2025, Vol 4, No 2, 499-508
https://doi.org/10.29329/pedper.2025.103

REVIEW ARTICLE Open Access

Assessing language proficiency: A Rasch analysis of ELT assessment tools

Neda Kianinezhad[1]

[1] Hakim Sabzevari University, Iran.

Abstract

This paper now addresses the important role of Rasch analysis in ensuring high-quality assessments in the English Language Teaching (ELT) context. Assessment is at the core of ELT; it verifies student learning and lends an ear to teaching methodologies. Current assessment methodologies often fail to measure learners' competencies accurately. Rasch analysis, on the other hand, provides a process to assess test items so that they are as fair and equal as possible in a given assessment. Rasch analysis converts raw scores into interval-level measurements, allowing for precise comparisons between learners. The tool is an important driver in identifying central items, such as item bias and DIF (Differential Item Functioning), that must be addressed for fair assessments. In this review, we present the strengths of Rasch Analysis in a wide range of ELT contexts, including reading, listening, speaking, and writing assessments. Although Rasch analysis has strengths, it is not without limitations, such as requiring large sample sizes and understanding the results' meaning. Even so, it helps in knowing more about how assessment quality can be improved. It concludes by summarizing the impact of Rasch on assessment, providing a fair and valid method for measuring from both learners' and educators' perspectives in various types of educational settings. Further studies should be conducted to demonstrate its utility in smaller-scale designs and to investigate its effects in conjunction with other assessment methodologies.

Keywords: Language proficiency, Rasch analysis, ELT assessment tools

Introduction

In English Language Teaching (ELT), assessment serves several purposes, including monitoring student learning, informing teaching practices, and evaluating approaches to teaching. As ELT assessment is a complex task in that many aspects of the assessment exist on a multidimensional scale, tools and methodologies need to be employed that can objectively evaluate learner performance, yet at the same time guarantee equity (i.e. getting each learner a fair chance, irrespective of their background) and reliability (i.e. maintaining consistency from one evaluation to the next, across time and across assessors). Among these approaches, Rasch analysis has proven itself to be a potent tool for assessing language, with a rigorous methodology for calibrating test items and validating assessment tools.

Moreover, introduced by Georg Rasch in 1960, Rasch analysis is a probabilistic model that maps raw test scores to interval-level measurements. In contrast to traditional methods, the underlying assumption of the Rasch statistic model is that there is a direct relationship between an

[1] Footnote

*Corresponding Author: kianinezhad.neda@gmail.com

Received 29.01.2025 Revised 21.03.2025 Accepted 07.04.2025

individual's ability and the item difficulty. This relationship influences the prediction of the response, meaning the probability of correctness depends on both the individual's ability and the difficulty of the test item. For instance, in a reading comprehension test, Rasch analysis can determine whether certain questions disproportionately favour native speakers over non-native speakers. This model provides detailed insights into the behaviour of test items and facilitates the creation of measurement scales that remain consistent across different populations (Bond Fox, 2015).

In fact, in ELT, assessment often includes a variety of constructs, including reading comprehension, listening fluency, and speaking ability. In this regard, Rasch analysis has some unique advantages, such as allowing measurement of problematic items, i.e., item bias, differential item functioning (DIF), and poorly fitting items, which are necessary for test validity and reliability (Alavi et al., 2011; Baghaei & Robitzsch, 2025; Kianinezhad, 2024; Kianinezhad & Kianinezhad, 2025; McNamara, 1996). Furthermore, this model can be used to design tests that are consistent with international standards, thereby making test items more specific and transferable to different linguistic and cultural contexts (Wright & Masters, 1982).

Accordingly, this paper discusses the critical contribution of Rasch analysis in leading the path of validity and reliability in English Language Teaching (ELT) assessment tools. The growing need for accurate and fair assessments of language proficiency calls for robust psychometric methods. Rasch analysis, a sophisticated model from item response theory (IRT), provides a strong framework for meeting this demand. In contrast to classical test theory (CTT), which produces level-specific data, it produces interval-level data, thus enabling the capture of finer learner capacities and characteristics of test items. Moreover, this review discusses the utility of Rasch analysis applied to different ELTs. It highlights the model's ability to address significant challenges, such as item bias, rater inconsistency, and differential item functioning (DIF). At the same time, the paper acknowledges the limitations of Rasch analysis, including its reliance on large sample sizes and its computational complexity. This review aims to highlight the key contributions of Rasch analysis in various ELT assessment applications by synthesising evidence from studies (Farlie, 2021; O'Brien, 1989; Hamp-Lyons, 1989) that employ Rasch analysis in diverse aspects of ELT assessment. It also provides a balanced perspective by discussing its challenges, offering a comprehensive understanding of its role in language assessment.

