Creative Education, 2019, 10, 2859-2868
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Brunswik's Lens Model: This Is How to Inspire Accurate Raters

Muhamad Firdaus bin Mohd Noh, Mohd Effendi Ewan bin Mohd Matore

Centre of Educational Planning and Policy, Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Selangor
Email: muhamad.firdausi@gmail.com

How to cite this paper: Noh, M. F. M., & Matore, M. E. E. M. (2019). Brunswik's Lens Model: This Is How to Inspire Accurate Raters. *Creative Education, 10,* 2859-2868.
https://doi.org/10.4236/ce.2019.1012212

Received: October 21, 2019
Accepted: November 26, 2019
Published: November 29, 2019

Abstract

Rating accuracy is one of the fundamental standards in educational assessment to ensure the quality and integrity. Inaccuracy in academic assessment engenders negative implications towards student's motivation and raters' credibility. Therefore, this paper seeks to provide a discussion on rating accuracy in educational assessment based on Brunswik's lens model. The model contends that raters' ratings are not completed directly but through the existence of many factors including raters' variability, rating scales and domains assessed. Raters' idiosyncrasy is scrutinized by describing varied sources that can threaten rating accuracy. This model explains how intervening factors moderate the relationship between candidates' capabilities and observed scores. The discussion may shed some light on the endeavors to inspire raters to be effective and uphold the values of reliable raters through the implementation of thoughtful rater training that incorporates scoring practices, exposure on rater bias and self-directed reflection. Future attempts are necessary for understanding the interaction among intervening factors that influence raters and differences of rating accuracy produced by internal and external raters.

Keywords

Rater Accuracy, Brunswik's Lens Model, High-Stakes Assessment, Quality Raters

1. Introduction

The challenge of any performance assessment lies on its ability in assuring students' scores to comply with standard measurement properties (Sundqvist, Wikström, Sandlund, & Nyroos, 2018). These properties refer to validity, reliability, accuracy and fairness maintained throughout the assessment procedures

(AERA, APA, & NCME, 2014). Validity is the extent to which the test realizes the intention of the test developer and how evidence is in support of the interpretation of candidates' marks (Baksh, Sallehhudin, & Hamin, 2019). It is achieved when items used intentionally measure what it purports to measure as described in the purpose of the assessment and thus can be trusted to categorize candidates based on their competencies in the measured domains. Reliability relates to how scores obtained by candidates are consistent even though the assessment occurs at different times, places and conditions (Coaley, 2009). Fairness in assessment is evident when assessment procedures do not give threat or create irrelevant variance towards any individuals in the intended population of candidates throughout all phases of assessment development, administration and interpretation. It will ensure that none of the candidates are assessed with bias, a major threat to fairness in assessment (Engelhard, Wang, & Wind, 2018) and all candidates are given equitable opportunity to manifest their capabilities. These properties are important factors in promising standards of education assessment especially for high-stakes testing as candidates' life-changing decisions hinged mainly on their performance in the test (Ameen, Sallehhudin, Kemboja, & Baksh, 2014). While maintaining these three factors in objective items appears to be easily monitored, that is not the case for subjective items as human raters are assigned to score candidates' answers. Subjective items scored subjectively by human raters give chance for irrelevant variances to jeopardize the three factors.

2. Brunswik's Lens Model

Lens model by Brunswik (1956) can be utilized in comprehending rating process involving human raters. This model accentuates indirect relationship between candidates' capabilities and scores given by raters. Raters cannot directly observe and assess candidates due to the existence of intervening factors including domains assessed, rating scales and raters' variability. In rater-mediated assessment, the model offers a theoretical framework to discuss the relationship between a latent variable, students' performance and an observed variable, score finalized by raters. Subjective items which require candidates to construct their own answers can be rated objectively based on rigid answer schemes or subjectively when raters are given flexibility (Haladyna & Rodrigues, 2013). In subjectively scoreditems, raters are normally provided with rubric, scoring procedures, marking criteria, general criteria of accepted answers and sometimes exemplar answer samples. It renders the rating process to be complex as candidates' answers are scored with the aid of varied cues and information. In addition, raters need to be able to recognize whether potential factors affecting their assessment is relevant or irrelevant (Engelhard et al., 2018) to discriminate candidates in the assessed domains fairly. These factors may include candidates' use of language embellishment (Saadat & Alavi, 2018), context of the assessment, nature of the items, layout of candidates' answers (Cooksey, Freebody, & Wyatt-Smith, 2007),

