourced raters was moderate (4.6) with strong reliability (.91), while severity and reliability for pairwise comparisons could not be calculated (because not all raters rated the same text comparisons).

To measure consistency, we report an infit statistic, which is an "information-weighted, inlier-pattern-sensitive, mean square fit statistic with expectation 1 and range 0 to infinity" (Linacre, 2021). Scores lower than 1 indicate too little variation or lack of independence in ratings, while scores higher than one may indicate noise or unmodeled excessive variation. For the purposes of this paper, we consider misfit to be infit scores greater than 1.5, and overfit to be scores lower than .5 (McNamara, Knoch, & Fan, 2019). Misfit for trained raters was reported in 64 of the 250 texts for the rating scale model and 45 texts for the partial credit model. Lower misfit was reported for the crowd-sourced scaled ratings (26 texts) and there were no misfitting texts with the crowd-sourced paired comparisons. For rater consistency, there was more misfitting (45 or 10%, vs. 37 or 10%) in crowd-sourced scaled raters versus crowd-sourced pairwise raters and overfitting as well (51 or 16% vs. 7 or 2%), resulting in a smaller percentage of raters behaving as independent experts (204, 66.2% versus 328, 88.2%).

Discussion

Inter-rater reliability as assessed by Kappa values indicated that the trained ratings showed moderate reliability (McHugh, 2012). The MRFM rating scale model for the trained raters showed strong separation of texts with high reliability for both the rating scale and partial credit models. The rating scale model reported extreme differences between the two raters in terms of severity with very high reliability, the opposite of what one hopes for among raters. The rating scale analysis also revealed that slightly over 25% of the texts received ratings that were misfitting, suggesting that these texts were not rated consistently between the two raters. Results were better for the partial credit model in terms of rater severity, but reliability was low. Misfit texts were also lower for the partial credit model, but still high at 18%. The high proportion of misfitting texts with trained raters suggests that either the training was insufficient to ensure high levels of agreement or, perhaps, that the 1-10 scale was too broad for raters to make reliable distinctions. Overall, the analysis indicates that the trained raters showed differences in severity and that their ratings led to many misfitted texts.

For the crowdsourced ratings, the MRFM analysis indicated that both scaled and paired comparisons resulted in fewer reliably distinguishable levels of lexical diversity. The rating scale model reported moderate differences between the crowd-source scaled raters in terms of severity with high reliability. The crowd-sourced scaled ratings led to 26 texts (10.4% of texts) that were misfit. The crowd-sourced pairwise scores reported no misfit texts. For rater consistency, the pairwise task led to few misfit raters. Overall, the findings indicate consistency in the crowd-sourced ratings, but hint that raters need to be pruned.

Study Two

The second study examines the internal and external validity of the human rating of lexical diversity by first conducting correlations among the ratings provided by trained and crowd-sourced raters. Second, correlations between ratings of lexical diversity and the speaking proficiency assigned to each JLE speaking sample were calculated. Lastly, linear models were developed to predict the human ratings using features related to lexical volume, abundance, and variety.

Human Ratings for Lexical Diversity

Trained Raters We calculated the average score after adjudication between the two normed raters as our measure of trained lexical diversity.

Crowd-Sourced Ratings (Scaled) To calculate reliable scaled scores, a follow-up Facets analysis was conducted after removing ratings from raters that met one or more of the following conditions: (a) they scored fewer than 5 texts; (b) they had infit scores greater than 1.5 and/or discrimination indices lower than -.50; (c) they only used the highest or lowest scores. Forty-four raters were removed in this way, leaving 254 for this analysis. The Facets analysis of the texts using these selected raters indicated the samples were separated into three stratas, $\chi 2\ (249) = 1428.300$, $p < .001$, with a reliability of .84. For each text, the Facets analysis provides an observed average score on the original rating scale, along with a fair average, which is the predicted average score if the text were rated by raters of average severity. That is, the fair average is adjusted to account for the severity of the actual raters who provided the scores.