Literature review

Rasch analysis: A foundational overview

Rasch analysis is a perennial technique in modern psychometrics and constitutes an ideal setting for assessing the quality of tests. As compared to classical test theory (CTT), which assumes a sum score to be dependent on a small amount of information from each item and from each examinee, Rasch models the item stability of correct answer as a function of examinee's ability level and item difficulty. For instance, imagine a dataset where Rasch analysis reveals biased items in a listening test, disproportionately affecting non-native speakers. In this paradigm, one scale of measurement is specified, within which both the subject's estimation and the item difficulty are discretised (i.e., in terms of logits [logit-odds units], respectively). One of the most attractive features of this model is its ability to produce interval-level data, which enables the meaningful assessment of differences in learning abilities across a range of learners and the identification of items with significantly differing results between learner subgroups (Farlie, 2021;

O'Brien, 1989; Hamp-Lyons, 1989).

Based on Kianinezhad (2024), several fundamental principles underpin Rasch analysis. One such principle is item difficulty, which refers to the probability of an item being answered correctly, irrespective of a person's skill level. Another principle is person ability, defined as the capacity of a subject to demonstrate their knowledge or skills in the tested construct or task. Next, Item fit statistics, such as INFIT (information-weighted fit) and OUTFIT (outlier-sensitive fit), are used to verify an item's appropriateness to the Rasch model's demands. Items whose values are very close to one are well-fitted items, while items with larger values are often not fitted (biased or ambiguous). A key issue is the assessment of reliability indices, such as person separation reliability, which reflects the test's ability to discriminate between levels of expertise among subjects. Since smaller standard error values are associated with greater accuracy in the estimation of these contrasts, larger standard error values are associated with reduced accuracy in the estimation of these contrasts (Farlie, 2021; O'Brien, 1989; Hamp-Lyons, 1989).

Therefore, Rasch analysis provides a range of models tailored to various types of data and assessment formats. In principle, the one-parameter logistic model (1PL) assumes equal discriminability for all items. The Partial Credit Model (PCM) also achieves the highest performance when items to be assessed are polytomous (i.e., questions with more than one plausible correct answer). The Rating Scale Model (RSM) is modified for ordinal data (i.e., the answer is ordinal in the sense that answer categories are at equal distances along a latent trait continuum). The specific form of the model depends on the measure to be taken and the attribute to be predicted. Items of a reading assessment multiple-choice, for example, excerpts from a reading assessment set can be characterized even more plausibly by the 1PL model or by the lower level process, e.g., the essay writing process, that is judged by its rubric performance, can be characterized even more plausibly by the RSM (Hamp-Lyons, 1989; Lee-Ellis, 2009; Khalaf, 2022). Hence, flexibility, accuracy, and the ability to derive deeper insights make Rasch analysis a powerful technique for creating and validating tests.

Limitations of Rasch analysis in ELT assessment

Although Rasch analysis is a valuable method for improving assessment quality, it has several limitations that should be considered, particularly when applied to English Language Teaching (ELT) assessments. To address this, future research could explore methods to adapt Rasch analysis for smaller classroom settings. One of the biggest criticisms of Rasch analysis is its requirement for a large sample size: it needs a substantial sample size to reliably and stably capture the parameters. With small sample sizes, the results become less reliable, and any claims constructed about dimensionality (item difficulty or learner ability) may not be accurate. This limitation, hence, means that Rasch analysis is less applicable for small-scale studies or classroom assessments (Holster, 2012).

Moreover, another key limitation is the complexity of interpreting the output from a Rasch analysis script. It provides numerous metrics, including logits, item fit statistics, and reliability indices, which can only be utilised after being analysed by someone familiar with the work. Many ELT practitioners may lack psychometric expertise or have limited access to sophisticated software tools, which can hinder the implementation of this method in a more advanced manner (Farlie, 2021).