raters' professionalism, characteristics of the testing such as topics covered (Südkamp, Kaiser, & Möller, 2012), and their performance in previous testing (Oudman, Pol, Bakker, Moerbeek, & Gog, 2018). The complexity of rating process is well-explained in the model as it delineates how an individual makes judgment upon what he or she observes as well as the process he or she engages in while making the judgment (Wind, Stager, & Patil, 2017). **Figure 1** shows the components in the lens model encompassing the process of how students' performance is assessed by raters through cues and raters' variability. The cues include how raters understand the domains assessed and rating scales, while raters' variability covers their background, professionalism and severity (Hammond, 1996). Thus, the quality of assessment is determined by the difference between candidates' real performance and marks obtained by candidates (Vögelin, Jansen, Keller, Machts, & Möller, 2019).

3. Rater Accuracy

Brunswik's lens model highlights raters' accuracy as one of the intervening factors towards the relationship between candidates' capabilities and marks they receive. Rater accuracy refers to a kind of effect brought by raters resulting in candidates receive marks that do not reflect their true abilities (Scullen, Mount, & Goff, 2000). It is raters' tendency to score students' answers far from the expected marks. Rater accuracy can be observed through five indicators which are severity/leniency, central tendency, randomness effect, halo effect and differential severity/leniency (Myford & Wolfe, 2004) (**Figure 2**).

Severity is raters' inclination to generate marks lower than other raters towards the same candidates. A severe rater underestimates candidates' capability by providing lower marks as compared to their warranted performance (Noor Lide, 2011). On the contrary, leniency is problematic as candidates are given higher marks than their true performance when a lenient rater scores their answers (Wind, 2018). Failure to manage severity and leniency will contribute to construct-irrelevant variability that can give impact on candidates' success in the assessment. Raters' ability to demonstrate average severity consistently will enable candidates of the same competencies to be awarded with similar scores. It is

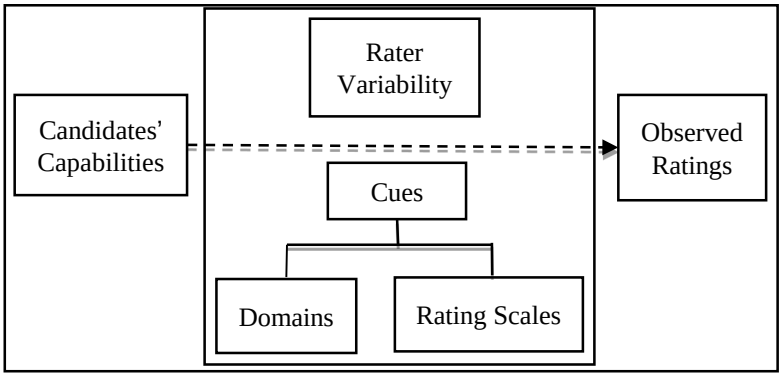

Figure 1. Brunswik's lens model as adapted from Engelhard and Wind (2018).

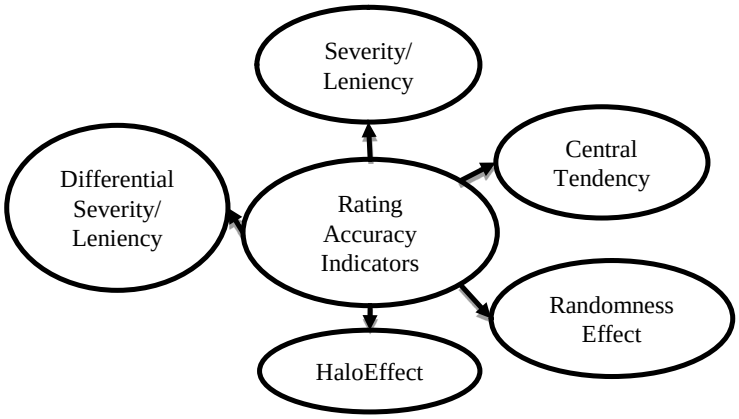

Figure 2. Rating accuracy indicators.

intolerable if two candidates of the same competencies receive different scores because their answers are rated by raters of dissimilar severity level. Next, central tendency happens when raters are attracted to use only a subset of rating point scales though candidates' performances may vary across the range of the scales (Wind & Schumacker, 2017) due to raters overuse central categories of ratings scales too frequent that they ignore the highest or lowest categories of the score points. Avoiding extreme score is raters' strategy in ensuring they are not considered as an outlier and distinct from other raters and perhaps to play safe (Wu, 2017). By doing so, they will not be questioned of why they score candidates with too high or too low marks and consequently challenges the fairness in rating as excellent candidates are given low marks and less capable candidates manage to get high marks even though they do not exhibit outstanding performance. Such unfair judgment will imperil rating accuracy and may as well reduce candidates' motivation.

