Examining Rater Performance on the CELBAN Speaking: A Many-Facets Rasch Measurement Analysis

Peiyu Wang
Queen's University

Karen Coetzee
Touchstone Institute

Andrea Strachan
Touchstone Institute

Sandra Monteiro
Touchstone Institute

Liying Cheng
Queen's University

Abstract

Internationally educated nurses' (IENs) English language proficiency is critical to professional licensure as communication is a key competency for safe practice. The Canadian English Language Benchmark Assessment for Nurses (CELBAN) is Canada's only Canadian Language Benchmarks (CLB) referenced examination used in the context of healthcare regulation. This high-stakes assessment claims proof of proficiency for IENs seeking licensure in Canada and a measure of public safety for nursing regulators. Understanding the quality of rater performance when examination results are used for high-stakes decisions is crucial to maintaining speaking test quality as it involves judgement, and thus requires strong reliability evidence (Koizumi et al., 2017). This study examined rater performance on the CELBAN Speaking component using a Many-Facets Rasch Measurement (MFRM). Specifically, this study identified CELBAN rater reliability in terms of consistency and severity, rating bias, and use of rating scale. The study was based on a sample of 115 raters across eight test sites in Canada and results on 2698 examinations across four parallel versions. Findings demonstrated relatively high inter-rater reliability and intra-rater reliability, and that CLB-based speaking descriptors (CLB 6-9) provided sufficient information for raters to discriminate examinees' oral proficiency. There was no influence of test site or test version, offering validity evidence to support test use for high-stakes purposes. Grammar, among the eight speaking criteria, was identified as the most difficult criterion on the scale, and the one demonstrating most rater bias. This study highlights the value of MFRM analysis in rater performance research with implications for rater training. This study is one of the first research studies using MFRM with a CLB-referenced high-stakes assessment within the Canadian context.

Résumé

Les compétences linguistiques dans la langue anglaise chez des infirmiers et infirmières ayant reçu leur éducation à l'étranger s'avèrent critiques à l'acquisition du permis professionnel d'exercer leur profession, car les compétences communicatives sont clé à la pratique sécuritaire. L'examen langagier des compétences de langue anglaise *The Canadian English Language Benchmark Assessment for Nurses* (CELBAN) demeure le seul examen langagier référentiel canadien auquel on fait référence dans le contexte canadien des règlements de contrôle du système de santé. Cet examen à enjeux élevés offre une preuve de compétence langagière de langue anglaise de la part des infirmiers et infirmières ayant reçu leur formation professionnelle à l'étranger et qui sont à la recherche d'un permis pour exercer leur profession au Canada, ainsi qu'une mesure de sécurité publique destinée aux régulateurs de la profession d'infirmiers et infirmières. Comprendre la qualité de la performance des évaluateurs/trices étant donné que les résultats servent à des décisions sur des enjeux importants demeure fondamental au maintien de la qualité de l'épreuve des compétences orales, car celle-ci implique le jugement et donc nécessite de fortes évidences de fiabilité (Koizumi, et coll. 2017). Cette étude a examiné la performance d'évaluateur/trice sur la composante des compétences orales du CELBAN en utilisant la mesure multifacette Rasch (MMFR). Spécifiquement, cette étude a identifié la fiabilité des évaluateurs/trices, la difficulté des critères, le parti pris de l'évaluation et l'usage de l'échelle de classification. Cette étude s'est basée sur un échantillon de 115 évaluateurs/trices dans huit centres d'évaluation au Canada et sur les résultats de 2.698 évaluations dans quatre versions parallèles. Les résultats démontrent une haute fiabilité relative entre évaluateurs/trices ainsi que sur le plan des intraévaluateurs/trices. De plus, les descripteurs des compétences orales de base des Compétences linguistiques canadiennes (CLC 6-9) ont fourni suffisamment d'information afin de permettre aux évaluateurs/trices de préciser le niveau de compétences du candidat / de la candidate. Il n'y a pas eu d'influence du site de l'examen ni de la version de celui-ci, ce qui offre de l'évidence de validité afin d'affirmer l'usage de cette épreuve pour des enjeux importants. La grammaire, une des huit critères, a été relevée comme étant celle la plus difficile sur l'échelle, et celle qui a mis en lumière le plus grand parti pris de la part des évaluateurs/trices. Cette étude accentue la valeur de l'analyse en effectuant la mesure multifacette Rasch dans des recherches de performance ayant des implications pour l'entraînement des évaluateurs/trices. Cette étude est parmi les premières se servant de la MMFR avec une évaluation à enjeux élevés à base des CLC dans le contexte canadien.

Examining Rater Performance on the CELBAN Speaking: A Many-Facets Rasch Measurement Analysis

In countries where the primary language used in health care is English, internationally educated health professionals are often required to demonstrate English language proficiency in order to qualify for professional practice. In Canada, the Canadian English Language Benchmarks Assessment for Nurses (CELBAN) fulfills this role for internationally educated nurses (IENs). The CELBAN was introduced in 2004 with the intent of facilitating the evaluation of IENs who were recruited specifically to help ease the current shortage of nurses in Canada (Epp & Lewis, 2004a; Jeans et al., 2005). A passing score from the CELBAN is recognized by Canadian nursing regulators as evidence of

English language proficiency for entry to practice level registration (see www.nnas.ca). The CELBAN focuses on assessing English language skills required for high-frequency nursing duties through task-based evaluation of reading, speaking, listening, and writing. Communication tasks contained within the CELBAN were developed based on an analysis of the language demands of the nursing profession in Canada (Epp & Lewis, 2004b) and simulate authentic tasks of a licensed nurse (Touchstone, 2018). Additionally, the CELBAN test score and assessment rubric align with Canadian Language Benchmarks (CLB) – a descriptive scale of communicative proficiency in English as a second language (Centre for Canadian Language Benchmarks [CCLB], 2013; further description of the CLB can be found at www.language.ca). Collecting evidence of validity for an assessment utilized for a high-stakes purpose such as entry to practice is a continuing practice. This study contributes to the CELBAN speaking score validation through the psychometric analysis of rater performance.

Literature Review

Establishing Evidence of Validity

Messick (1995) defines the adequacy of the inferences made through test scores to be reliant on the multiple sources of empirical evidence: content, substantive, structural generalizability, external and consequential validity. The CELBAN's design is supported by a comprehensive language benchmarking of the demands of the nursing profession (Epp & Stawychny, 2002) which identified the target language use (Bachman & Palmer, 1996; Douglas, 2001) and the constructs to be measured. This benchmarking analysis was anchored in the Canadian Language Benchmarks (CLB): "independent standards that describe a broadly applied theory of language ability" (CCLB, 2013, p.14), which supports a theory-based, substantive validity claim. Multiple language use functions are sampled through a series of communicative tasks in the CELBAN for appropriate domain coverage, and performance is evaluated through a CLB-referenced rubric. The CELBAN's test development process is documented and openly available for public reference (Epp & Lewis, 2004b). These test specifications delineate its constructs and structure for the purposes of ongoing test development (Touchstone Institute, 2018). Test renewal is chronicled through *Facts & Figures* reports available to the public (Touchstone Institute, 2016) and describes how CELBAN retains construct validity through consultations with nursing professionals. Additionally, the two-rater model and ongoing inter-rater reliability measures for the CELBAN speaking component are designed to support valid score interpretations. Although systematic quality assurance processes contribute to ongoing validation of the constructs evaluated by the CELBAN, limited research evidence is publicly accessible.

