Scale Validation in Applied Linguistics: Methodological Trends and Challenges

Citation:

Cong-Lem, N. (2025). Scale validation in applied linguistics: Methodological trends and challenges. *Research Methods in Applied Linguistics*, *4*(2), 1–8. https://doi.org/10.1016/j.rmal.2025.100217

Abstract

Scale validation plays a critical role in ensuring the reliability and applicability of measurement instruments in applied linguistics research. This review examines scale validation studies published in *Research Methods in Applied Linguistics*, with the aim of identifying methodological trends and areas in need of refinement. Fourteen studies were analyzed with attention to their validation frameworks, statistical techniques, and approaches to reliability and external validation. Findings reveal a growing use of Exploratory Structural Equation Modeling (ESEM) and bifactor modeling to support construct validation in multidimensional scales, while Confirmatory Factor Analysis (CFA) remains prevalent for theory-driven, unidimensional constructs. However, external validation—particularly predictive validity and independent sample cross-validation—remains limited, reducing the generalizability of many validated instruments in domains such as language assessment and second language acquisition. Although reliability assessment is evolving through the use of Rasch modeling and Generalizability Theory, Cronbach's alpha continues to dominate, despite its known limitations in complex constructs. Given the practical constraints faced by researchers, the review advocates for a flexible, goal-aligned approach to validation that emphasizes foundational steps such as construct conceptualization, pre-validation, and structural modeling. Enhancing predictive validity, incorporating independent sample cross-validation, and improving methodological transparency can further strengthen the rigor and practical relevance of scale development. While the scope is limited to a single journal, this review offers a roadmap for improving validation practices in applied linguistics and contributes to more robust research in language assessment, teacher education, and second language acquisition.

Keywords: *scale validation, construct validity, reliability, research methods, structural equation modelling, applied linguistics*

Introduction

Scale validation is an essential process in applied linguistics research, ensuring that measurement instruments accurately assess constructs such as language anxiety, motivation, self-efficacy, and interactional competence. In applied linguistics, a scale refers to a psychometric instrument used to measure latent psychological or affective constructs such as motivation, enjoyment, or self-efficacy, typically through self-report survey items. This differs from a test, which directly assesses learners' observable performance or knowledge in a specific language domain (Bachman & Palmer, 1996).

As linguistic and psychological constructs become more complex, researchers have shifted from traditional reliability measures toward more sophisticated validation methodologies. This shift has been documented in a growing body of work exploring both conceptual and statistical advancements in scale development (e.g., Dörnyei & Taguchi, 2010; Fulcher, 2019; McNeish, 2018). Recent studies have explored a range of advanced validation approaches to improve measurement precision and structural accuracy in applied linguistics.

These include Confirmatory Factor Analysis (CFA), which tests hypothesized factor structures; Exploratory Factor Analysis (EFA), which identifies underlying dimensions without predefined models; and Exploratory Structural Equation Modeling (ESEM), a flexible technique that integrates the strengths of both EFA and CFA. Other methods such as Rasch modeling, Generalizability Theory (G-theory), and Item Response Theory (IRT) have also gained prominence for assessing item-level functioning and reliability.

As these methodologies evolve, it becomes necessary to take stock of how validation is being approached in applied linguistics and to identify emerging trends that may shape future research. Despite this methodological expansion, there remains no systematic review focused specifically on how scale validation is being conducted and reported in a targeted journal context. Inconsistencies in the use of external validation techniques (Clark & Watson, 2019; Fulcher, 2013), uneven adoption of modern psychometric tools (McNeish, 2018; Dunn et al., 2014; Alamer, 2022), and limited attention to generalizability across samples (Isbell & Son, 2023; Ji et al., 2022) point to a lack of consolidated guidance in the field.

To address this need, this review synthesizes recent studies on scale validation published in Research Methods in Applied Linguistics, with a focus on the dominant themes and methodological choices guiding this work. RMAL, as a dedicated methodological journal, has quickly become a central venue for scale validation research, making it an ideal source for examining both progress and persistent challenges. Examining how validation has been conceptualized within this journal provides valuable insights into the statistical techniques used to establish reliability and validity. Additionally, it helps to identify recurring issues and areas that require further development in validation practices. By systematically analyzing recent studies, this review contributes to a clearer understanding of the methodological landscape of scale validation in applied linguistics.