Then, raters' failure to be consistent in applying one or more rating scales in comparison to other raters departs from randomness effect (Eckes, 2015). In ranking candidates based on their performance, raters who demonstrate randomness effect tend to rank candidates in contrasting order than the other raters (Myford & Wolfe, 2004). Raters' propensity to rate with randomness may anticipate that they are unable to discriminate candidates based on their ability well and their inability to understand how one should rate candidates' work. Halo effect occurs when raters are inclined to generate equal marks to every domain assessed for a candidate and unable to make fine discrimination among the domains especially when they are tasked to rate using analytical scoring. It is also evident when a particular domain gives influence on another domain either resulting in higher or lower total marks (Bijani, 2018). It is the consequences of raters allowing their global conception of candidates to domineer their judgment and ignore the fact that candidates may be heterogeneous in their mastery of different domains (Hennington, Bradley, Crews, & Hennington, 2013). Raters' inclination towards halo effect engenders inability to segregate candidates based on their competencies in accordance to different domains. Correspondingly,

candidates may be awarded with different marks if they are to be rated using analytical scoring and will challenge the accuracy of raters' rating. On top of that, candidates cannot be guided in the domains that in actual they have not mastered.

Finally, differential severity/leniency leads to awarding certain groups of candidates with higher or lower marks than they are entitled according to measurement principles. This phenomenon takes place when raters are too severe or too lenient in assessing a certain group of candidates but manage to control their severity/leniency when marking other groups of candidates (Engelhard, 2007). It challenges rater accuracy standard because each candidate should be judged solely on their capabilities using the guidelines provided.

4. Factors Influencing Rater Accuracy in Language Testing

A growing body of literature has investigated factors influencing raters in scoring candidates' works. The employment of human raters to score students' answers is ubiquitous in any assessment setting either to be used for formative or summative purpose. It is exceptionally common in high-stakes assessment of language subjects to judge students' production in speaking or writing test.

Firstly, the effect of raters' first language has proven to be inconsistent. Zhang & Elder (2014) learned that there is no significant difference between ratings done by native speaker raters and non-native speaker raters. Both raters' inter-rater and intra-rater reliability achieve the same standards. On the contrary, Jabeen (2016) discovered that inter-rater reliability among non-native raters are too low in comparison to native speaker raters.

Secondly, raters' familiarity towards candidates also shows contradictory findings. Winke & Gass (2013) who assigned raters that have ESL candidates' first languages as their second languages figured out that such raters tend to be bias and compromise their ratings' reliability. Lee (2017) concurred that raters who are familiar with candidates' first languages score candidates with higher marks as compared to other raters who are not familiar of candidates' first languages. However, Zhao (2017) observed the opposite when raters regardless of their familiarity with candidates' first language generate indifferent ratings.

In terms of raters' rating experience, Isaacs & Thomson (2013) reported that experienced raters obtained higher level of agreement among them as against novice raters. Groups of raters with different level of rating experience can generate the same quality of ratings when a proper training is provided (Attali, 2016). By assigning experienced raters to rate in four different scoring sessions with training provided prior to each session, Davis (2016) discovered that raters exhibited invariable pattern of ratings throughout the sessions. It is argued that rating experience is more powerful in determining raters' ability to attain rating accuracy.

Investigating on the influence of raters' teaching experience, Hsieh (2011) concluded that teachers were more severe especially in assessing domains like

"accentedness" and "comprehensibility" as compared to non-teachers. Conversely, Song, Wolfe, Less-Petersen, Sanders, & Vickers (2014) observed differences of ratings between teachers and non-teachers were very minimal and not significant. Meanwhile, Zhao (2017) who has grouped study samples based on the length of teaching service years reported that experienced teachers were more severe towards scoring specification as opposed to novice teachers. More recently, Weilie (2018) corroborated findings by Song et al. (2014) by discovering the same outcomes. Contradictory findings from the studies discussed above propose that teaching experience might be a valid factor in establishing accurate rating among raters.

