DOI: https://doi.org/10.4038/jls.v9i2.55

Journal of Language Studies (JLS), Volume 9, Issue II, Dec 2025

The Role of Pilot Studies in the Standardisation of Measurement Instruments in Quantitative Language Education Research

S.G.S. Samaraweera

Department of English Language Teaching

University of Ruhuna

Abstract

Pilot studies have long been accepted as an important initial step in the conduct of quantitative research but the systematic function of pilot studies in the standardisation of measurement instruments has been largely neglected in the field of language education. The aim of this investigation is to explore the contribution of pilot studies to the establishment of reliability, validity, and feasibility of quantitative measurement instruments utilised in language education research. The study employs a mixed-methods design in which documentary analysis was conducted on six doctoral pilot study reports and qualitative analyses were conducted on semi-structured interviews and focus groups conducted with 25

experienced researchers. The results suggest that pilot research serves an essential purpose in improving the clarity of items, minimising measurement error, improving the psychometric quality and resolving logistical difficulties before conducting main data collection. The findings also suggest that the effective use of pilot testing occurs when it is implemented as an ongoing and integrated process versus a singular methodological task. This study contributes to the methodological understanding of pilot studies in language education by providing both empirical and experiential data from a notably under-researched area within the field and offers insights for practitioners as they adhere to rigorous measurement development and standardisation through systematic pilot testing.

Keywords: Language education research, Measurement tools, Pilot studies, Quantitative research, Standardisation

1. Introduction

Quantitative research plays a central role in language education by enabling researchers to investigate language acquisition, instructional practices, learner characteristics, and educational outcomes through systematic measurement and statistical analysis. It provides objective, reliable, and generalisable evidence that informs educational theory, policy, and practice. However, the credibility of this evidence depends fundamentally on the quality of the measurement instruments used to collect data.

In language education research, numerous latent constructs, including language proficiency, vocabulary knowledge, motivation, anxiety, learning strategies, and perceptions of instructional effectiveness, are investigated. Because these constructs cannot be measured directly, they must be operationalised through carefully designed measurement instruments such as standardised tests, questionnaires, surveys, rating scales, and structured observation protocols (Brown, 2014). The quality of research findings depends on the psychometric

soundness of these instruments, as poorly designed instruments characterised by ambiguous wording, weak construct alignment, or cultural bias can produce unreliable and invalid results, leading to inaccurate conclusions (DeVellis, 2017). These challenges are particularly significant in language education due to the diversity of learners, multilingual learning environments, and varying sociocultural contexts. Consequently, developing reliable, valid, and culturally appropriate measurement instruments requires a systematic and iterative process of instrument development and standardisation. This process typically involves item generation, expert review, pilot testing with representative samples, and the establishment of reliability and validity evidence to ensure that the instrument accurately measures the intended constructs and produces consistent, comparable findings across different research contexts (Boateng et al., 2018; Crocker & Algina, 2006).

Instrument piloting plays a vital role in the research design process by assessing the feasibility, validity and

reliability of both research instruments and research methods prior to large-scale collection of research data (Thabane et al., 2010; van Teijlingen et al., 2001). Although the primary purpose of pilot studies is to be conducted as an initial research phase, they enable researchers to refine their research instruments iteratively based upon information obtained from the pilot study process (e.g., unclear question items, problems with response format, unclear instruction, potential issues related to participant fatigue, administration issues, etc.). Researchers also evaluate how pilot-study participants interpret questionnaire items using techniques such as cognitive interviews, think-aloud protocols, and exploratory factor analysis (Boateng et al., 2018; Dörnyei, 2010). In addition, pilot testing is essential to ensuring the quality of the data collected, as it informs sampling, recruitment, data collection and ethical procedures while also informing changes made to improve psychometric quality and overall methodological rigor in the research methodology (Bryman, 2016; DeVellis, 2017; Kim, 2011). As such, pilot studies minimise measurement error, strengthen the

research design and improve the credibility and replicability of quantitative findings in culturally and linguistically diverse educational settings.

Although pilot studies are widely recognised as an important component of quantitative research, their role in the systematic standardisation of measurement instruments has received limited attention in language education research. In addition, past studies have examined the isolated methodological elements of pilot testing without incorporating the results of documentary review and researcher experience into explaining how pilot studies can contribute to the overall standardization of instruments at different stages of the research process. The absence of these conceptual and empirical integrations results in a lack of comprehensive methodological guidance for researchers to use when designing quantitative studies in the language education domain. To fill this gap, the current study examines how pilot studies contribute to the standardisation of quantitative measurement instruments in language education research.