This review has two interrelated objectives: first, to examine the dominant themes that characterize validation studies in RMAL; and second, to analyze the methodological choices that underpin these studies. Understanding these themes is crucial for identifying broader trends in the field, such as shifts in validation frameworks, increased emphasis on multidimensional modeling, or greater scrutiny of context-specific validation evidence. At the same time, exploring the validation methods employed—ranging from EFA and CFA to ESEM and Rasch modeling—provides insight into the practical techniques used to implement these frameworks. These complementary goals allow for a comprehensive picture of current validation practices and the extent to which they meet evolving psychometric standards in applied linguistics.

By consolidating findings from recent validation studies, this review serves as a valuable resource for researchers seeking to develop or refine measurement instruments in applied linguistics. Identifying patterns in how validation is conducted not only informs best practices but also helps establish methodological consistency within the field. Moreover, by pinpointing gaps in current validation strategies, this synthesis contributes to ongoing discussions about improving the rigor and applicability of measurement tools (Norris & Ortega, 2003; Clark & Watson, 2019; Sudina & Plonsky, 2021). Ultimately, a clearer understanding of scale validation practices enables applied linguists to make more informed methodological choices, ensuring that research instruments provide reliable and meaningful insights into language learning and assessment (Chen, Li, & Hui, 2025; Isbell & Son, 2023).

Methodology

Search Strategy and Selection Criteria

This review follows a systematic approach to examining scale validation studies published in Research Methods in Applied Linguistics (RMAL). The primary objective is to identify emerging themes and methodological trends in validation practices within applied linguistics research.

A structured search was conducted using Elsevier's online database, applying the search string (scale OR instrument OR questionnaire OR survey) AND (validation OR validating OR dimension OR factor) within the *Research Methods in Applied Linguistics* journal. The search was limited to empirical articles published between January 2021 and January 2025 to ensure a focused and contemporary synthesis. This search yielded 139 results, ensuring comprehensive coverage of studies related to validation. To maintain a focused scope, only peer-reviewed empirical studies that explicitly conducted scale validation were considered; conceptual papers, review articles, and studies where validation was not a primary focus were excluded.

Inclusion and Exclusion Criteria

Abstracts and full texts were screened to ensure that each study met specific inclusion and exclusion criteria. Studies were included if they developed or validated a scale, instrument, questionnaire, or survey and applied statistical validation techniques. Accepted methods included both traditional approaches such as EFA and CFA, as well as advanced techniques like ESEM, Rasch Modeling, G-theory, and IRT. To be included, studies were also required to provide explicit discussions of reliability, dimensionality, and validity.

Conversely, studies were excluded if they only referenced validation without formally applying validation methodologies, focused on theoretical discussion without empirical validation, or lacked sufficient methodological detail. After applying these criteria, 14 studies were selected for in-depth analysis.

Data Extraction and Analysis

Key details were extracted from each selected study, including the constructs measured, theoretical frameworks guiding validation, and statistical techniques employed. The analysis focused on identifying the validation approaches used, such as EFA, CFA, ESEM and bifactor modeling for multidimensional constructs, Rasch modeling for measurement precision, G-theory for reliability assessment, and external validation techniques such as regression analysis.

The analysis did not rely on a formal coding framework. Instead, validation methods were systematically compared across studies to identify patterns in methodological choices. This comparative approach allowed for the identification of common trends, innovations, and potential gaps in scale validation practices.

To support transparency and facilitate cross-study comparison, a summary table is included in Appendix A, providing an overview of each included study's validation approach. The table summarizes key features such as the construct measured, theoretical foundation, statistical validation techniques (e.g., EFA, CFA, ESEM, Rasch), reliability indicators, and use of external validation.