Finally, examining the effect of rater training, Kim (2015) has adopted three groups of raters; novice raters, intermediate raters and experienced raters and assigned them to rate in three scoring sessions using analytical scoring with rater training provided prior to each scoring sessions. Initially, experienced raters performed better than the other two groups of raters, but all the rater groups eventually managed to achieve targeted accuracy after the third scoring session. Meanwhile, Davis (2016) employed 20 experienced teachers without rating experience to receive rater training and score candidates' answer samples. It was learned that raters' severity remained consistent before and after the training yet level of agreement among raters improved at the end of the scoring sessions. Bijani (2018) and Huang, Kubelec, Keng, & Hsu (2018) have chosen samples from experienced raters and inexperienced raters. Both studies concluded that the two groups of raters were able to attain the same standard of inter-rater reliability after rater training was given. Results from the mentioned studies suggested that rater training is influential in producing quality ratings especially when raters have no or limited experience in rating.

5. Rater Accuracy in Malaysian Education System

In Malaysian context, rater accuracy is much related to the recent revamp of teachers' role in high-stakes assessment who are now assigned to be internal raters for their own students in Pentaksiran Tingkatan Tiga (PT3) starting from 2014. The decision has since brought about changes as the assessment for lower secondary school students is no longer centralized nationally but managed respectively by schools. Thus, schools now have the sole authority in choosing question sets, administering assessment-taking, marking and scoring students' answers and finally reporting students' results. Even though PT3 is managed by schools, the result is still used for high-stakes purposes such as to stream students for their streamlines in the upper secondary schools. Therefore, turning a blind eye to PT3 raters' accuracy is not an option.

It is argued that rating accuracy is questionable when internal raters are appointed among candidates' own teachers (Sundqvist et al., 2018) because teachers have long played the role of knowledge disseminators who make daily interaction with their students. The nature of teachers' every-day tasks in classrooms

to enhance students' learning and monitor their progression towards learning objectives is contradictory to the task of a rater who needs to judge and decide students' future. Thus, how do we ensure teachers play both roles as knowledge disseminator and internal raters well? Interestingly, such a conundrum can be well-explained by Brunswik's lens model. While teachers continue to function as students' teachers in classrooms as usual, awareness about elements discussed in the model will enable teachers to become a quality rater. If teachers can comprehend that the process of rating students' answers in assessment is indirect and their judgment is accomplished with the existence of many factors, accuracy in their ratings may be achieved. Consequently, even though PT3 is scored by candidates' teachers, the standard for a high-stake assessment can be reached.

6. Implication and Conclusion

Variability among raters impacting rater accuracy is well explicated in the lens model. The model signifies that candidates' performance is not directly observed as raters are also influenced by raters themselves, assessed domains and rating scales. Factors rooted from raters can significantly impact the overall marks received by candidates. The factors include raters' first language, raters' familiarity with candidates' language, rating experience, teaching experience and rater training. Even though rating error is inevitable, efforts need to be made to minimize such errors. Teacher raters need to meet up the standard of a good rater by considering factors such as teaching experience and rating experience. Next, rater training should be designed to yield raters with good exposure, understanding, determination and rating practice. Apart from including explanation about scoring scales, rubric, nature of the assessment items and scoring procedures, raters need to be exposed to potential factors that can risk their rating accuracy. They need to be able to discriminate aspects that should and should not be considered when scoring. Rater training should incorporate rating practice and feedback provision so that they are able to apply the how-to in operational scoring. Consideration should also be given to provide raters with basic knowledge on how to analyze their own ratings for them to avoid rating errors.

The earlier discussion has reached corroboration that it is impossible to eliminate the fact that raters come to scoring scenes with their own background, professionalism and variability. Hence, the first step to manage their idiosyncrasy and mitigate rater effects is to acknowledge that raters are heterogeneous rather than struggling to achieve fictitious homogeneity. Then, raters should not allow themselves to be dominated by their own variability while scoring. Also, the education authority needs to focus on the appointment process of raters, the quality of rater training provided to raters, provision of continuous mentoring, feedback and guidance to raters, administration of monitoring programs and analysis of raters' performance periodically. The discussion on rater accuracy should spark researchers' interest to compare the quality of internal versus external raters. Variation among raters needs to be explored in relation to how

they lead to rater accuracy especially in a specific assessment setting. Contradictory findings from existing studies on factors affecting rater accuracy may suggest that a new research is needed to confirm the interaction among the factors. A mixed method research is also required to understand raters' cognitive process while scoring by combining statistical analysis and qualitative method such as think-aloud protocols and stimulated recalls.