Crowd-Sourced Ratings (Pairwise) From the original paired comparison Facets analysis, we removed 67 raters who either had completed fewer than five comparisons or had infit scores greater than 1.5 and/or discrimination indices lower than -.50. We conducted a follow-up Facets analysis of the text using these selected raters, which indicated the samples were separated into three stratas, $\chi 2\ (249) = 849.300$, $p < .001$, with a reliability of .76. We used a Bradley-Terry model (Bradley & Terry, 1952) to calculate pairwise comparison scores for lexical diversity for these raters. The Bradley-Terry model ranked each sample by lexical diversity complexity

based on each sample's probability to be rated more highly than other samples based on the crowd-sourced ratings. The model creates a maximum likelihood estimate which iteratively converges towards a unique maximum that defines the ranking of the sample (i.e., the samples with the highest lexical diversity have the highest probability). The probability values are then used as estimates of lexical diversity. We used the Python package *choix* to calculate Bradley-Terry models. After calculating the Bradley-Terry model, visual analyses indicated that the scores for two samples were obvious outliers. These were removed from the data analyses leaving a final sample of 248.

Calculations of Volume, Abundance, and Variety

We used the Tool for the Automatic Analysis of Lexical Diversity (TAALED, Kyle et al., 2020) to calculate three dimensions of lexical diversity for each JLE speaking sample: volume, abundance, and variety. Volume was measured through the number of word tokens while abundance was measured using number of types. We selected the Measure of Textual Lexical Diversity (MTLD; McCarthy & Jarvis, 2010) for all words in each sample to measure variety. MTLD is a type-token ratio (TTR) measure that is stabilized for text length.

Statistical Analyses

Bivariate Pearson correlations were run using the cor.test() function in R 3.6.1 (R core team, 2016) between all ratings of lexical diversity to examine similarities among the different rating methods. Similar correlations were run between all ratings of lexical diversity and the speaking proficiency scores assigned to each JLE sample to examine associations. To investigate predictive validity, we used ten-fold cross-validated stepwise linear models to predict human ratings of lexical diversity using indices from TAALED. To do this, we used the leapSeq function in Leaps and Caret (Kuhn et al., 2020). We also ran simple linear regression models using the lm() function (R core team, 2016) so that we could compute the relative importance of the indices in each model using the calc.relimp() function in the relaimpo package (Grömping, 2007).

Results

Correlations among Lexical Diversity Ratings Pearson product moment correlations were conducted between the trained scaled, the crowd-sourced scaled and scaled fair, and the crowd-sourced pairwise rankings for the 248 texts included in the Bradley-Terry modeling (recalling that two texts were remove as outliers in the Bradley-Terry model). All correlations were significant ($p < .001$) and showed strong effects. The strongest correlation was reported between the crowd-sourced scaled and scaled fair ratings (see Table 1).

Correlations between Lexical Diversity Ratings and Speaking Proficiency Ratings Pearson product moment correlation tests were conducted between the human scores of lexical diversity, and the speech proficiency scores given to the 248 participants in the subsample of the JLE that included all human ratings of lexical diversity. All correlations were significant ($p < .001$), met assumptions, and showed strong effects. The strongest correlation was reported for crowd-sourced scaled (fair) scores and the lowest was reported for the trained scaled ratings (see Table 2).

Table 1: Correlations among lexical diversity ratings

Variable	2	3	4
1. TSR	0.645	0.779	0.717
2. CPR	1	0.735	0.702
3. CSFR		1	0.935
4. CSR			1

TSR = trained raters, CPR = Crowd-sourced pairwise ratings, CSFR = Crowd-sourced scaled fair ratings, CSR = Crowd-sourced scaled ratings

Table 2: Correlations between lexical diversity ratings and speaking proficiency ratings

Variable	Speaking Proficiency
1. TSR	0.533
2. CPR	0.568
3. CSFR	0.631
4. CSR	0.609

Linear Models to Predict Human Ratings Four linear models were developed to predict the four different lexical diversity ratings (i.e., trained scaled, crowd-sourced scaled, crowd-sourced scaled fair and crowd-sourced pairwise). In all cases, there was strong multicollinearity between number of token and number of types, and number of tokens was removed, and number of types was retained (based on the strength of correlation with trained raters scores). In all models, VIF values indicated no problems with multicollinearity (all < 2), and visual and statistical examinations of the residuals indicated they were normally distributed.