Rationale for Methodological Approach

By focusing exclusively on studies from RMAL, this review ensures consistency in research quality and provides a controlled assessment of validation practices within a leading applied linguistics journal. This targeted sampling approach allows for a focused evaluation of how validation is conceptualized and implemented. In doing so, it contributes to ongoing discussions on best practices in scale development. The synthesis of findings highlights prevailing methodological approaches and identifies areas in need of further refinement in validation techniques.

Findings

This section presents the findings of the review, organized around three major themes identified across the included studies: (1) approaches to assessing multidimensionality and construct structure, (2) the use and limitations of external validation methods, and (3) techniques employed to evaluate reliability and response patterns. Each theme is discussed

in detail to illuminate key methodological trends and highlight areas for improvement in current validation practices.

Multidimensionality and Structural Validation Approaches

A key theme emerging from the selected studies is the increasing emphasis on assessing the multidimensionality of constructs in applied linguistics. Traditional methods such as EFA and CFA remain widely used, but there is a clear shift toward more flexible and comprehensive modeling techniques. ESEM and bifactor modeling are increasingly favored for their ability to capture the complexity of constructs with multiple related dimensions.

This trend is evident in four of the fourteen studies (29%) (Alamer, 2022; Kruk et al., 2023; Nakanishi & Takeuchi, 2024; Teng & Teng, 2024). Alamer (2022) applied bifactor ESEM to validate the Basic Psychological Needs in Second Language (BPN-L2) Scale, demonstrating that the model provided a more precise representation of both general and specific factors than standard CFA. Similarly, Kruk et al. (2023) revisited the Boredom in Practical English Classes–Revised (BPELC-R) Scale, comparing multiple validation approaches. Their findings showed that CFA alone failed to provide acceptable fit indices and discriminant validity, whereas ESEM and bifactor ESEM offered a more nuanced assessment of the scale's structure. This methodological shift is particularly relevant for studies assessing affective and motivational constructs. For instance, Nakanishi and Takeuchi (2024) validated the Foreign Language Enjoyment Scale (FLES) for Young Learners, finding that a three-factor bifactor ESEM model better accounted for the complexity of enjoyment in language learning. Teng and Teng (2024) confirmed the superiority of bifactor ESEM in modeling self-efficacy beliefs in peer feedback, showing its predictive utility across multiple writing subdomains.

Despite recent methodological advancements, some studies continue to rely on CFA without incorporating more flexible analytic approaches such as ESEM or bifactor modeling. For example, Chen et al. (2025) and Ji et al. (2022) used CFA to examine multidimensional constructs, but their findings revealed challenges in capturing overlapping dimensions. These cases highlight the value of more adaptable models when dealing with complex structures. In contrast, studies such as Khademi (2023), Leeming and Harris (2024), and Yamashita (2024) employed alternative methods—such as multidimensional scaling or Rasch modeling—without applying factor analytic techniques, suggesting a broader need for comparative model testing in future research.

External Validation and Predictive Utility

While internal validation methods such as factor analysis dominate scale validation studies, fewer studies explicitly assess external validity. Establishing whether a scale predicts relevant outcomes is crucial for demonstrating its real-world applicability. Studies that go beyond internal consistency and factor structure by linking validated scales to external measures provide stronger evidence for their utility in applied linguistics research.

In this review, four out of 14 studies (29%) incorporated external validation. Chen et al. (2025) examined how emotions measured by the AEQ-SF predicted English language learning achievement. Using regression analyses, they found that negative emotions such as boredom and hopelessness had stronger predictive effects on test scores than positive emotions, highlighting the differential impact of emotional constructs on academic performance. Similarly, Phipps (2023) validated two L2 self-efficacy instruments using Rasch analysis and examined their predictive power in relation to learners' oral proficiency. The study reported that self-efficacy in speaking significantly correlated with students' communicative performance, supporting the scale's external validity.

Another approach was employed by Isbell and Son (2023), who used Explanatory Item Response Models to explore how item characteristics influenced difficulty in an elicited imitation test. Their findings showed that linguistic complexity and item length predicted learner performance, providing further validation for their measurement instrument. By

contrast, six out of 14 studies (43%) relied solely on internal validation metrics. Leeming and Harris (2024), for example, conducted regression analysis with Rasch-derived scores to examine the predictive power of motivational constructs for TOEIC scores. However, the explained variance was small, and external validity was not a central focus of their study. This suggests that while predictive modeling was applied, its potential was not fully explored.