Acknowledgements

This research was supported by the Universiti Kebangsaan Malaysia under the Dana Penyelidikan FPend code GG-2019-034 and PP-FPEND-2019.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing: National Council on Measurement in Education.* Washington DC: American Educational Research Association.

Alla Baksh, M. A. K., Mohd Sallehhudin, A. A., & Siti Hamin, S. (2019). Examining the Factors Mediating the Intended Washback of the English Language School-Based Assessment: Pre-Service ESL teachers' Accounts. *Pertanika Journal of Social Sciences and Humanities, 27,* 51-68. http://www.pertanika.upm.edu.my/Pertanika%20PAPERS/JSSH%20Vol.%2027%20(1)%20Mar.%202019/4%20JSSH-2622-2017.pdf

Attali, Y. (2016). A Comparison of Newly-Trained and Experienced Raters on a Standardized Writing Assessment. *Language Testing, 33,* 99-115. https://doi.org/10.1177/0265532215582283

Bijani, H. (2018). Investigating the Validity of Oral Assessment Rater Training Program: A Mixed-Methods Study of Raters' Perceptions and Attitudes before and after Training. *Cogent Education, 33,* 1-20. https://doi.org/10.1080/2331186X.2018.1460901

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments.* Berkeley, CA: University of California Press.

Coaley, K. (2009). *An Introduction to Psychological Assessment and Psychometrics.* Los Angeles, CA: SAGE.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as Judgment-in-Context: Analyzing How Teachers Evaluate Students' Writing. *Educational Research and Evaluation, 13,* 401-434. https://doi.org/10.1080/13803610701728311

Davis, L. (2016). The Influence of Training and Experience on Rater Performance in Scoring Spoken Language. *Language Testing, 33,* 117-135. https://doi.org/10.1177/0265532215582282

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (2nd ed.). New York: Peter Lang.

Engelhard, G., Wang, J., & Wind, S. A. (2018). A Tale of Two Models: Psychometric and Cognitive Perspectives on Rater-Mediated Assessments Using Accuracy Ratings. *Psychological Test and Assessment Modeling, 60,* 33-52.

http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/3_PTAM_Engelhard__Wang__Wind__2018-03-10__1855.pdf

Haladyna, T. M., & Rodrigues, M. C. (2013). *Developing and Validating Test.* New York: Routledge. https://doi.org/10.4324/9780203850381

Hennington, C. S., Bradley, L. J., Crews, C., & Hennington, E. A. (2013). *The Halo Effect: Considerations for the Evaluation of Counselor Competency* (pp. 1-10). Vistas Online. http://counselingoutfitters.com/vistas/VISTAS_Home.htm

Hsieh, C. N. (2011). Rater Effects in ITA Testing: ESL Teachers' versus American Undergraduates' Judgments of Accentedness, Comprehensibility, and Oral Proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9,* 47-74. https://michiganassessment.org/wp-content/uploads/2014/12/Spaan_V9_FULL.pdf#page=55

Huang, L., Kubelec, S., Keng, N., & Hsu, L. (2018). Evaluating CEFR Rater Performance through the Analysis of Spoken Learner Corpora. *Language Testing in Asia, 8,* 1-17. https://doi.org/10.1186/s40468-018-0069-0

Isaacs, T., & Thomson, R. I. (2013). Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions. *Language Assessment Quarterly, 10,* 135-159. https://doi.org/10.1080/15434303.2013.769545

Jabeen, R. (2016). *An Investigation into Native and Non Native English Speaking Instructors' Assessment of University ESL Student's Oral Presentation.* Mankato, MN: Minnesota State University. https://cornerstone.lib.mnsu.edu/etds/647/

Kim, H. J. (2015). A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment. *Language Assessment Quarterly, 12,* 239-261. https://doi.org/10.1080/15434303.2015.1049353

Lee, H. (2017). *The Effects of Rater's Familiarity with Test Taker's L1 in Assessing Accentedness and Comprehensibility of Independent Speaking Tasks.* Seoul: Department of English Language and Literature, Seoul National University. http://s-space.snu.ac.kr/handle/10371/139643

