EVALUATING THE VALIDITY AND RELIABILITY OF SPEAKING RUBRICS IN MULTILINGUAL CONTEXTS

Student: ***Toshpolatova Sabina***
Scientific adviser: ***Knonsaidova Maktuba***
Chirchiq State Pedagogical University
Sabinabekmirzayeva04@gmail.com

Abstract: *This thesis investigates the validity and reliability of speaking assessment rubrics used in multilingual educational settings. As classrooms become increasingly diverse, language instructors face the challenge of evaluating oral performance fairly and consistently across speakers of multiple linguistic backgrounds. The study examines how rubric design, rater training, and cultural-linguistic assumptions affect the accuracy and fairness of speaking assessments. Through a mixed-methods approach combining rubric analysis, inter-rater reliability testing, and qualitative interviews with assessors, the research identifies key sources of validity threats and proposes evidence-based recommendations for improving rubric quality. Findings suggest that culturally neutral descriptors, clear performance benchmarks, and systematic rater calibration are critical to achieving equitable and consistent speaking assessment in multilingual contexts.*

Keywords: *speaking rubrics, validity, reliability, multilingual assessment, language testing, rater agreement, oral performance*

Introduction

The assessment of oral language proficiency is among the most complex tasks in applied linguistics. Speaking rubrics — structured scoring guides that define levels of performance across multiple criteria — are widely used in language classrooms, proficiency examinations, and teacher training programs worldwide. Despite their prevalence, the quality of these rubrics varies considerably, and their validity and reliability in multilingual contexts remain underexplored.

Multilingual learners bring diverse phonological systems, discourse conventions, and communicative norms to the speaking classroom. Standard rubrics, often designed with monolingual native-speaker norms as the benchmark, may fail to capture the full communicative competence of multilingual users. This not only undermines the fairness of assessment but also misrepresents students' actual language abilities, potentially leading to consequential decisions based on biased scores.

This thesis addresses three core research questions: (1) Do currently used speaking rubrics demonstrate sufficient construct validity for multilingual learner populations? (2) What levels of inter-rater reliability are achieved when trained raters apply these rubrics? (3) What modifications to rubric design and rater training can improve both validity and reliability in multilingual assessment contexts?

1.1 Background and Motivation

The growth of multilingual education globally has accelerated interest in fair and inclusive assessment practices. Studies by Fulcher 2003, Knoch 2009, and Schissel et al. 2019 highlight persistent concerns about the cultural embeddedness of rating scales and the subjectivity of rater judgments. In many educational systems, particularly those transitioning from monolingual to multilingual instructional models, legacy rubrics developed for homogeneous learner populations continue to be applied without critical review.

The motivation for this study stems from observed inconsistencies in speaking assessment outcomes at the institutional level and from a broader need to align assessment practices with contemporary theories of multilingual communicative competence, including the concept of translanguaging Garcia & Wei, 2014 and plurilingual competence Council of Europe, 2001.

1.2 Scope and Limitations

This study focuses on rubrics used in EFL and ESL academic speaking contexts at the tertiary level. The data were collected from university-level settings. The study does not evaluate computerized speech assessment systems or automated scoring tools, which represent a separate and emerging area of research.

Additionally, findings should be interpreted with caution regarding their generalizability beyond the specific cultural and institutional contexts examined.

Literature Review

2.1 Validity in Language Assessment

Validity, in the context of language testing, refers to the degree to which a test measures what it claims to measure Messick, 1989. Modern validity theory, as articulated by Kane 2006, conceptualizes validity as an argument — a coherent set of inferences from test scores to intended interpretations and uses. For speaking rubrics, validity encompasses content validity whether the rubric criteria cover the full domain of speaking ability, construct validity whether the rubric reflects theoretically sound constructs of oral communication, and consequential validity whether rubric-based scores lead to fair and appropriate decisions.

In multilingual contexts, construct validity is particularly at risk when rubrics privilege native-speaker norms of fluency, pronunciation, and pragmatic behavior. Research by Levis 2005 and Jenkins 2007 argues that intelligibility — rather than nativeness — should serve as the primary standard for pronunciation assessment. Similarly, Kramsch 2009 questions the normativity of monolingual communicative competence models in diverse language learning environments, calling for assessment frameworks that recognize the hybrid, fluid nature of multilingual communication.