Rater Performance: Rater Cognition and Error Variance

Rater-based assessments such as speaking and writing are susceptible to multiple sources of error variance (Bachman et al., 1995; Gingerich et al., 201; Sebok & Syer, 2015). McNamara (1996) highlighted four dimensions of rater variability: rater consistency, rater leniency (or severity), rater's use of rating scale, and rater bias. These

four dimensions have been examined by language researchers in relation to various variables such as rating experience (Brown, 2000), rating context (Lumley & McNamara, 1995), rater type (Kim, 2009), task types and rating criteria (Wigglesworth, 1993), and examinees' gender (Eckes, 2005). Kondo-Brown (2002) evaluated rater bias of Japanese writing performance assessment through a Many-Facets Rasch Measurement (MFRM) analysis and concluded that raters demonstrated severe and lenient rating patterns but maintained consistency in general (sometimes referred to as hawks and doves). Eckes conducted an MFRM analysis of writing and speaking performance assessments and revealed relatively more consistency in raters' overall rating compared to their use of rating scales. Eckes (2008, 2012) proposed that raters' use of rating scales related to their rater type, and rating can be regarded as a routine (i.e., fixed) process that was formed by their past rating experience and belief of rating scale importance. Cai (2015) later confirmed this correlation in speaking assessments and added that rater type can also affect rater bias during the speaking rating process.

Lim (2011) conducted a longitudinal study of writing assessments through MFRM analysis to investigate the development and maintenance of rating quality for both novice and experienced raters. The results showed that novice raters improve rating quality faster compared to experienced raters, and both groups maintain consistency in quality over time. In a more recent study, Davis (2016) examined the effect of training on rater scoring patterns in the Test of English as a Foreign Language Internet-based Testing (TOEFL iBT) speaking test using MFRM analysis. The results indicated that experienced raters had achieved the desired severity and internal consistency prior to training, but training increased inter-rater reliability.

In rater-based assessments, two raters may assign identical scores on the same rating criteria with different rating perceptions. Orr (2002) evaluated the rating performances of 32 trained raters in a speaking test and found that raters did not focus on the same aspects of rating criteria, and applied varied assessment standards while assigning the scores. Han (2018) suggest that raters in second language (L2) speaking assessments tend to rely more on certain rating criteria (e.g., content, grammar, organization) than others. Lumley (2002) conducted MFRM analysis to assess rater performances in writing assessment and found that raters tend to rate severely on grammar. Caban (2003), Lee (2018), and McNamara (1990) provided similar findings revealing that speaking raters interpreted each scoring category in a rather different way.

The structure or cognitive demands created by the rating scales can influence how raters apply the rubric (Tavares et al., 2013). Raters may be unable to differentiate between analytical elements and therefore might assign similar scores when the elements are too similar to each other (Johnston et al., 2009). If a holistic rating scale is utilized to evaluate a construct underlying several skill dimensions, raters may fail to assign an appropriate score due to the fact that they are confused about the priorities of each dimension composing the score (Barkaoui, 2010). In another study, raters reached high agreement on the upper half of the rating scale and low agreement on the lower half of the rating scale (Yan, 2014). In this case, raters who consistently assign above average scores across all examinees are considered lenient, while raters who constantly score below average scores are regarded as severe. Raters' leniency or severity may change over time (Wolfe et al., 2007), vary across rubric dimensions (Eckes, 2005), and be inconsistent across scoring levels (Yan, 2014). These sources of rater-based variance may lead to inaccurate decisions,

yet there have been no studies examining the influence of rater-based variability on CELBAN scores.

Measuring Error Variance in Rater-Based Assessments

Evaluating an assessment, like the CELBAN, for evidence of validity requires a measure of inter-rater (agreement between two or more persons rating one examinee) and intra-rater reliability (agreement between ratings by one person rating various examinees) (Bramley, 2007). The rationale behind this is a core Classical Test Theory (CTT) principle when evaluating any construct: if a construct is defined appropriately and is observable by an objective observer, then raters should agree as to what they observe, allowing an approximation of the true score (the concept of a true score is consistent with CTT which assumes that it is possible to measure the actual, or error-free ability of examinees; Streiner et al., 2015). Perhaps more importantly, a construct should be rated similarly by the same person across different time points. Using classical test theory approaches to psychometrics, we can evaluate the reliability of the data using measures of agreement; coefficients like Kappa, intraclass correlation, or Spearman correlation. (Streiner et al., 2015). The level of agreement can also be communicated with a measure of internal consistency, such as Cronbach's Alpha (Cronbach, 1951) which can be viewed as a special case of intraclass correlations (Shrout & Fleiss, 1979; Streiner et al., 2015). However, CTT internal consistency measures sometimes fail to identify systematic inter-rater differences such as when raters are consistently lenient or severe across all items (Newton, 2009).

Many-Facets Rasch Measurement analysis is an extension of basic Rasch analysis that analyzes two facets, typically, examinees and items (Baylor et al., 2011; Reckase, 1997). Performance assessments typically not only include examinees and items/tasks, but also other facets such as raters, scoring criteria, and possibly many more. Micko (1969) and Kempf (1977) were the earliest researchers proposed to extend the basic Rasch model by considering three or more facets. Many-Facet Rasch analysis has received increased attention and is commonly employed in the fields of language testing, educational and psychological measurement (Barkaoui, 2014; Linacre & Wright, 1989). The approach has been regarded as "a standard part of the training of language testers and is routinely used in research on performance assessment" (McNamara, 2011, p. 436). Many-Facet Rasch measurement model (MFRM) is useful when analyzing test data affected by three or more facets such as examinees, raters, and evaluation criteria. It combines multiple facets into the same scale allowing users to compare various factors on the same reference scale. Eckes (2011) identified four main reasons that MFRM analysis is advantageous. First, MFRM can produce in-depth information that includes rater severity, rater self-consistency and rater bias that relate to examinees, raters, and evaluated criteria facets. Second, the analysis procedure is simple and quick, and details can be derived through just a single run of analysis. Third, MFRM can deal with data that contain missing responses. Fourth, it considers differences in rater severity and criterion difficulty.