Myford, C., & Wolfe, E. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part IL. *Journal of Applied Measurement, 5,* 189-227. https://www.researchgate.net/profile/Carol_Myford/publication/9069043_Detecting_and_Measuring_Rater_Effects_Using_Many-Facet_Rasch_Measurement_Part_I/links/54cba70e0cf298d6565848ee.pdf

Noor Lide, A. K. (2011). Judging Behaviour and Rater Errors: An Application of the Many-Facet Rasch Model. *GEMA Online Journal of Language Studies, 11,* 179-197. http://ejournals.ukm.my/gema/article/view/49

Oudman, S., Pol, J. Van De, Bakker, A., Moerbeek, M., & Gog, T. Van. (2018). Effects of Different Cue Types on the Accuracy of Primary School Teachers' Judgments of Students' Mathematical Understanding. *Teaching and Teacher Education, 76,* 1-13. https://doi.org/10.1016/j.tate.2018.02.007

Saadat, M., & Alavi, S. Z. (2018). The Effect of Type of Paragraph on Native and Non-Native English Speakers' Use of Grammatical Cohesive Devices in Writing and Raters' Evaluation. *3L: Language, Linguistics, Literature, 24,* 97-111. https://doi.org/10.17576/3L-2018-2401-08

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology, 85,* 956-970. https://doi.org/10.1037/0021-9010.85.6.956

Song, T., Wolfe, E. W., Less-Petersen, M., Sanders, R., & Vickers, D. (2014). *Relationship*

between Rater Background and Rater Performance. https://pdfs.semanticscholar.org/f745/93340d40576dc303eed2c7998806ef27554a.pdf

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology, 104,* 743-762. https://doi.org/10.1037/a0027627.supp

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2018). The Teacher as Examiner of L2 Oral Tests: A Challenge to Standardization. *Language Testing, 35,* 217-238. https://doi.org/10.1177/0265532217690782

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The Influence of Lexical Features on Teacher Judgments of ESL Argumentative Essays. *Assessing Writing, 39,* 50-63. https://doi.org/10.1016/j.asw.2018.12.003

Weilie, L. (2018). To What Extent Do Non-Teacher Raters Differ from Teacher Raters on Assessing Story-Retelling. *Journal of Language Testing & Assessment, 1,* 1-13. http://clausiuspress.com/assets/default/article/2018/08/29/article_1535590233.pdf https://doi.org/10.23977/langta.2018.11001

Wind, S. A. (2018). Examining the Impacts of Rater Effects in Performance Assessments. *Applied Psychological Measurement, 43,* 159-171. https://doi.org/10.1177/0146621618789391

Wind, S. A., & Schumacker, R. E. (2017). Detecting Measurement Disturbances in Rater-Mediated Assessments. *Educational Measurement: Issues and Practice, 36,* 44-51. https://doi.org/10.1111/emip.12164

Wind, S. A., Stager, C., & Patil, Y. J. (2017). Exploring the Relationship between Textual Characteristics and Rating Quality in Rater-Mediated Writing Assessments: An Illustration with L1 and L2 Writing Assessments. *Assessing Writing, 34,* 1-15. https://doi.org/10.1016/j.asw.2017.08.003

Winke, P., & Gass, S. (2013). The Influence of Second Language Experience and Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation. *TESOL Quarterly, 47,* 762-789. https://doi.org/10.1002/tesq.73

Wu, M. (2017). Some IRT-Based Analyses for Interpreting Rater Effects. *Psychological Test and Assessment Modeling, 59,* 453-470. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/04_Wu.pdf

Yahya Ameen, T., Mohd Sallehhudin, A. A., Kemboja, I., & Alla Baksh, M. A. K. (2014). The Washback Effect of the General Secondary English Examination (GSEE) on Teaching and Learning. *GEMA Online Journal of Language Studies, 14,* 83-103. https://doi.org/10.17576/GEMA-2014-1403-06

Zhang, Y., & Elder, C. (2014). Investigating Native and Non-Native English-Speaking Teacher Raters' Judgments of Oral Proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy and Practice, 21,* 306-325. https://doi.org/10.1080/0969594X.2013.845547

Zhao, K. (2017). *Investigating the Effects of Rater's Second Language Learning Background and Familiarity with Test-Taker's First Language on Speaking Test Scores.* Provo, UT: Brigham Young University.