Plus, Rasch is a unidimensional model, which means that items measure only a single latent dimension. However, several ELT instruments are multidimensional and aim to assess multiple skills, including reading comprehension, vocabulary, grammar, and writing fluency tasks. Using a unidimensional Rasch model on these kinds of complex assessments may lead to faulty results. Under this analysis, multi-dimensional IRT (IRT) models might be more suitable as they can better capture the jag designed for multiple constructs in language tests (Inoue, 2016).

Rasch analysis also emphasizes psychometric properties that can overshadow other important aspects of assessment, including item difficulty and person ability. Using the single task as a unit of analysis, factors such as the actuality of tasks, the real-world utility of tests and ease of use in classrooms are rarely addressed adequately by Rasch Analysis. They are essential components in designing meaningful and feasible assessments in ELT, but are superfluous to the analytical Rasch framework.

As a result, some of the limitations mentioned earlier can be minimised by integrating Rasch analysis with other assessment methods to maximise the quality of assessment. For example, combining Rasch analysis with evidence of content validity, task authenticity, and practicality helps to create assessments that are both psychometrically valid and practically useful. This approach can be used, for instance, to deliver a genuine language test assessment.

Accordingly, this method ensures that assessments are accurate and fair to learners, while also being meaningful for educators. Hence, Rasch analysis is an incredible tool for improving the reliability and validity of ELT assessments, but it comes with its own limitations. Recognising these shortcomings and adopting a balanced approach can help address the full scope of Rasch analysis from an objective standpoint in language teaching and assessment.

Advantages of Rasch analysis in ELT assessment

Rasch analysis offers numerous benefits for English Language Teaching (ELT) assessment, particularly when compared to classical test theory (CTT). Its primary advantage lies in its ability to generate interval-level data, enabling more precise and interpretable comparisons of learner abilities. Unlike CTT, which provides only an ordinal ranking, Rasch analysis allows scores to be meaningfully represented on a common scale, regardless of the type of items or the testing conditions. This is achieved by modelling item difficulty and learner ability on the same linear scale, ensuring that performance discrepancies are assessed with greater accuracy. As a result, Rasch analysis provides deeper insights into learners' development and their level of expertise (Farlie, 2021). For example, a placement test using Rasch analysis can group learners by true proficiency levels, avoiding issues caused by flawed or biased items.

Another important feature of Rasch analysis is that it enables the production of a single measurement scale, independent of the test items employed. This reduces the variability of assessment and the potential for bias, which is especially crucial in ELT, since learners are typically speakers of linguistically and culturally diverse backgrounds. Although traditional assessment methodologies carry a risk of unfairly benefiting particular groups, Rasch analysis can promote fairness and reliability across all learners, regardless of their backgrounds or learning environments (Farlie, 2021).

Furthermore, one of the strongest features of Rasch analysis is its ability to detect and correct item bias through differential item functioning (DIF) analysis. DIF analysis reveals test items that

differ differentially between one group of learners and another, even if overall ability is equal. As an example, a test question may be more accessible to speakers of one language than to speakers of others. Such bias can result in unfair assessments. By using Rasch analysis, it is possible to identify problematic items and adapt or eliminate them, thereby increasing test validity and potentially enhancing test fairness (Turkan, 2012; Medvedev, 2019).

Likewise, the increases in reliability and validity offered by Rasch analysis have value for learning and teaching. Reliable assessments mean that learners' abilities are measured accurately, leading to fairer placement decisions and more effective instructional planning. For example, placement tests that utilise Rasch analysis are better equipped to group learners by their true proficiency levels, thereby avoiding issues caused by flawed or biased test items. Furthermore, program evaluations using this kind of judgment are stronger, as they represent more naturalistic student development (O'Brien, 1989).

Rasch analysis also improves the interpretability of assessment results. Since the scores are on the same scale, they can be used to cross-compare and track learners' achievement across different test administrations. This functionality is handy in high-stakes assessments or longitudinal studies, where stability and equivalence are paramount (Lee-Ellis, 2009). All in all, by generating accurate data, detecting inaccuracies, and ensuring fair measurement for diverse populations of learners, it contributes to improved pedagogical decision-making in teaching, placement, and program assessment. These benefits make Rasch analysis a vital tool for language teachers and researchers, leading to better outcomes for learners and the ELT community as a whole.