For the trained scaled ratings, the ten-fold stepwise linear model indicated that the best tuned model included both number of types and MTLD. This model reported an R^2 = .747 and RMSE = .745. These two variables were used in a simple linear model that indicated a significant relationship with a large effect, F (2, 245) = 349.2, $p < .001$, $R^2_{adjusted}$ = .738), explaining approximately 74% of the variance in LD scores. The model parameters are summarized in Table 3. The relative importance metrics indicate that approximately 82% of the explained variance can be attributed to the single abundance measure (number of types), while approximately 18% of the explained variance can be attributed to lexical variety (MTLD).

For the crowd-sourced pairwise ratings, the ten-fold stepwise linear model indicated that the best tuned model included both variables. This model reported an R^2 = .498 and

RMSE = .694. These two variables were used in a simple linear model that indicated a significant relationship with a large effect, F (2, 245) = 112.100, $p < .001$, $R^2_{adjusted} = .474$), explaining approximately 47% of the variance in LD scores. The model parameters are summarized in Table 3. The relative importance metrics indicate that approximately 73% of the explained variance can be attributed to the single abundance measure (number of types), while approximately 27% of the explained variance can be attributed to MTLD.

For the crowd-sourced fair ratings, the ten-fold stepwise linear model indicated that the best tuned model included both variables. This model reported an $R^2 = .609$ and RMSE = .477. The two variables were used in a simple linear model that indicated a significant relationship with a large effect, F (2, 245) = 200.100, $p < .001$, $R^2_{adjusted} = .617$), explaining approximately 62% of the variance in LD scores. The model parameters are summarized in Table 3. The relative importance metrics indicate that approximately 68% of the explained variance can be attributed to the abundance measure (number of types), while the remaining variance (approximately 32%) of the explained variance can be attributed to the lexical variety variable (MTLD).

Table 3: Models to predict ratings of lexical diversity

Ratings	Relative Importance	Estimate	*t*
TSR			
Intercept		4.462	98.147**
Type count	0.825	1.224	22.61**
MTLD	0.175	0.106	1.977*
CPR			
Intercept		0.038	0.856
Type count	0.728	0.554	10.727**
MTLD	0.272	0.182	3.519**
CSFR			
Intercept		4.585	151.978**
Type count	0.682	0.475	13.429**
MTLD	0.318	0.209	5.911**
CSR			
Intercept		4.727	155.54**
Type count	0.671	0.4	11.315**
MTLD	0.329	0.187	5.275**

** $p < .001$, * $p < .050$

For the crowd-sourced scaled ratings, the ten-fold stepwise linear model indicated that the best tuned model included both variables. This model reported an $R^2 = .546$ and RMSE = .476. The two variables were used in a simple linear model that indicated a significant relationship with a large effect, F (2, 245) = 146.400, $p < .001$, $R^2_{adjusted} = .541$), explaining approximately 54% of the variance in LD scores. The model parameters are summarized in Table 3. The relative importance metrics indicate that approximately 67% of the explained variance can be attributed to the abundance measure (number of types), while the remaining variance (approximately 33%) of the explained variance can be attributed to the lexical variety variable (MTLD).

Discussion

Study two reported strong correlations ($r > .650$) between all the different ratings of lexical diversity with the lowest correlations between crowd-sourced pairwise ratings and trained scaled ratings. As expected, the strongest correlation was reported between crowd-sourced scaled (fair) and crowd-sourced scaled ($r = .935$). The strength of the relationships indicates that raters in each approach are likely scoring lexical diversity in a similar manner.

We also examined correlations between human ratings of speaking proficiency from the JLE samples and the human ratings of lexical diversity. In all cases, the correlations were strong and positive indicating that higher lexical diversity scores given by raters were related to greater speaking proficiency scores. The correlations indicated that crowd-sourced scaled (fair) and crowd-sourced scaled scores reported the highest correlations ($r > .60$). The lowest correlation was reported for trained scaled scores ($r = .53$).

We also assessed correlations among the human scores for lexical diversity with computational measures of volume (tokens), abundance (types), and variety (MTLD). For all human ratings, we found strong correlations between judgments of lexical diversity and the number of tokens and types as well as MTLD scores. Correlations indicated that number of types showed the strongest associations with human ratings of lexical diversity followed by number of tokens and then MTLD for all types of human ratings.