Research Questions

The current study adopted the Many-Facet Rasch Measurement analysis approach proposed by Linacre and Wright (1989) to examine the CELBAN Speaking data in order to identify score patters and rater behaviours on the CELBAN in terms of rater reliability, criteria difficulty, rating bias, and use of rating scale. This study examined rater performance by addressing the three questions as follows:

1. What are the levels of inter-rater and intra-rater reliability? To what extent do raters differ in rating severity and leniency?
2. In what ways do raters show systematic bias patterns when applying the rating criteria?
3. How does the rating scale discriminate performance categories and levels?

Methods

Assessment Context

The CELBAN Speaking Test

The CELBAN speaking test features eight tasks that engage examinees in discussions and role-plays. Questions and topics of the CELBAN speaking test begin with concrete daily routine topics and move to abstract, hypothetical and less predictable topics. The discussion tasks elicit health-related discourse and role-play tasks prompt typical and commonly occurring interactions for authentic health contexts. The speaking test format is a 20-30 minute face-to-face interview facilitated by two trained CELBAN raters who take turns as the interlocutor and the evaluator. Two speaking raters assign their scores independently – referencing CLB 6 to CLB 9 – and these scores are recorded in the scoring sheets and entered into the database as two discrete decisions. There are eight rating criteria: communication (the ability to produce the appropriate language); intelligibility (the clarity of speech); grammar (grammar accuracy); vocabulary (the variety and accuracy of general and health-related vocabulary); fluency (speech flows); organization and cohesion (idea connection and support); initiative (take initiative and establish rapport); and use of communication strategies (acknowledgement, clarification, affirmation, etc.). A final score is arrived at once the two raters (Rater A and Rater B) have completed their independent evaluations and conferred that they arrived at the same final score. If the final scores differ, raters will deliberate until they reach an agreement. If two raters cannot reach consensus, a third rating (Rater C) will be needed. The minimum CELBAN scores required for nursing registration are set by Canadian Nursing Regulators. For speaking, the cut-off benchmark is CELBAN 8.

Description of the Data

The data drew upon 2018 test results across eight test sites and included a total sample of 2698 examinees and 115 raters. A holistic score (range from 6 to 9) and eight analytical scores (communication, intelligibility, grammar, vocabulary, fluency, organization & cohesion, initiative, strategies) were detailed in the dataset for each

examinee. This study analyzed the independent scores of all 115 raters. The dataset included 2698 examinees with up to 3 repeated tests. This may affect the results of this study but was considered a minor risk as the number of examinees that took multiple tests was small, and raters are randomly assigned to the examinees.

Sample Background and Demographics

Given the context where the CELBAN is applied as a high-stakes assessment for internationally educated nurses, examinees typically originate from countries where English is either not used or is used as a secondary official language. However, for confidentiality and test security reasons, examinees' demographic information was not included in the analysis.

Data Analysis

This study applied the Many-Facets Rash Measurement (MFRM) using Facets Software (Linacre, 2014) to evaluate rater performance and the quality of the rating scale. To provide additional validity evidence of CELBAN rating, the results were also used to analyze whether the test sites and test versions were sources of error variance. Specifically, a five-facet MFRM was applied to the data. The five facets included in this study were as follows:

Facet 1 = IENs*(N=2698); *Facet 2* = Raters (N=115); *Facet 3* = Test Site (N=8); *Facet 4* = Test Versions (N=4); and *Facet 5* = Speaking Criteria (N=8)

Results

For clarity, the results are reported according to each of the five facets first to answer research question 1 followed by rater bias analysis and rating scale measurement to answer research questions 2 and 3.

Examinee Measurement Report (Facet 1)

The first facet analysis shows the examinees' proficiency level. Table 1 includes the examinee facet (Facet 1) from five examinees with different levels of proficiency. The observed average (column 2) in the table indicates the raw average scores assigned by raters, and the fair average (column 3) shows the expected average scores that were assigned by a rater with an average severity. The proficiency measure (column 4) specifies the examinees' proficiency on the logit scale, and model SE (column 5) reveals the errors of examinees' proficiency estimates. Specifically, Examinee 1783 had the highest proficiency estimate (9.51 logits, SE = 1.85), and Examinee 608 had the lowest estimate (–9.15 logits, SE = 1.85). The strata value of 5.37 with a high separation reliability of 0.93 suggests that among the 2969 examinees included in the analysis, there are about five statistically distinct classes of examinee proficiency.

Table 1
Examinee Measurement (Facet 1)

Examinee	Observed Average	Fair Average	Proficiency Measure	Model SE
No. 1783	9.00	8.99	9.51	1.85
No. 738	8.94	8.97	8.06	1.04
No. 2629	7.19	7.16	−1.58	0.51
No. 1944	6.56	6.50	−4.38	0.49
No. 608	6.00	6.03	−7.94	1.85

Note. Separation 3.78, Strata 5.37, Separation Reliability (not inter-rater) 0.93

Rater Measurement Report (Facet 2)

Table 2 presents the raters' measurement report in rating the examinees' speaking performances. The rater measurement analysis yielded 115 raters' measurement estimates in total, for better data visualization, only raters with most and least rating severity (five each) were ranked and reported in Table 2. The rating count (column 2) identifies the total rating the raters performed, and severity measurement (column 3) describes the rater severity estimates, and it appears in order from most lenient to most severe. The model SE (column 4) indicates the errors of rater severity estimates, and infit/outfit MnSq (column 5 and 6) reveal rater fit statistics.

In an MFRM analysis, reliability estimates are reflected through exact/expected agreement (*inter-rater*) and rater fit statistics (*intra-rater*). In Table 2, the observed agreement opportunities (inter-rater) were 25539 of exact agreement 17666 with a percentage of 69.2%. The expected agreement is 15396.3 with a percentage of 60.3%. We can see the observed agreement is slightly higher than the expected agreement, which met the indicator of good *inter-rater* reliability and proved raters did not do their rating in an independent way (Linacre, 2018).

Rater fit statistics include infit MnSq and outfit MnSq that indicate the extent to which ratings provided by a given rater match the expected ratings that were generated by this model indicating intra-rater reliability. Rater fit values greater than 1.0 indicate more variation than expected in the ratings; this kind of misfit is called underfit. By contrast, rater fit values less than 1.0 indicate less variation than expected, meaning that the ratings are too predictable or provide redundant information; this is called overfit. Wright and Linacre (1994) highlighted that reasonable rater fit values should range between 0.4 and 1.2. Overall, the rater fit values were within the acceptable range which means raters made consistent ratings and used the rating scale in a consistent way.