Applications of Rasch analysis in ELT assessment

Rasch analysis has proven to be a valuable tool for English Language Teaching (ELT) assessment, improving several facets, including reading, listening, speaking, and writing. By offering a standardized framework for evaluating the quality of tests, Rasch analysis helps ensure that language assessments are fair, reliable, and valid.

Reading assessments and Rasch analysis

In the field of reading assessments, Rasch analysis has been a major contributor to increasing the quality of reading comprehension tests. Research also demonstrates its usefulness in determining item difficulty, detecting bias, and assessing general test validity. For example, the tool has been used to investigate whether a single score is a suitable measure of a child's reading competence and whether the test accurately measures reading understanding (Chen, 2018).

One of the standout features of Rasch analysis is its ability to identify items that may be too easy or too challenging for the intended audience. By revising or removing such items, educators can create more balanced assessments that genuinely reflect the diverse abilities of learners (O'Brien, 1989). Additionally, the differential item functioning (DIF) analysis, an integral part of Rasch analysis, helps identify biased items that could disadvantage specific groups, such as students from different language backgrounds or those with varying levels of reading experience (Bakri, 2022; Bakri, 2023). Through the resolution of these problems, we can create fair and valid assessment of reading for all students.

Listening assessments and Rasch analysis

Listening assessments also gain significant advantages from Rasch analysis. Researchers have applied this approach to investigate factors like item bias, rater effects, and the influence of audio characteristics on performance. For example, it can reveal whether specific accents disadvantage second-language listeners, contributing to more inclusive tests. By analysing the relative difficulty of various listening tasks, such as retrieving specific information or making inferences, Rasch analysis provides insight into the cognitive demands of different items.

Additionally, it examines the impact of variables such as speech accent, speech rate, and background noise on test performance. For instance, Rasch analysis can reveal if certain accents are difficult for listeners exposed to a second language, thereby contributing to the development of more inclusive listening tests. These results aid those who develop such tasks in creating fairer and more realistic listening scenarios.

Speaking assessments and Rasch analysis

Rasch analysis encompasses several models, such as the one-parameter logistic model (1PL), the partial credit model (PCM), and the rating scale model (RSM). Each of these models is tailored to specific types of data and assessment formats. When discussing speaking assessments, Rasch analysis, particularly in the form of Many-Faceted Rasch Measurement (MFRM), has played a crucial role in enhancing test quality and integrity. MFRM enables the simultaneous assessment of multiple sources of variation, such as rater variance, item difficulty, and test-taker ability. This approach can be used to identify raters who tend to be overly generous or too pedantic in their evaluations, as well as items on which a subject's ratings may be biased (Choi, 2021; Leeming, 2022). For instance, MFRM can identify raters who are overly lenient or strict, ensuring fair evaluations.

Rasch analysis has also been instrumental in the creation and calibration of speaking assessment rubrics. By analyzing the repeatability of raters' use of scoring criteria, researchers can ensure that rubrics are unambiguous and consistently interpreted by different raters. This reduces subjectivity in scoring and enhances the reliability and validity of speaking evaluations (Bijani, 2017).

Writing assessments and Rasch analysis

Writing assessments, which are sometimes judged on subjectivity, have also been greatly enhanced through the use of Rasch analysis. Researchers have applied this approach to assess the psychometric properties of rubrics for essay grading, where issues of rater bias and variations in prompts also play a role in test results (Shirazi, 2019).

Additionally, using Rasch analysis, items (e.g., essay prompts) that are too easy or too difficult for the test-taker can be identified and adjusted to ensure balanced test-taking. For example, it can identify essay prompts that are disproportionately difficult for certain groups, improving test non-discrimination. Additionally, it helps detect raters who may consistently score leniently or strictly, allowing for targeted training (O'Brien, 1989). This approach has also refined the selection criteria to be clearly defined and uniformly applied, thereby increasing reliability and fairness in the assessment process (Aryadoust, 2017; Lee-Ellis, 2009).