Noting strong correlations between number of types and tokens for all the human ratings, all linear models to predict human ratings were developed using only number of types and MTLD scores. In all the models, the strongest predictor according to the variance explained was number of types (i.e., abundance) followed by MTLD scores (i.e., variety). The amount of variance explained by the models ranged from 74% (for the trained scaled ratings) to 47% (the crowd-sourced pairwise scores). The relative importance measures indicated that, in all cases, the number of unique types explained the greatest amount of variance in the lexical diversity scores. Thus, raters were more likely influenced by number of unique types when assigning lexical diversity scores than lexical variety such that samples with more unique words were given higher lexical diversity scores. The number of unique word types seemed to have the lowest impact on crowd-source scaled raters. In terms of MTLD (i.e., variety), all models had positive coefficients for MTLD scores, indicating that greater lexical variety led to greater lexical diversity scores. MTLD explained the most variance according to relative importance scores for the crowd-sourced scaled scores and the lowest amount of variance for the trained scaled scores, indicating that that crowd-sourced

raters may have focused more strongly on lexical variety than trained, trained raters.

Overall, the results seem to provide stronger validation of the lexical diversity scores reported by the trained scaled ratings and the crowd-sourced scaled (fair) ratings. The crowd-sourced scaled (fair) ratings reported the highest correlation with speaking proficiency and the trained scaled ratings reported the strongest associations with computational indices of lexical diversity. As well, lexical diversity indices predicted the greatest amount of variance in the trained scaled ratings, but most of this variance related to type counts. The weakest results appear to be reported for the crowd-sourced pairwise scores which demonstrated the lowest correlations with speaking proficiency scores and the least amount of variance in the linear model predicting the lexical diversity scores.

Conclusion

The results from this data indicate that crowd-sourced scaled ratings showed strong reliability in terms of text and rater strata with relatively low misfitted texts. The crowd-sourced scaled (fair) ratings also showed the strongest associations with speaking proficiency scores for the speech samples and the scores were strongly predicted by lexical diversity features derived from the texts. Also, unlike the trained rater scores, the crowd-sourced scaled (fair) ratings were not as strongly predicted by abundance measures.

This study provides evidence that collecting a large quantity of cognitive ratings from a crowd-sourced population may be advantageous in low-stakes data collection like assessing lexical diversity. Compared to a simple panel of two trained raters, a crowd-sourcing approach that collects data from a more diverse participant pool that better represents the demographic and knowledge base of English language users may reduce sampling bias and lead to more reliable ratings. Of course, not all crowd-sourced raters were reliable, and some measures had to be taken to prune the participants to increase reliability. However, with proper measures in place, a crowd-sourcing approach may prove to be as reliable as smaller panels of raters. As well, a crowd-sourcing approach affords faster data collection that is more cost-efficient. We can envision crowd-sourcing assessments being used for low stakes assessments that measure language proficiency, language production, or text complexity. While opportunities may exist for using crowd-sourced raters in high stakes assessment, more research would need to be done in terms of norming or training raters and assigning tasks that deal with more specific aspects of language assessment.

Future work should focus on collecting human ratings from a larger sample of trained raters to ensure bias did not exist within the single pair of raters used here. Additionally, the approaches used need to be replicated on language samples other than second language speech transcriptions.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learners using computational indices. *Language Testing, 28* (4), 561-580.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31-51.

Grömping, Ulrike. "Relative importance for linear regression in R: the package relaimpo." *Journal of statistical software* 17 (2007): 1-27.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management, 12*(2), 119-125.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553.

Kim, H. J. (2015) A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment, Language *Assessment Quarterly, 12*(3), 239-261.

Kim, M., & Crossley, S. A., & Kim, B. (2022). Second language reading and writing in relation to first language, vocabulary knowledge, and learning backgrounds. *International Journal of Bilingual Education and Bilingualism*, *25(6)*, 1992-2005

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., & Benesty, M. (2020). Package 'caret'. *The R Journal, 223*.

Linacre, J. M. (2021). *A User's Guide to FACETS Rasch-Model Computer Programs Program Manual* 3.83.5.

Litman, L., Robinson, J. & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49, 433–442.

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

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment: The role of measurement.* Oxford University Press.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vanhove, J., Bonvin, A., Lambelet, A., & Berthele, R. (2019). Predicting perceptions of the lexical richness of short French, German, and Portuguese texts using text-based indices. *Journal of Writing Research, 10(3*), 499-525.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–78.