Table 2
Rater Measurement (*Facet 2*)

	Rating Count	Severity Measure	Model SE	Infit MnSq	Outfit MnSq
Rater 86	96	−1.35	0.23	1.38	1.34
Rater 113	8	−0.91	0.71	0.99	0.97
Rater 48	400	−0.82	0.11	1.01	1.01
Rater 19	656	−0.75	0.09	0.80	0.75
Rater 89	120	−0.74	0.22	0.87	0.87
Rater 33	456	0.70	0.10	0.88	0.84
Rater 24	496	0.72	0.10	0.91	0.89
Rater 60	392	0.87	0.11	0.88	0.87
Rater 114	32	0.92	0.38	0.89	0.84
Rater 105	32	1.46	0.39	1.18	1.09

Note. Strata 3.33, Separation Reliability (not inter-rater) 0.83
Fixed (all same) chi-square: 1367.4, d.f.: 114, significance (probability): 0.00
Inter-rater agreement opportunities: 25539, Exact agreements: 17666=69.2% Expected: 15396.3=60.3%

The rater measurement analysis provides this study with both group-level and individual-level rater severity information. At the *individual-level*, severity measurement explained raters' rating severity patterns, where positive values indicate severity and negative values represent leniency. From Table 2, we can see that rater severity ranges from logit −1.35 to 1.46, with rater 86 as most lenient (−1.35) and rater 105 as most severe (1.46). Based on a rough guideline that average rater severity estimates is −1.0 to 1.0 logits, most of the raters in this study were neither severe nor lenient ("average" or "normal"), except for rater 86 and rater 105. The fixed chi-squared test result further provides *group-level* severity evidence. The fixed-chi square statistics (see Table 2) tell us the difference between the expected data and observed data. The observed data was the raw scores from the CELBAN tests and expected data is the expected score that is assigned by raters with average rating severity. In this study, rater severity is significantly different *(Q=1376.4, df=114, p<.001*). The strata value of 3.33 with the reliability of 0.83 tells us, 115 raters in this study could be clustered into three statistically distinct levels in terms of severity. This finding suggests that CELBAN rater's performance is in a largely consistent pattern in terms of severity and leniency.

Test Site (Facet 3) and Test Version (Facet 4) Report

Table 3 and Table 4 present the test site and test version measurement reports. As we can see from the test site measure report (Table 3), small differences were observed between test sites in terms of the observed average (7.63-7.99) and fair average scores (7.86). On the logit scale, the influence of site ranges from 0.00 (SE+0.01) to 0.00 (SE+0.003). This result presents no influence due to test sites. This is the same with the test version measurement (Table 4), where small differences were observed between test versions in terms of the observed (7.79-7.81) and fair average score (7.86). On the logit scale, the influence of site ranges from 0.00 (SE+0.00) to 0.00 (SE+0.002).

Table 3
Test Site Measurement (Facet 3)

	Observed Average	Fair Average	Measure	Model SE	Infit MnSq	Outfit MnSq
Site 1	7.99	7.86	0.00	0.03	1.07	1.08
Site 2	7.89	7.86	0.00	0.03	0.96	0.93
Site 3	7.79	7.86	0.00	0.02	0.93	0.91
Site 4	7.76	7.86	0.00	0.03	0.82	0.79
Site 5	7.70	7.86	0.00	0.02	1.04	1.04
Site 6	7.98	7.86	0.00	0.05	1.05	1.09
Site 7	7.90	7.86	0.00	0.04	1.10	1.10
Site 8	7.63	7.86	0.00	0.03	1.14	1.16

Note. RMSE 0.03, Adj (True) S.D. 0.00, Separation 0.00, Strata 0.33, Reliability 0.00
Fixed (all same) chi-square: 0.0, d.f.: 7, significance (probability): 1.00

Table 4
Test Version Measurement (Facet 4)

	Observed Average	Fair Average	Measure	Model SE	Infit MnSq	Outfit MnSq
Version 1	7.79	7.86	0.00	0.02	0.99	0.98
Version 2	7.79	7.86	0.00	0.02	1.01	1.01
Version 3	7.79	7.86	0.00	0.02	1.00	1.00
Version 4	7.81	7.86	0.00	0.02	1.00	0.99

Note. RMSE 0.02, Adj (True) S.D. 0.00, Separation 0.00, Strata 0.33, Reliability 0.00
Fixed (all same) chi-square: 0.0, d.f.: 3, significance (probability): 1.00

Criterion Measurement Report (Facet 5)

Table 5 presents a detailed description of the eight criteria used to measure examinees' speaking abilities. The measure (column 2) indicates the difficulty for the examinee to receive a high score. The table sorts the criteria from the most difficult [*grammar* (1.07)] to the easiest [*initiative* (–0.43)]. The result suggests that it was very difficult for examinees to obtain high scores in *grammar*, with a difficulty of 1.07 logits, while it was much easier for them to get high scores in *initiative*, with a difficulty of −0.43 logits. Due to the relatively large number of responses used for estimating each difficulty measure (total count per criterion was 5639), the measurement precision was very high. For each criterion, MnSq fit indices stayed well within very narrow quality control limits (i.e., 0.90 and 1.10). This is evidence supporting the assumption of unidimensional measurement, as implied by the Rasch model. That is, these criteria worked together to define a single latent dimension.

Table 5
Criterion Measurement Report (Facet 5)

	Measure	Model SE	Infit MnSq	Outfit MnSq
Grammar	1.07	0.03	0.99	1.00
Organization and Cohesion	0.27	0.03	0.92	0.91
Fluency	0.06	0.03	1.14	1.14
Vocabulary	−0.22	0.03	0.96	0.94
Communication	−0.23	0.03	0.77	0.73
Strategies	−0.24	0.03	0.94	0.93
Intelligibility	−0.28	0.03	1.31	1.33
Initiative	−0.43	0.03	0.96	0.94

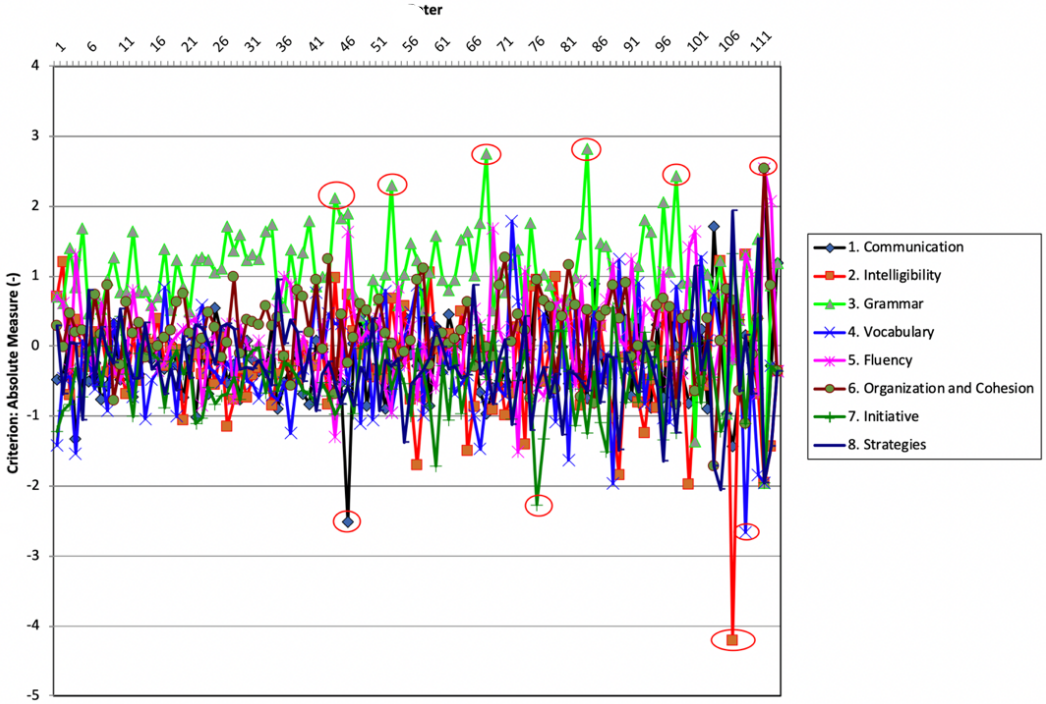
Note. RMSE 0.03, Adj (True) S.D. 0.49, Separation 16.60, Strata 22.47, Reliability 1.00
Fixed chi-square 2021.2, d.f.: 7, significance (probability): 0.00