Addressing rater effects and bias through Rasch analysis

Rater effects and bias are significant issues within ELT assessment, particularly in subjective performance tasks (e.g., speaking, writing). Biasedness and reliability may be compromised because of factors like rater severity, leniency, inconsistency in the use of scoring rubrics, and external factors (e.g., the test-taker's gender or mother tongue, respectively). The Many-Facet Rasch Model (MFRM) offers a practical resolution to these issues.

In contrast to a simplified Rasch model, MFRM accounts for more than one dimension, including participants, items, and subjects, in its model. This provides researchers with the ability to make comparisons of rater behaviour, item difficulty, and test-taker performance simultaneously. The knowledge derived can help determine sources of variance and identify problems related to rater severity or leniency (Holster, 2012; Leeming, 2022; Wang, 2020).

For example, MFRM may be possible to detect raters whose scores are systematically higher or lower than those of their peers. It can also detect biases, such as when a rater tends to favour or penalise certain groups of test-takers or specific types of items. Moreover, this data is quite helpful in enhancing scoring practices via more skilled rater training and calibration. Studies confirm that training programs designed using MFRM findings can significantly enhance rater consistency and reduce bias (Bijani, 2017, 2019), resulting in a more reliable and equitable assessment process for everyone involved.

MFRM and its role in rater training

MFRM provides rich, granular feedback, making it an ideal resource for improving rater training programs. From performance data, MFRM produces a separate report for each rater, flagging their individual gains and losses. This allows for training interventions that are specifically tailored to the needs of each rater.

For example, a rater who consistently assigns lower scores might benefit from additional guidance on interpreting the scoring rubric or understanding specific aspects of the task. Additionally, when a rater has a bias towards certain groups or items, personalized feedback can be used to alert the rater to the bias and help them address it (Wang, 2020; Sundqvist et al., 2020).

Therefore, research indicates that this personalized approach to rater training improves scoring consistency and impartiality (Shirazi, 2019). By imposing control over both rater bias and rater variance, MFRM enables the portfolio of assessments to be not only more reliable but also fairer, benefiting both raters and test-takers.

Technological advancements and Rasch analysis in ELT assessment

ELT assessment methodology has changed with the introduction of technology (e.g., computer-based assessments [CBA] and automated scoring systems). These technologies offer a range of advantages, including high efficiency, low cost, and the ability to mitigate biased scoring. However, these technological breakthroughs need to be rigorously tested to demonstrate their effectiveness and equity. Rasch analysis is a valid method for this and similar purposes.

With Rasch analysis, clinicians or language assessment practitioners can analyse the psychometric properties of items related to their use of computer-based assessments (CBAs). That is, it ensures that each item functions as intended and does not exhibit any bias. Rasch analysis also provides a robust platform for comparing the performance of human raters to

automated scoring systems. For instance, Kopparapu (2024) employed Rasch analysis to evaluate automated scoring algorithms, highlighting their strengths and limitations.

Equally important, research using Rasch analysis has also shown differential levels of correspondence between human and automated scoring in ELT assessments. These results emphasize the potential of automatic scoring systems, but also bring to light their limitations. For instance, whereas automated systems are highly proficient in quantifiable aspects (e.g., grammar, vocabulary) of language, they are less effective in evaluating more qualitative aspects (e.g., fluency, coherence) in speaking evaluations. Rasch analysis is an essential step in determining these deficits and provides a framework for developing more robust and accurate automatic scoring methodologies (Nelson, 2023).

In addition, the recent progress of natural language processing (NLP) and machine learning (ML) provides the possibility for addressing these. Besides, these technology-based systems can also contribute to further improving the accuracy and effectiveness of computerised scoring systems. However, as these systems continue to evolve, Rasch analysis will remain a crucial tool for evaluating the performance of systems and ensuring they continue to meet the high standards required for fair and valid ELT assessment.

Conclusion

In summary, Rasch analysis is a valuable tool that enhances the validity and equity of assessments in English Language Teaching (ELT). It does not matter if reading, hearing, talking, or writing, Rasch analysis allows for making assessments much more precise and equitable tests. In contrast to the conventional approach, it transforms raw test scores into rich, valid data, enabling educators to measure students' abilities efficiently. It also helps identify issues such as biased questions, unfair test items, and inconsistencies in scoring, ensuring assessments are fair for all students, regardless of their background.