This gap in external validation highlights the need for future research to place greater emphasis on demonstrating how well validated scales predict real-world language learning outcomes.

Reliability Assessment and Response Patterns

Reliability assessment is a cornerstone of scale validation, but approaches to measuring reliability vary considerably across studies. While many studies continue to report Cronbach's alpha—despite widespread recognition of its limitations in estimating measurement precision—others increasingly complement or replace it with more robust alternatives. These include composite reliability (CR), McDonald's omega, Generalizability Theory (G-theory), and Rasch modeling, which together provide a more comprehensive evaluation of scale reliability.

For example, Lee and Ye (2023) applied G-theory to assess the reliability of the Foreign Language Classroom Anxiety Scale (FLCAS) and found that traditional coefficients such as Cronbach's alpha tended to overestimate reliability when test–retest consistency was considered. Their findings suggest that G-theory offers a more robust reliability estimate by accounting for multiple sources of variance, including occasion and rater effects. Similarly, Syquia and Leeming (2024) employed G-theory to examine inter-rater reliability in performance-based assessments, highlighting the method's utility for decomposing variance in rater-mediated contexts.

Rasch modeling was employed in several studies—specifically Isbell and Son (2023), Leeming and Harris (2024), Phipps (2023), and Yamashita (2024)—to examine item fit, response category functioning, and person-item reliability. For instance, Yamashita (2024) applied both the Rasch Partial Credit Model and the Rating Scale Model to identify problematic items with poor category functioning, while Phipps (2023) used Rasch analysis to refine response categories in self-efficacy instruments. These applications demonstrate how Rasch modeling can enhance measurement precision at the item level.

Most studies reported reliability using multiple indicators. For example, Nakanishi and Takeuchi (2024) relied on McDonald's omega and item-total correlations, while Teng and Teng (2024) reported alpha, CR, and omega. Only two studies—Khademi (2023) and Ji et al. (2022)—reported Cronbach's alpha as the sole measure of internal consistency, without accompanying modern reliability metrics. Although Yabukoshi (2024) also reported alpha, it was appropriately supplemented with CR and AVE. This limited reliance on alpha alone suggests a positive shift toward more nuanced reliability reporting, though continued attention to measurement precision remains necessary.

Discussion

Summary of Findings

The findings highlight three key methodological trends in scale validation studies in applied linguistics. First, there is a growing adoption of advanced factor analytic techniques, with a noticeable shift from standard CFA to ESEM and bifactor modeling. This transition allows for a more precise understanding of multidimensional constructs, particularly in areas such as motivation, emotions, and self-efficacy. However, CFA remains widely used in studies where theoretical models are well-defined, suggesting that future research may benefit from continued exploration of the contexts in which each method is most appropriate.

Second, limited emphasis on external validation remains a notable gap. While some studies successfully link validated scales to real-world language learning outcomes, many focus

solely on internal structural validation. This gap indicates a need for future research to incorporate predictive validity tests, regression analyses, and cross-validation to establish the broader applicability of validated instruments.

The third key trend involves the varied approaches to reliability assessment. While traditional reliability metrics such as Cronbach's alpha remain dominant, some studies have adopted more sophisticated approaches, including G-theory and Rasch modeling, to improve measurement precision. Expanding the use of these methods would enhance the robustness of scale validation studies in applied linguistics.

Together, these findings contribute to a broader understanding of validation practices in applied linguistics and underscore the importance of methodological rigor in scale development. Prioritizing external validation, embracing flexible factor analytic techniques, and using advanced reliability estimation methods can help ensure that validated scales provide meaningful insights into language learning and assessment.

Critique of Key Trends and Contributions of the Review

The evolution of scale validation in applied linguistics reflects significant methodological advancements. In particular, the increasing use of sophisticated techniques such as ESEM, bifactor modeling, and Rasch analysis has enhanced construct validation by enabling more nuanced measurement of multidimensional constructs. However, despite these advances, many studies continue to emphasize internal validity while paying limited attention to external validation.