Rater Bias Interaction

Overall, the bias interaction of rater by rating criterion analysis identified 6 biases out of 920 possible iterations (115 raters × 8 criteria). From the rater bias diagram (Figure 1), we can see that *vocabulary* is the easiest criterion while *grammar* is the most difficult criterion for examinees. Many raters have systemic bias towards *grammar*.

Based on an acceptable range of -2.00 to 2.00 logits, there are 10 raters showing systematic bias against the rating criterion. There were five raters who exhibited a significantly severe bias towards *grammar*, one rater towards *communication*, one rater towards *initiative*, one rater towards *intelligibility*, and one rater towards *intelligibility and fluency*. Table 6 shows an individual rater bias information of one particular rater (Rater 32).

Figure 1
Rater-Criterion Bias Diagram



To identify rater biases towards the rating criteria, this study conducted the rater criterion bias interaction analysis. At the individual level, rater bias analysis (Table 6) provided more detailed statistical information of rating bias for rater training. This analysis can provide in-depth information for individual rater calibration. This study randomly picked one rater's results as an example for data interpretation. Table 6 lists Rater 32's overall difficulty measures, number of ratings in total, total of scores assigned (observe score), expected scores, and the average variance between the observed and expected scores. Bias measure refers to the criterion bias measure on the logit scale. Positive values of bias measure indicate observed scores are higher than expected based on the model, and vice versa. Specifically, Rater 32 assigned higher than expected scores on *strategies, fluency, vocabulary, and intelligibility*, while lower than expected scores on *initiative, organization/cohesion, grammar, and communication*. Also, the t-value can be utilized to identify bias interactions. Based on the control limit of −2 to +2, rater 32 displayed bias towards *initiative* criterion.

Table 6
Rater 32 Criterion Bias

Criterion	Difficulty Measure	N of Ratings	Observed Score	Exp. Score	Obs-Exp Average	Bias Measure	SE	t	p
Strategies	−0.24	167	1302	1298.90	0.02	−0.09	0.17	−0.53	0.5960
Initiative	−0.43	167	1293	1305.31	−0.07	0.36	0.17	2.13	0.0345
Organization & Cohesion	0.27	167	1278	1281.05	−0.02	0.08	0.17	0.51	0.6117
Fluency	0.06	167	1259	1288.27	0.04	−0.19	0.17	−1.13	0.2589
Vocabulary	−0.22	167	1309	1298.11	0.07	−0.32	0.17	−1.86	0.0640
Grammar	1.07	167	1242	1250.78	−0.05	0.23	0.16	1.41	0.1598
Intelligibility	−0.28	167	1305	1300.14	0.03	−0.14	0.17	−0.83	0.4054
Communication	−0.23	167	1297	1298.52	−0.01	0.04	0.17	0.26	0.7953

Rating Scale Measurement

Table 7 reveals the measurement report of the four categories (CLB 6, 7, 8, 9) that raters used to measure examinees' speaking abilities. This table represents data in 5 columns ranked by the level of score from CLB 6 to CLB 9. The observed (column 1) indicates the total account assigned by raters for each rating scale and its associated percentage. The average measure (column 2) describes the average examinees' speaking abilities for each scale. The expected measure (column 3) reflects the expected speaking abilities computed from the model. The outfit MnSq (column 4) is the mean square fit statistics for each scale category. The Rasch-Andrich threshold (column 5) is the step calibration which shows how difficult it is to choose one rating on the scale.

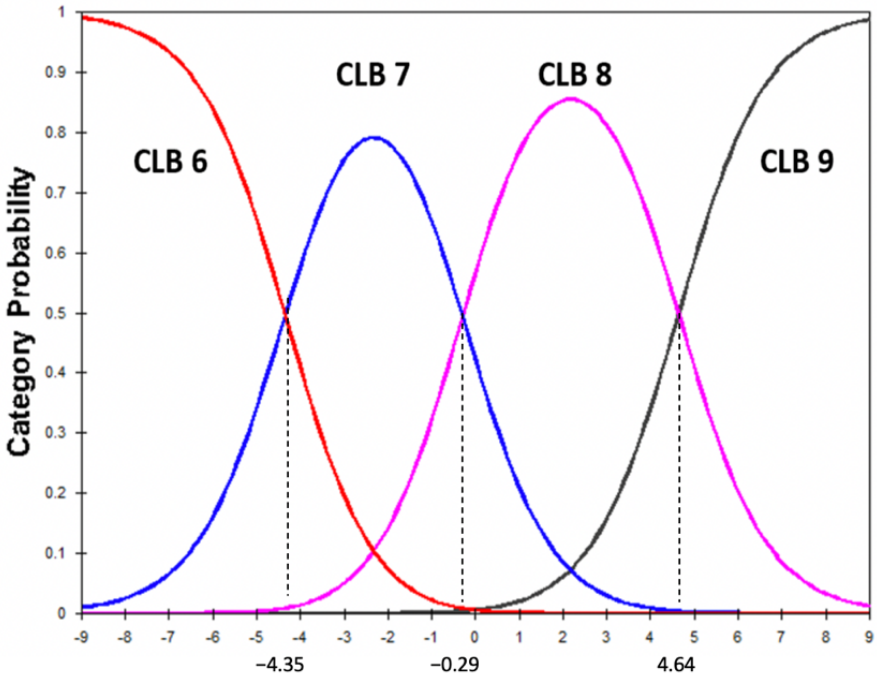
In Table 7, we can see that the least assigned rating was CLB 6 with a frequency of 983 and accounts for two percent of the total ratings, while the most assigned rating was CLB 8 which has a frequency of 26366 and eleven percent of the total account. The average and expected measure values in column 3 and 4 indicate CLB 7 (−0.58 logits), CLB 8 (1.52 logits), and CLB 9 (4.48 logits) were an exact match to the expected values the Rasch model predicted. The CLB 6 was assigned slightly lower than the expected value from the model. The outfit MnSq value for all four rating categories are close to the expected value of 1.0.