2.2 Reliability and Rater Variability

Reliability refers to the consistency of scores across raters, occasions, and tasks. Inter-rater reliability IRR is a central concern in speaking assessment, as human raters inevitably bring subjective judgments to the scoring process. Eckes 2011 identifies several rater effects — including halo effects, leniency/severity biases, and central tendency errors — that systematically distort scores. Generalizability theory Brennan, 2001 and Many-Facet Rasch Measurement Linacre, 2020 provide powerful statistical frameworks for estimating and controlling rater variance.

Studies consistently report moderate to low IRR values for holistic speaking rubrics, with Cohen's kappa values often falling below 0.70 Weigle, 2002. Analytic rubrics tend to yield higher reliability coefficients but require more rater training and impose greater cognitive load on assessors. The trade-off between reliability and practicality is a recurring tension in rubric design literature.

2.3 Speaking Rubrics in Multilingual Settings

Research on rubric performance in multilingual contexts is relatively recent but growing. Shohamy 2011 critiques the ideological assumptions embedded in standard language testing frameworks, arguing that tests designed around monolingual norms reproduce linguistic hierarchies that disadvantage multilinguals. More practically, studies by Grabowski 2009 and De Costa et al. 2017 demonstrate that raters often penalize multilingual speakers for accent features that do not impede communication, reflecting prescriptivist biases rather than objective performance criteria. Emerging frameworks such as Dynamic Assessment (Poehner, 2008) offer more equitable approaches by foregrounding communicative effectiveness over surface-level accuracy.

Methodology

3.1 Research Design

This study employs a mixed-methods design, integrating quantitative analysis of rater agreement data with qualitative investigation of rater perceptions and rubric content. The mixed-methods approach enables triangulation of findings and provides both statistical robustness and interpretive depth Creswell & Plano Clark, 2018. The sequential explanatory design proceeds from quantitative data collection and analysis to qualitative follow-up, with the latter used to explain and contextualize statistical findings.

3.2 Participants and Materials

Six existing speaking rubrics were selected for analysis, drawn from widely used proficiency frameworks IELTS, CEFR, TOEFL iBT as well as locally developed institutional rubrics. Twelve trained raters — six EFL/ESL instructors and six graduate students in applied linguistics — independently scored 30 audio

recordings of student presentations by speakers with L1 backgrounds including Arabic, Mandarin, Russian, Uzbek, French, and Spanish. All raters attended a two-hour orientation session but did not receive rubric-specific calibration training prior to the first scoring round, allowing baseline reliability levels to be measured authentically.

3.3 Data Collection and Analysis

Quantitative data were analyzed using intraclass correlation coefficients ICC and Many-Facet Rasch Measurement via FACETS software to estimate rater severity, rubric dimensionality, and score dependability. ICC values were interpreted using Koo and Mae's 2016 benchmarks: values below 0.50 indicating poor reliability, 0.50-0.75 moderate, 0.75-0.90 good, and above 0.90 excellent. Qualitative data from semi-structured interviews with raters were analyzed using thematic analysis Braun & Clarke, 2006, focusing on raters' interpretive strategies, perceived rubric ambiguities, and cultural assumptions in scoring decisions.

Findings and Discussion

4.1 Validity Findings

Content analysis of the six rubrics revealed significant variation in the constructs operationalized. Criteria related to pronunciation ranged from accent-focused descriptors — which implicitly reference native-speaker norms — to intelligibility-based descriptors emphasizing listener comprehension. Rubrics derived from the CEFR demonstrated stronger alignment with plurilingual competence models, while locally developed rubrics showed greater construct ambiguity and cultural specificity. In particular, terms such as "natural delivery," "appropriate register," and "standard grammar" were interpreted inconsistently by raters assessing speakers from non-Western discourse traditions.

Rater interviews confirmed that vague criterion labels generated interpretive inconsistency. Several raters reported unconsciously applying L1-influenced standards when assessing pragmatic features such as turn-taking, directness, and politeness conventions. These findings corroborate Bachman and Palmer's 2010)”

argument that rubric construct validity depends not only on criterion design but also on the cultural interpretive frameworks raters bring to the scoring task.