It is important to recognize that CFA remains a valuable and appropriate tool, especially when researchers work with unidimensional, well-established constructs and have strong theoretical expectations about factor structure. In such cases, CFA provides a robust and parsimonious method for hypothesis testing and model fit evaluation. Yet, in more complex or emergent construct domains, alternative approaches like ESEM and bifactor modeling may be better suited to account for cross-loadings and hierarchical structures.

While many studies rigorously assess factor structures and model fit indices, fewer evaluate whether their instruments can predict real-world language learning outcomes (Chen et al., 2025; Phipps, 2023). As highlighted by Clark and Watson (2019) and Norris and Ortega (2003), construct validation requires evidence of both internal structure and external correspondence. Without predictive validity and cross-validation across independent samples, the practical utility of many validated scales may be limited. This review contributes to the field by systematically identifying this gap and reinforcing the need for an integrated validation framework—one that supports both statistical rigor and practical relevance in applied linguistics research.

The limited application of external validation in applied linguistics may be attributed to several factors. First, many studies rely on self-report data, which do not easily link to independent outcome measures. Second, researchers often lack access to performance data or longitudinal outcomes that would enable predictive validity testing. Finally, there is a disciplinary emphasis on psychometric rigor within internal validation frameworks, which may unintentionally deprioritize validation strategies that assess real-world applicability.

Another issue emerging from this analysis is the inconsistent approach to reliability assessment. While more studies are incorporating G-theory and Rasch modeling to refine reliability estimates (Lee & Ye, 2023; Yamashita, 2022), Cronbach's alpha continues to dominate as the primary metric, despite widespread recognition of its limitations (McNeish, 2018). Critics argue that alpha overestimates reliability in multidimensional scales and fails to account for unequal factor loadings or measurement error (McNeish, 2018; Dunn et al., 2014). Many studies report alpha coefficients as sufficient evidence of reliability without addressing measurement error sources or alternative consistency estimates. This reliance on outdated metrics suggests a need for broader methodological awareness and more systematic reporting of reliability across validation studies. This review highlights the value of

shifting towards multi-faceted reliability assessments, ensuring that scale validation practices align with contemporary psychometric standards.

Beyond methodological choices, the absence of standardized validation procedures across studies remains a challenge. Some studies employ rigorous multi-phase validation processes, including cross-validation and item refinement based on model fit indices (Alamer, 2022; Nakanishi & Takeuchi, 2024), while others rely solely on factor analysis without further verification. This inconsistency makes it difficult to compare validation studies and evaluate the robustness of different instruments. This review underscores the importance of developing standardized validation guidelines for applied linguistics, ensuring that researchers follow a systematic approach that enhances comparability and replicability across studies.

Steps for Future Scale Validation Studies

In light of the methodological gaps identified—particularly the inconsistent use of external validation and advanced psychometric methods—a structured, multi-stage validation framework drawing on the reviewed best practices can support more rigorous and transparent scale development in applied linguistics.

The validation process begins with a strong theoretical foundation. As a first step, researchers are encouraged to engage in careful construct conceptualization to ensure theoretical clarity prior to item development. Studies such as Ji et al. (2022) demonstrated how strong theoretical grounding facilitated the validation of the Peabody Picture Vocabulary Test–5 (PPVT-5) for bilingual learners. Pre-validation procedures, including expert review, cognitive interviews, and pilot testing, are recommended to refine item clarity and construct relevance before full-scale data collection (Chen et al., 2025).

After data collection, structural validation becomes a critical step in assessing the underlying factor structure of the scale. This stage is strengthened by employing a range of analytic techniques that extend beyond traditional CFA. Given the increasing use of ESEM and bifactor modeling in applied linguistics research (Alamer, 2022; Kruk et al., 2023), future studies are encouraged to systematically compare different factor models to determine the most appropriate structure. Rasch modeling can be used to refine individual item functioning, ensuring that response categories operate effectively and that measurement invariance holds across different demographic groups (Yamashita, 2022).