Table 7
Rating Scale Measurement

	Observed		Average Measure	Expected Measure	Outfit MnSq	Rasch-Andrich	
	Frequency	%				Measure	S.E.
CLB 6	983	2%	−3.11	−3.19	1.1		
CLB 7	12473	28%	−0.58	−0.58	1.0	−4.35	0.04
CLB 8	26366	59%	1.52	1.52	1.0	−0.29	0.01
CLB 9	4709	11%	4.48	4.48	1.0	4.64	0.02

Figure 2 provides a graphical description of the Rasch-Andrich threshold order from the data analysis. The figure illustrates the category probability curves for the CLB (6-9) scale that raters used to assess examinees' speaking proficiency. The horizontal axis indicates the examinee proficiency; the vertical axis presents the probability of being rated in each scale. There is one curve and one peak for each category, but the peaks appear to be different and distinct. The dash lines where adjacent rating scales cross indicate the Rasch-Andrich threshold. We can see in the figure that the threshold is nicely ordered from left to right based on the levels of the rating scale though with overlapping, and the threshold appears to progress from −4.35 logits to 4.64 logits. Linacre (2002) suggests that the size of the increase in rating scale threshold values should range from minimal 1.4 logits to maximal 5.0 logits. As can be seen in Table 7 Rasch-Andrich threshold column, the increase between adjacent rating scales (CLB 7 to CLB 8, CLB 8 to CLB 9) are 4.06 and 4.93, which all stayed within the acceptable range.

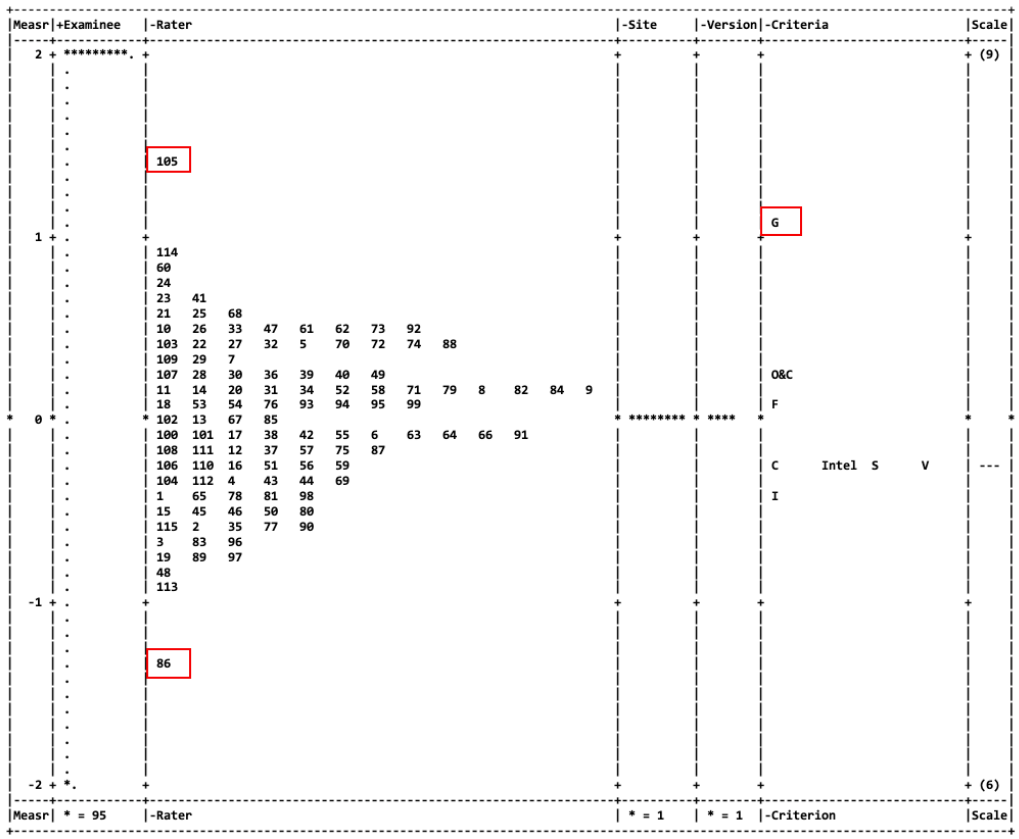
Figure 2
Category and Level Probability



Discussion and Implications

The study examined the speaking rating behaviours, function of rating scale, and rating biases related to rating criteria in the CELBAN test. Figure 3, MFRM Wright Map, is a graphical display of the measurement estimates that were directly taken from the FACETS output.

Figure 3
MFRM Wright map



In the Wright Map (Figure 3), all measures of the five facets (examinees, raters, sites, versions, and criteria) are vertically positioned on the same dimension, with logits as the measurement units. The logit scale appears as the first column labelled as Measr to define the units underlying the Rasch model. Similar to how rulers have inches or centimetres as their standardized units, logits are the units of the Rasch model used to locate the marks on our measuring instruments. The second column (+Examinee) shows the locations of the examinee proficiency estimates. The plus sign indicates that examinee was positively oriented, which means that the higher-scoring examinees were located at the top and lower-scoring examinees were located at the bottom. As can be seen from the bottom, each star indicates a total of 95 examinees. Through this figure, one can clearly visualize the proportion of examinees who met the required level of proficiency for the nursing profession. The examinees with extreme scores are located at the top and bottom rows, respectively.

The third column (-Rater) maps out raters' level of severity on the Rasch logit scale, and the minus sign indicates the negatively oriented measures – severe raters appear higher in the column and lenient raters appear lower (rater 86 as most lenient and rater 105 as most severe). Although only a small number of raters fall outside of the normal fit ranges, in practice the data help with quality assurance procedures in that it facilitates specific feedback and shapes calibration activities for these raters.

The fourth and fifth columns (-Site and –Version) describe the test site and test version. Results across the different test sites and four versions of the exam align on the logit scale, indicating that test results are not influenced by test sites or exam versions.

The sixth column (-Criteria) presents the locations of the criterion measures. Again, this facet is negatively oriented, which means that the criteria appearing higher in the column were more difficult than those criteria located lower in the column. Hence, it was much more difficult for CELBAN examinees to receive a high score on grammar compared to other criterions. The seventh column (Scale) maps the four-level CELBAN scale to the logit scale. The lowest scale category 6 and highest scale category 9. With this column, we can visualize the location of all estimates correspond to the CELABN rating scale.