Additionally, the significant advantage of Rasch analysis lies in its potential to expose and rectify issues related to test design, specifically identifying items that are problematic or biased. It leads to more accurate assessments that actually reflect measures of language ability, providing both educators and students with a more accurate assessment of their language abilities. Furthermore, Rasch analysis also contributes to refining test items and test rubrics so that they meet the criteria of both consistency and validity, thereby having an overall impact on the accuracy of the tests. However, Rasch analysis does have some limitations. It performs better with large sample sizes, and it points out that the interpretation of results is complex. Therefore, this could be a barrier for educators who lack training in psychometrics. Although there are these difficulties, when applied in conjunction with other techniques, Rasch analysis generates a comprehensive assessment of the quality of an assessment, considering both statistical validity and operational significance.

As pedagogical technologies continue to evolve, Rasch analysis will remain a valuable tool for evaluating automated scoring systems and ensuring that they deliver fair and accurate results. Ultimately, Rasch analysis enhances language testing by making assessments more precise, fair, and reliable, benefiting both learners and teachers in the long run.

Implications and future directions

The findings of this research carry significant implications for the future of assessment in English

Language Teaching (ELT). Rasch analysis underscores the critical need for more reliable and equitable assessment practices. Its ability to detect and rectify issues such as item bias, poorly calibrated questions, and rater inconsistencies enables a more accurate measurement of true language proficiency. This is particularly vital for promoting fair learning environments in heterogeneous classrooms with students from diverse cultural and linguistic backgrounds. By leveraging Rasch analysis, educators and institutions can work toward ensuring that every learner is assessed fairly and accurately, regardless of their background.

Looking ahead, several promising avenues for future research emerge. First, there is a need to adapt Rasch analysis for smaller classroom settings and niche learner populations to enhance its practicality for widespread use among teachers. Second, integrating Rasch analysis with complementary frameworks, such as task-based language assessment, could provide a more holistic measurement of naturalistic language production, thereby bridging the gap between theoretical constructs and real-world language use. Third, as educational technology continues to evolve, research should explore how Rasch analysis can be embedded within digital assessment platforms, including AI-driven tools and automated scoring systems, to ensure their neutrality and accuracy.

Additionally, expanding the application of Rasch analysis in multilingual and multicultural contexts represents another crucial direction for future research. By focusing on Differential Item Functioning across diverse populations, researchers can develop assessments that are not only statistically robust but also culturally sensitive and responsive to the needs of these populations. In conclusion, while Rasch analysis holds immense potential for revolutionising ELT assessment, realising this potential requires continued innovation in its applications, integration with emerging methodologies, and proactive addressing of its current limitations. Through such efforts, the field can advance toward assessments that genuinely capture language competence in ways that are objective, valid, and meaningful for all learners.

References

Alavi, S. M., Ali, R. A., & Amirian, S. M. R. (2011). Academic discipline DIF in an English language proficiency test. *Journal of English Language Teaching and Learning*, *7*, 39–65.

Aryadoust, V. (2017). Adapting levels 1 and 2 of Kirkpatrick's model of training evaluation to examine the effectiveness of a tertiary-level writing course. *Innovation in Language Learning and Teaching*, *11*(2), 107–120. https://doi.org/10.1080/1554480X.2016.1242426

Baghaei, P., & Robitzsch, A. (2025). A tutorial on item response modelling with multiple groups using TAM. *Educational Methods and Practice*. *(Yayına Hazırlık Aşamasında)*

Bakri, H. (2022). Evaluating and testing English language skills: Benchmarking the TOEFL and IELTS tests. *International Journal of English Linguistics*, *12*(3), 99–110. https://doi.org/10.5539/ijel.v12n3p99

Bakri, H. (2023). Evaluating and testing English language skills: Benchmarking the TOEFL and IELTS tests. *Bulletin of Faculty of Languages and Translation*. https://doi.org/10.21608/bflt.2023.319543

Bijani, H., & Khabiri, M. (2017). The impact of raters and test-takers' gender on oral proficiency assessment: A case of multifaceted Rasch analysis. *Journal of Teaching Language Skills*, *36*(1), 115–138. https://doi.org/10.22099/jtls.2017.25897.2290

Bijani, H. (2019). Evaluating the effectiveness of the training program on direct and semi-direct oral proficiency assessment: A case of multifaceted Rasch analysis. *Cogent Education*, *6*(1), 1670592. https://doi.org/10.1080/2331186x.2019.1670592