Assessing reliability requires equal methodological care. Moving beyond Cronbach's alpha is advisable for future studies seeking more accurate and nuanced reliability estimates. They are encouraged to integrate composite reliability, McDonald's omega, or G-Theory-based estimates (e.g., Lee & Ye, 2023). Rasch-based reliability estimates, as illustrated in L2 self-efficacy validation studies (Phipps, 2023), represent a valuable tool for accounting for response and group-level variability and merit wider application.

Validation efforts should also extend beyond internal psychometric qualities to include external validation. This often-overlooked step is essential for establishing whether an instrument meaningfully predicts relevant outcomes. While some studies, such as Chen et al. (2025), have successfully linked achievement emotions to standardized test scores, many others omit this crucial aspect of validation. Only a few, including Yabukoshi (2024), have incorporated predictive analyses—for instance, by correlating self-regulated listening strategy scores with TOEIC Listening performance—though sample sizes were limited. Integrating predictive validity measures, such as correlations between scale scores and external indicators like student performance or engagement, can enhance the practical utility of validated instruments (Fulcher, 2013). In addition, the use of independent sample cross-validation plays an important role in determining whether instruments generalize beyond the original development context.

To improve transparency and consistency across the field, standardized reporting practices are needed. As a final step, researchers and journal editors may consider adopting clearer

guidelines for documenting reliability and validity evidence. These should move beyond simplistic alpha-based reporting and require documentation of the broader validation framework used. Establishing minimum reporting standards would raise the methodological bar for applied linguistics research and ensure that validated instruments provide meaningful insights (McNeish, 2018; Dunn et al., 2014).

At the same time, it is important to recognize the practical constraints faced by many validation studies. Implementing a comprehensive framework may not always be feasible due to limited time, funding, or access to representative samples. Rather than expecting full implementation in every project, researchers can prioritize key steps that align with their specific goals and context. For early-stage scale development, this might include construct conceptualization, expert review, pilot testing, and robust structural validation (e.g., ESEM or bifactor modeling), with predictive validity or cross-sample replication addressed in follow-up research.

By adopting this flexible yet principled approach, future research can enhance the methodological rigor of scale validation in applied linguistics, ensuring that measurement instruments are both statistically robust and practically relevant. This review contributes to that effort by identifying where practices need refinement, outlining a roadmap for improved validation, and reinforcing the value of sound measurement in advancing research and pedagogy.

Limitations

This review is not without limitations. First, it focuses exclusively on articles published in a single journal—*Research Methods in Applied Linguistics*—which, while methodologically focused, may not fully represent validation practices across the broader field. Second, the synthesis relied on qualitative judgment in coding and thematic analysis, which, despite careful checking, introduces some potential for reviewer bias. Third, the relatively small number of included studies (N = 14) limits the generalizability of the findings. Future reviews could extend this work by including multiple journals, broader timeframes, and more diverse methodological traditions.

Conclusion

This review highlights the increasing sophistication of scale validation practices in applied linguistics, particularly the growing adoption of ESEM and bifactor modeling to refine the assessment of multidimensional constructs. These methodological advancements have contributed to more precise construct validation, allowing researchers to better account for complex factor structures. However, the emphasis on internal validation continues to outweigh efforts to establish external validity, which remains an area for further development. While many studies present strong psychometric evidence through factor analysis and reliability estimation, a more integrated approach incorporating predictive validity and independent sample cross-validation would further enhance the applicability of validated instruments beyond research settings.

Future research should build upon these methodological advancements by systematically incorporating predictive validity testing, cross-validation across diverse learner populations, and longitudinal studies to examine the stability and practical relevance of validated scales. Expanding the use of alternative reliability measures, such as G-theory and Rasch modeling, can help provide a more comprehensive understanding of measurement precision. At the same time, practical constraints must be acknowledged, and researchers may prioritize key validation steps aligned with their study's aims and stage—focusing on conceptual clarity, pre-validation, and structural modeling where comprehensive validation is not feasible. Greater methodological transparency and standardization in validation reporting would also contribute to comparability across studies, fostering cumulative knowledge-building in applied linguistics research.