All in all, the study findings indicate that the CELBAN speaking test yielded a relatively high inter- and intra-rater reliability. The severity and leniency of raters from across the country stayed within an acceptable range, so the effectiveness of CELBAN rating scale can be considered as high quality. The overall severity measures are closely distributed to model logit 0.00 (M=0.00, SD=0.18), suggesting that all raters are in the acceptable range of severity and leniency. There was no observed influence of test sites and test versions on test performance either, which provides a strong argument for test validity. These findings help to answer our first research question.

Rating pattern classification and rater-criterion interaction analyses examine rater systematic bias patterns when applying the rating criteria (i.e., research question 2). Based on the rating pattern classification (Engelhard, 2013; Linacre, 2018), rater fit values above 1.50 can be classified as “noisy” or “erratic”, and those with fit values below 0.50 may be classified as “muted”. In other words, “noisy” describes raters who assign extreme unexpected scores, while “muted” indicates raters who demonstrate less variance in their rating patterns than expected. In the present analysis, most raters exhibited “acceptable” rating patterns, but there were six “noisy” (raters 73, 82, 108, 45, 77, 90) raters. For noisy raters and raters with significant rating bias, additional training and calibration is necessary before each administration to support them in re-establishing an internalized set of criteria. In addition to the analysis of rater 32, the rater-criterion interaction analysis results suggest that trainers monitor the performance of rater 105 (severe) and rater 86 (lenient). However, these two raters’ rating accounts are comparatively less than other raters (rater 105—four times; rater 86—twelve times) and such information needs to be considered when interpreting the results. There are 6 “noisy” (raters 73, 82, 108, 45, 77, 90) raters in need of calibration and monitoring as they are more likely to assign extreme unexpected scores. Moreover, training and monitoring are needed for several raters with criteria-specific systematic biases: raters 45, 54, 69, 85, 99 for grammar performance (severe), rater 47 for communication performance (lenient), rater 77 for initiative performance (lenient), Rater 108 for intelligibility performance (lenient) and rater 113 for organization and cohesion performance (severe). Using these results, it is possible to build individual rater profiles that contain information about raters’ performance statistics (fit statistics) and the record of bias tendency (degree of severity/leniency). Trainers can use this profile to not only capture the modifications that raters made along the way but also as a resource for individual feedbacks and rater selection.

The results of criterion measurement suggest that *initiative, intelligibility, strategies, communication, vocabulary* were *easier* criteria than *fluency, organization and cohesion, and grammar*. Due to the relatively large number of responses used for

estimating each difficulty measure (total count per criterion was 5639), the measurement precision was very high. For each criterion, MnSq fit indices stayed well within very narrow quality control limits (i.e., 0.90 and 1.10). This study identified *grammar* as the most difficult criterion for examinees and as the criterion against which most raters have systemic biases; such a result is consistent with earlier and similar findings (McNamara, 1990). The differences observed among rating criterion suggest that the raters hold different views of the meanings of each of the scoring categories towards the relative severity and consistency with which they are applied. An in-depth investigation and review of grammar descriptor is needed to minimize this group-level rater bias. Interviews may be conducted with raters to understand their perceptions and difficulties while applying this criterion. Moreover, group level and individual level rater training is required to support raters adjusting this rating bias and provide consistent feedback.

At the group level, the bias interaction analysis identified 6 biases out of 920 possible iterations (115 raters × 8 criteria). Many raters showed systemic bias towards the grammar criterion. Based on an acceptable range of $-2 \leq \text{logits} \leq 2$, there are 10 raters showing systematic bias against the rating criterion. In detail, raters 45, 54, 69, 85, 99 exhibited a significant severe bias towards *grammar*. Rater 47 exhibited a significant leniency bias towards *communication*. Rater 77 exhibited a significant leniency bias towards *initiative*. Rater 108 exhibited a significant leniency bias towards *intelligibility*. Rater 113 exhibited a significant bias towards *intelligibility and fluency*.

Rater bias may never be eliminated due to rater variance as raters differ in age, gender, and background (McNamara, 1996). Raters' interpretation and use of the same rating scale may vary from rater to rater, and it is unrealistic to require that all raters produce identical rating behaviours and results. However, there are steps that the test can work on to reduce rater bias and improve internal consistency in subsequent ratings. Clear criteria descriptors across levels can effectively assist raters in differentiating the various levels of performance (Moskal & Leydens, 2000), therefore, a review and update of the descriptor may contribute to a reduction in such bias. To facilitate the best use of the rating rubric and consistent use of the rating criteria across scale levels, scoring criteria should (1) ensure the clarity of descriptor, (2) clearly differentiate through descriptors, (3) be consistent in language across scale levels (Tierney & Simon, 2004), and (4) be descriptive and avoid quantifying the performance with wording like "less" "some" "a lot" (Moskal, 2000). In addition to adjusting the criteria descriptor, continued rater training can also reduce rater bias.

To answer research question 3, we turn to Linacre (2002) as an overall guide with eight indicators for assessing the effectiveness of rating scale categories. The study followed these steps to evaluate the psychometric quality of CLB rating scales. Guideline # 1, "At least 10 observations of each category", refers to a minimum of 10 observations for each CLB scale. Guideline # 2, "Regular observation distribution", focuses on detecting irregularities in observation frequency across scales. Guideline # 3, "Average measures advance monotonically with category", means higher categories produce higher average measurement and vice versa. Guideline #4, "Outfit mean-squares less than 2.0", refers to rating scale fit statistics which details the variance between observation and expected value. Guideline #5, "Step calibrations advance", means Rasch-Andrich thresholds should advance monotonically with categories. Guideline #6, "Rating imply measures, and measures imply ratings", states that observed measure of the rating scale should

approximate the expected values predicted by the Rasch model. Guidelines #7 and #8 delineate that step difficulties advance by at least 1.4 logits, but less than 5.0 logits.

In Table 7, the least assigned category was CLB 6 with a frequency of 983, which means that all the rating scales have more than 10 observations. The observed frequency in this study follows a unimodal distribution that smoothly increases from 983 to 26366 and decreases from 26366 to 4709, which indicates that it is unproblematic in terms of rating quality. In this study, average measure monotonically increased corresponds to the rating scale. Outfit values of all rating scales in Table 7 are below 2.0 and close to the ideal value 1.0. This result indicates that the CLB rating scale had an excellent model fit. Figure 2 provides a threshold order of category probability curves for CLB rating scales. As we can see, each scale has one curve and peak and the thresholds are nicely ordered from left to right. The scale thresholds are clearly progressed from CLB 6 and CLB 7 to CLB 9 with −4.35 to 4.64, therefore, this confirms that the thresholds advance monotonically. As we can see from the column of Rasch-Andrich Measure in Table 7, the distances between calibrations of adjacent CLB scales all stayed within the recommended range. In conclusion, the CLB scale analysis meets all indicators of high-quality scale suggested by Linacre (2002).