4.2 Reliability Findings

ICC values for overall speaking scores ranged from 0.61 to 0.83 across the six rubrics, indicating moderate to good inter-rater reliability. Analytic rubrics consistently outperformed holistic rubrics on reliability measures. FACETS analysis revealed significant rater severity differences logit range: -1.2 to +1.6, suggesting that without calibration, score comparability across raters cannot be assumed. Reliability was lowest on pronunciation and pragmatic appropriateness subscales — precisely the dimensions most susceptible to cultural bias. Rater training in a second round reduced severity discrepancies by approximately 30%, demonstrating the practical value of systematic calibration.

The findings point to three key principles for rubric improvement in multilingual contexts. First, criterion anchoring: descriptors should be formulated around concrete, observable communicative behaviors rather than evaluative labels open to cultural interpretation. Second, intelligibility-based standards: pronunciation criteria should explicitly prioritize comprehensibility to a diverse audience rather than proximity to native-speaker phonological norms. Third, multilingual competence recognition: rubrics should include indicators that acknowledge translanguaging and cross-linguistic pragmatic awareness as positive communicative resources rather than deficiencies. These principles align with the ALTE Code of Practice for inclusive language assessment.

Conclusion

This thesis has demonstrated that speaking rubrics commonly used in educational settings exhibit significant validity and reliability limitations when applied to multilingual learner populations. The construct assumptions embedded in many rubrics — particularly regarding pronunciation norms and pragmatic conventions — reflect monolingual ideologies that do not account for the communicative repertoires of multilingual speakers.

The study makes three principal contributions. It provides empirical evidence of rubric-related validity threats in multilingual contexts; it quantifies inter-rater reliability across rubric types using robust psychometric methods; and it offers a set of design principles grounded in contemporary multilingual language assessment theory. These contributions serve both practical and theoretical purposes, offering assessors actionable guidelines while advancing scholarly discourse on equitable assessment.

Future research should explore the role of rater language background in scoring multilingual speakers, investigate AI-assisted rubric calibration tools, and examine student perspectives on assessment fairness. As language education continues to evolve toward plurilingual and translanguaging pedagogies, assessment frameworks must keep pace — ensuring that rubrics serve not as gatekeeping instruments but as equitable tools for recognizing and developing multilingual communicative competence.

REFERENCES

Bachman, L. F., & Palmer, A. S. (2010). Language Assessment in Practice. Oxford University Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101.

Brennan, R. L. (2001). Generalizability Theory. Springer.

Council of Europe. (2001). Common European Framework of Reference for Languages. Cambridge University Press.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and Conducting Mixed Methods Research (3rd ed.). SAGE.

De Costa, P. I., Tigchelaar, M., & Cui, Y. (2017). Ideology and emotion in L2 pronunciation. TESOL Quarterly, 51(3), 715-724.

Eckes, T. (2011). Introduction to Many-Facet Rasch Measurement. Peter Lang.

Fulcher, G. (2003). Testing Second Language Speaking. Longman.

Garcia, O., & Wei, L. (2014). Translanguaging: Language, Bilingualism and Education. Palgrave Macmillan.

Grabowski, K. (2009). Investigating construct validity of a test measuring grammatical and pragmatic knowledge in speaking. TC Working Papers in TESOL & Applied Linguistics, 9(2).

Jenkins, J. (2007). English as a Lingua Franca: Attitude and Identity. Oxford University Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 17-64). ACE.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating processes. Assessing Writing, 14(2), 97-115.

Koo, T. K., & Mae, M. Y. (2016). A guideline for selecting and reporting ICC. Journal of Chiropractic Medicine, 15(2), 155-163.

Kramsch, C. (2009). The Multilingual Subject. Oxford University Press.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. TESOL Quarterly, 39(3), 369-377.

Linacre, J. M. (2020). FACETS: Rasch Measurement Computer Program (Version 3.83). Winsteps.com.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). ACE.

Poehner, M. E. (2008). Dynamic Assessment: A Vygotskian Approach to L2 Development. Springer.

Schissel, J. L., Leung, C., & Chalhoub-Deville, M. (2019). The construct of multilingualism in language testing. Language Assessment Quarterly, 16(4-5), 373-388.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. Modern Language Journal, 95(3), 418-429.

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