This review also acknowledges limitations, including its focus on a single journal, a modest number of studies, and reliance on qualitative synthesis. Despite these boundaries, the review provides a foundation for improving validation practice. It also emphasizes that while advanced methods are valuable, CFA remains appropriate in contexts involving well-defined, unidimensional constructs**.** Strengthening validation practices through a more holistic and flexible approach may help ensure that measurement instruments continue to evolve in both methodological rigor and practical relevance. By balancing internal and external validation efforts, future research can further enhance the reliability and applicability of validated scales, ultimately supporting more nuanced language assessment, pedagogical innovation, and theoretical advancements in applied linguistics.

References

Alamer, A. (2022). Exploratory structural equation modeling (ESEM) and bifactor ESEM for construct validation purposes: Guidelines and applied example. *Research Methods in Applied Linguistics*, *1*(1), 100005. https://doi.org/10.1016/j.rmal.2022.100005

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Brown, J. D., & Hudson, T. (2001). *Criterion-referenced language testing* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139524803

Chen, N., Li, Y. M., & Hui, A. N. N. (2025). Exploring foreign language classroom emotions and their impact on achievement in China: Validating the Achievement Emotions Questionnaire—Short Form and resupply. *Research Methods in Applied Linguistics*, *4*(1), 100158. https://doi.org/10.1016/j.rmal.2024.100158

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Fulcher, G. (2013). *Practical language testing*. Routledge.

Isbell, D. R., & Son, Y.-A. (2023). Explanatory item response models for instrument validation: A tutorial based on an elicited imitation test. *Research Methods in Applied Linguistics*, *2*(3), 100080. https://doi.org/10.1016/j.rmal.2023.100080

Ji, X. R., Li, G., & Gunderson, L. (2022). Validation of the PPVT–5 for Chinese-English bilingual learners: Application of cross-classified mixed effects model. *Research Methods in Applied Linguistics*, *1*(2), 100013. https://doi.org/10.1016/j.rmal.2022.100013

Khademi, A. (2023). Investigating test content structure using multidimensional scaling. *Research Methods in Applied Linguistics*, *2*(2), 100047. https://doi.org/10.1016/j.rmal.2023.100047

Kruk, M., Pawlak, M., Shirvan, M. E., & Soleimanzadeh, S. (2023). Revisiting boredom in practical English language classes via exploratory structural equation modeling. *Research Methods in Applied Linguistics*, *2*(1), 100038. https://doi.org/10.1016/j.rmal.2022.100038

Lee, K., & Ye, Y. (2023). Investigating the reliability of foreign language classroom anxiety scale (FLCAS): An application of generalizability theory. *Research Methods in Applied Linguistics*, *2*(1), 100036. https://doi.org/10.1016/j.rmal.2022.100036

Leeming, P., & Harris, J. (2024). The language learning orientations scale and language learners' motivation in Japan: A partial replication study. *Research Methods in Applied Linguistics*, *3*(1), 100096. https://doi.org/10.1016/j.rmal.2024.100096

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Nakanishi, Y., & Takeuchi, O. (2024). Validating the foreign language enjoyment scale for young learners: An exploratory structural equation modeling approach. *Research Methods in Applied Linguistics*, *3*(3), 100167. https://doi.org/10.1016/j.rmal.2024.100167

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (1st ed., pp. 716–761). Wiley. https://doi.org/10.1002/9780470756492.ch21

Phipps, J. (2023). The validation of two L2 self-efficacy instruments using Rasch analysis. *Research Methods in Applied Linguistics*, *2*(3), 100084. https://doi.org/10.1016/j.rmal.2023.100084

Sudina, E., & Plonsky, L. (2021). Academic perseverance in foreign language learning: An investigation of language-specific grit and its conceptual correlates. *The Modern Language Journal, 105*(4), 829−857. https://doi.org/10.1111/modl.12738

Syquia, J., & Leeming, P. (2024). Assessing interactional competence through group discussion: A mixed methods validation. *Research Methods in Applied Linguistics*, *3*(3), 100144. https://doi.org/10.1016/j.rmal.2024.100144