Conclusion

This study has implications for rater training, ongoing quality assurance and test design. First, the results of this study can be used as a resource providing positive evidence of rater reliability and validity of the CELBAN speaking test. Second, the study can serve as an example for those interested in utilizing Many-Facets Rasch Measurement analysis to assess rater performance within a context of CLB-based assessments. Third, the results offer an informative resource for rater training and calibration, as well as rubric adjustments.

In practice, CELBAN's quality assurance team utilizes these results in planning and implementing additional training and calibration for examinees. Routine quality checks apply these types of psychometric analyses which are performed by a psychometric team. Results are then transferred to the examinee trainer, who collates a report for each examiner which includes both qualitative and quantitative feedback, and a training and calibration plan, if required. In terms of rubrics adjustments, these occur as part of test renewal initiatives. Any changes to the rubrics are made carefully and pilot tested to ensure that results do not vary significantly from those based on earlier versions of the rubric.

Given that CELBAN is a unique example of a CLB-based assessment used in a high-stakes environment, this study contributes to a nuanced understanding of the functioning of CLB-based rating scales in a formal assessment context. Overall, this study indicates a satisfactory degree of intra-rater consistency. The reliability of test results is of critical importance not only to IENs who rely on the results so they can proceed along their professional registration trajectory but also to nursing regulators who must ensure that third party assessments are objective, impartial and fair (http://www.fairnesscommissioner.ca/).

Correspondence should be addressed to Peiyu Wang, Karen Coetzee, Andrea Strachan, Sandra Monteiro and Liying Cheng.
Email: peiyu.wang@queensu.ca; k.coetzee@tsin.ca; a.stratchan@tsin.ca; s.monteiro@tsin.ca; liying.cheng@queensu.ca

Acknowledgements

We would like to thank Christine Amstory (Queen's University) for helping us with the French abstract.

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.

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

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74.

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In Kunna, A. (Ed.), *The companion to language assessment* (pp. 1301-1322). Wiley-Blackwell

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory and Rasch models for speech-language pathologists. *American Journal of Speech-Language Pathology*.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues, *Research Matters, 4*, 22–28.

Brown A (2000). An investigation of the rating process in the IELTS oral interview. In R. Tullon (Ed.), *IELTS Research Reports 2000* (Vol. 3, pp. 49–84). IELTS Australia.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *University of Hawai'i Second Language Studies Paper, 21* (2), 1–44.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly*, *12*(3), 262-282.

Centre for Canadian Language Benchmarks. (2013). *Theoretical framework for the Canadian Language Benchmarks/Niveaux de compétence linguistique canadiens.* Immigration, Refugees and Citizenship Canada.

Cronbach, L.J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika, 16* (3), 297–334.

Douglas, D. (2001). Language for Specific Purposes assessment criteria: where do they come from? *Language Testing*, *18*(2), 171-185.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117-135.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197–221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185.

Eckes, T. (2011*). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Peter Lang.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Epp, L. & Stawychny, M. (2002). *Benchmarking the English Language Demands of the Nursing Profession Across Canada*. Centre for Canadian Language Benchmarks.

Epp. L., & Lewis, C. (2004a). *Developing an Occupation-specific Language Assessment Tool using the Canadian Language Benchmarks*. Centre for Canadian Language Benchmarks.

Epp. L., & Lewis, C. (2004b). *The Development of CELBAN (Canadian English Language Benchmark Assessment for Nurses): A Nursing-specific Language Assessment Tool.* Centre for Canadian Language Benchmarks.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Academic Medicine, 86*(10), S1–S7.

Han, C. (2018). Using rating scales to assess interpretation: Practices, problems and prospects. *Interpreting*, *20*(1), 59-95.

Jeans, M. E., Hadley, F., Green, J., Da Pratt, C. (2005) *Navigating to Become a Nurse in Canada: Assessment of International Nurse Applicants.* Canadian Nurses Association

Johnston, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Koizumi, R., Okabe, Y., & Kashimada. (2017). A Multifaceted Rasch Analysis of Rater Reliability of the Speaking Section of the GTEC CBT. *ARELE: Annual Review of English language Education in Japan, 28*, 241-256. https://doi.org/10.20581/arele.28.0_241

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Kempf, W. F. (1977). Dynamic models for the measurement of traits in social behavior. In W. F. Kempf, E. B. Andersen, & B. H. Repp (Eds.), *Mathematical models for social psychology* (pp. 14–58). Huber.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), 187-217.

Lee, K. R. (2018). Different Rating Behaviors between New and Experienced NESTs When Evaluating Korean English Learners' Speaking. *Journal of Asia TEFL*, *15*(4), 1036-1050.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement*, *3*(1), 85–106.

Linacre, J. M. (2014). *Facets: Many-Facet Rasch-measurement (Version 3.71.4*) [software]. MESA Press.

Linacre, J. M. (2018). *Winsteps® Rasch measurement computer program user's guide*. winsteps.com

Linacre, J. M., & Wright, B. D. (1989). The "length" of a logit. *Rasch Measurement Transactions, 3*(2), 54–55.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246-276.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language testing*, *12*(1), 54–71.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52-76.

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

McNamara, T. (2011). Applied linguistics and measurement: A dialogue. *Language Testing, 28*(4), 435–440.

Messick, S. (1995). Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist, 50* (9), 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Micko, H. C. (1969). A psychological scale for reaction time measurement. *Acta Psychologica*, *30* (1969), 324–335.

Moskal, B. M. (2000). Scoring rubrics: what, when, and how? *Practical Assessment, Research, & Evaluation, 7*(3). http://pareonline.net/getvn.asp?v=7&n=3

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical assessment, research & evaluation*, *7*(10), 71-81.

Newton, P. E. (2009). The reliability of results from national curriculum testing in England, *Educational Research, 51*(2), 181–212.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25–36.

Sebok, S. S., & Syer, M. D. (2015). Seeing things differently or seeing different things? Exploring raters' associations of noncognitive attributes. *Academic Medicine, 90*(11), S50–S55.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use*. Oxford University Press.

Tavares, W., Boet, S., Theriault, R., Mallette, T., & Eva, K. W. (2013). Global rating scale for the assessment of paramedic clinical competence. *Prehospital emergency care, 17*(1), 57–67.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, *9*(2), 1–10.

Touchstone Institute (2016) *CELBAN Speaking Test Renewal. CELBAN Facts & Figures, Issue 3*. Touchstone Institute.

Touchstone Institute (2018) *CELBAN Test Specifications - Internal and confidential.* Touchstone Institute.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305–319.

Wolfe, E. W., Myford, C. M., Engelhard, G., Jr. & Manolo, J. R. (2007*). Monitoring reader performance and DRIFT in the APR English Literature and Composition*

Examination using benchmark essays (Research Report 2007-2). The College Board.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing*, *31*(4), 501–527.