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Chen, P., & Fu, J. (2018). Examining measurement properties of the revised Preschool Language Assessment for use with Mandarin-speaking children. *Language Assessment Quarterly*, *15*(2), 152–167. https://doi.org/10.1080/15434303.2018.1529176

Choi, H., & Lee, H. (2021). Effects of examinees' perceptions of interlocutor proficiency on paired oral assessment results. *The Journal of Asia TEFL*, *18*(1), 23–40. https://doi.org/10.18823/ASIATEFL.2021.18.1.2.23

Farlie, M., Johnson, C., Wilkinson, T. J., & Keating, J. (2021). Refining assessment: Rasch analysis in health professional education and research. *Focus on Health Professional Education*, *22*(2), 16–28. https://doi.org/10.11157/fohpe.v22i2.569

Hamp-Lyons, L. (1989). Applying the partial credit method of Rasch analysis: Language testing and accountability. *Language Testing*, *6*(1), 1–9. https://doi.org/10.1177/026553228900600109

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *Language Awareness*, *25*(1–2), 4–20. https://doi.org/10.1080/09571736.2015.1130079

Khalaf, M. A., & Omara, E. M. N. (2022). Rasch analysis and differential item functioning of English Language Anxiety Scale (ELAS) across sex in the Egyptian context. *BMC Psychology*, *10*(1), Article 1. https://doi.org/10.1186/s40359-022-00955-w

Kianinezhad, N. (2024). Validation of the adapted attitudes toward online teaching scale for Iranian EFL teachers: A Rasch model analysis. *Education Mind*, *3*(1), 31–45.

Kianinezhad, N., & Kianinezhad, M. (2025). Comparative evaluation of C-test reliability using classical and modern psychometric methods. *Language Education & Assessment*, *8*(1), 2279–2295.

Kopparapu, S., & Panda, A. (2024). Unified spoken language proficiency assessment system. In *Proceedings of Oriental COCOSDA International Conference on Speech Database and Assessments*. https://doi.org/10.1109/O-COCOSDA64382.2024.10800105

Lee-Ellis, S. (2009). The development and validation of a Korean C-test using Rasch analysis. *Language Testing*, *26*(2), 245–274. https://doi.org/10.1177/0265532208101007

Leeming, P., & Harris, J. (2022). Measuring speaking proficiency growth in the language classroom: An investigation of practical approaches for teachers. *Language Teaching Research*. https://doi.org/10.1177/13621688221130856

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

Medvedev, O., Sheppard, C. L., Monetta, L., & Taler, V. (2019). The BNT-38: Applying Rasch analysis to adapt the Boston Naming Test for use with English and French monolinguals and bilinguals. *Journal of Speech, Language, and Hearing Research*, *62*(10), 3690–3701. https://doi.org/10.1044/2018_JSLHR-L-18-0084

O'Brien, M. L. (1989). Psychometric issues relevant to selecting items and assembling parallel forms of language proficiency instruments. *Educational and Psychological Measurement*, *49*(2), 321–334. https://doi.org/10.1177/0013164489492007

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Shirazi, M. A. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open*, *9*(1), 1–13. https://doi.org/10.1177/2158244018822377

Sundqvist, P., Sandlund, E., Skar, G. B., & Tengberg, M. (2020). Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology*, *8*(1), 1–23. https://doi.org/10.46364/njmlm.v8i1.605

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education*, *34*(13), 1967–1987. https://doi.org/10.1080/09500693.2012.705046

Wang, P., Coetzee, K. L., Strachan, A., Monteiro, S., & Cheng, L. (2020). Examining rater performance on the CELBAN speaking: A many-facets Rasch measurement analysis. *Canadian Journal of Applied Linguistics*, *23*(2), 72–95. https://doi.org/10.37213/cjal.2020.30436

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.

Wang, P., Coetzee, K. L., Strachan, A., Monteiro, S., & Cheng, L. (2020). Examining rater performance on the CELBAN speaking: A many-facets Rasch measurement analysis. Canadian Journal of Applied Linguistics, 23(2), 72–95. https://doi.org/10.37213/cjal.2020.30436

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. MESA Press.