Teng, M. F., & Teng, L. S. (2024). Validating the multi-dimensional structure of self-efficacy beliefs in peer feedback for L2 writing: A bifactor-exploratory structural equation modeling approach. *Research Methods in Applied Linguistics*, *3*(3), 100136. https://doi.org/10.1016/j.rmal.2024.100136

Yabukoshi, T. (2024). Validating a scale to measure self-regulated learning strategies for independent listening beyond the EFL classroom. *Research Methods in Applied Linguistics*, *3*(1), 100090. https://doi.org/10.1016/j.rmal.2023.100090

Yamashita, T. (2022). Analyzing Likert scale surveys with Rasch models. *Research Methods in Applied Linguistics*, *1*(3), 100022. https://doi.org/10.1016/j.rmal.2022.100022

Appendix A

No.	Study	Constructs Measured	Framework	Techniques	Reliability	External Validation
1	Alamer (2022)	Basic Psychological Needs (autonomy, competence, relatedness)	Self-Determination Theory	CFA, Bifactor CFA, ESEM, Bifactor ESEM	Omega (.55–.83)	Yes – Predicts autonomous motivation
2	Chen et al. (2025)	Foreign language emotions (e.g., hope, pride, anxiety)	Control-Value Theory	CFA (4 models), measurement invariance	Alpha (.76–.89), Omega (.79–.89)	Yes – Predicts CET-4 achievement
3	Isbell & Son (2023)	Item difficulty in Elicited Imitation Test	Construct representation validity	LLTM, LRSM (EIRM)	R¬≤ = .13 (LLTM), R¬≤ = .61 (LRSM)	Yes – Linguistic predictors of item difficulty
4	Ji, Li, & Gunderson (2022)	Receptive vocabulary (PPVT-5)	Flexible validation (Ji & Wu, 2021)	Cross-classified mixed effects models	Estimated .99	Yes – Explained by lexical/demographic predictors
5	Khademi (2023)	Content complexity in IELTS and TOEFL prompts	Validity framework (AERA, APA, NCME)	Multidimensional Scaling (MDS)	RSQ: TOEFL = .74, IELTS = .58	No explicit prediction; comparison-based
6	Kruk et al. (2023)	Boredom in English classes	Control-Value Theory, attention deficit theory	CFA, ESEM, Bifactor CFA/ESEM	Omega: .90–.92 (specific), .65 (global)	Yes – Predicted self-evaluation
7	Lee & Ye (2023)	Foreign Language Anxiety (FLCAS)	G-Theory-based reliability	Univariate and multivariate G-Theory	Gen. coef. .91 ,Üí .64 across occasions	No
8	Leeming & Harris (2024)	Language learning motivation (SDT)	SDT & Organismic Integration Theory	Rasch analysis, regression	Person/item separation; threshold fit	Yes – Predicted TOEIC scores

9	Nakanishi & Takeuchi (2024)	Foreign Language Enjoyment	Positive Psychology, Flow, Broaden-and-Build	EFA, ESEM, Bifactor ESEM	Alpha = .944, Omega = .958 (composite)	No – Only internal structure
10	Phipps (2023)	Speaking self-efficacy and sources	Bandura‚Äôs Self-Efficacy Theory	Rasch modeling (WINSTEPS)	Person: .96, Item: .99	No performance-based validation
11	Syquia & Leeming (2024)	Interactional competence	MFRM, G-Theory	Many-Facet Rasch Model, G-study	Person: .96; Criteria: .94	No external criterion used
12	Teng & Teng (2024)	Self-efficacy in peer feedback (5 dimensions)	Sociocognitive Theory, SRL	CFA, ESEM, Bifactor ESEM	Omega = .91–.95	Yes – Predicts L2 writing performance
13	Yabukoshi (2024)	SRL strategies for listening	Zimmerman‚Äôs SRL model	CFA, regression	Alpha = .86	Yes – Predicts TOEIC listening scores
14	Yamashita (2024)	Foreign Language Anxiety (8-item FLCAS)	Messick‚Äôs validity theory	Rasch RSM, PCM	Alpha = .89	No – Focus on item functioning