2. Literature Review

Traditionally, pilot study has been defined as small-scale preliminary investigation; conducted in advance of a main study to assess the feasibility of research procedures and obtain information concerning potential methodological challenges (Van Teijlingen & Hundley, 2001; Leon et al., 2011). In this context, pilot studies were viewed primarily as tool to assess the feasibility of research protocols such as recruitment, methods used to collect data and logistical arrangements. However, more recent methodological literature identifies pilot studies as an integral component of the broader research design, contributing to the systematic improvement and standardisation of measurement instruments used in research (Boateng et al., 2018).

Pilot studies provide both interpretive and communicative value beyond psychometrics. Through techniques such as cognitive interviewing (Willis, 2005), pilot testing reveals how respondents interpret item content while also highlighting the mismatch between the

intended versus the perceived meaning of items. Garner et al. (2012) illustrate how pilot testing provides clarification regarding instructions and formatting; additionally, frameworks based on sound research practice advocate for a combination of iterative sampling, feedback from stakeholders, and structured analysis processes (Burns & Grove, 2010; Lee et al., 2015; O'Cathain et al., 2015). Thus, the pilot study is increasingly considered not only as a feasibility assessment but also as part of an ongoing cycle of research development that provides for the refinement of instruments, improvement of research procedures and enhancement of the methodological quality of research prior to collection of large amounts of data.

Pilot studies have been shown to significantly enhance the psychometric quality of measurement instruments. As quantitative research on language education depends on instruments to collect data on latent constructs, it is critical for researchers to ascertain the measurement instruments' reliability and validity if they wish to produce credible results (DeVellis, 2017). Pilot testing

allows researchers to assess the internal consistency, test-retest reliability, content validity, and construct validity of the measurement instruments (Cicchetti, 1994; Cronbach et al., 1955; Messick, 1989; Polit & Beck, 2006). Researchers can revise their measurement instruments based on actual empirical data collected through pilot testing (Boateng et al., 2018; Drennan, 2003). Researchers are able to decrease measurement error, increase the precision of their measurement instruments, and increase the overall quality and credibility of their quantitative research through repeated cycles of evaluation and refinement of the measurement instruments.

Pilot studies play an important role in language education research, as measurement instruments measure several complicated and multidimensional constructs, such as language proficiency, motivation, anxiety, vocabulary knowledge, learning strategies and perceptions of instructional practices. All of these constructs are influenced by their linguistic, cultural and educational context; therefore, developing context appropriate

instruments is important in language education research. Pilot studies allow researchers to assess the language clarity, cultural appropriateness, construct equivalence and participant's understanding of the instrument(s) being used prior to implementing the instrument(s) on a larger scale (Brown, 2014). Additionally, these studies allow researchers to evaluate the administrative procedures for implementing a research study, including recruiting participants, collecting data, timing, logistics, and instrument usability, before the implementation of the main research study (Leon et al., 2011). This is especially critical in multicultural and multilingual educational contexts, where culturally inappropriate or poorly adapted instruments may introduce bias into the measurement process, and undermine the validity of the research. Given that the constructs studied in language education are multidimensional and culture-specific (e.g., language proficiency, learner engagement, strategy use, and cognitive processing), measurement in this field must consider both the psychometric properties of the measurement instruments and the linguistic idiosyncrasies of the participants, together with relevant

sociocultural and contextual factors. The use of pilot studies reduces the number of procedural mistakes, increases the number of participants engaged in the research study, and allows for more efficient and effective data collection during the main research study. Therefore, pilot studies have a significant impact on the creation of valid, reliable, and contextually relevant measurement instruments for use in language education research.

Pilot studies are acknowledged as important within research methodology; however, the majority of prior research has focused on assessing the feasibility of conducting pilot studies and refining measurement tools rather than how pilot studies contribute to the overall systematic measurement standardisation process that is commonly found in language education research. Furthermore, most prior literature on the methodological functions of conducting pilot testing has treated each study individually without incorporating both the empirical evidence from document analysis and the practical experiences of researchers within each study to

provide a more thorough theoretical framework about pilot testing's contribution to developing instruments. A lack of theoretical framework and integration of both empirical and theoretical evidence limits researchers' access to comprehensive methodological guidance when conducting quantitative research within the field of language education.

The current study attempts to address this void by conducting both documentary analysis (to collect a body of literature on pilot studies and their contributions to developing quantitative measurement instruments) and interviews (to collect practical experiences associated with pilot studies) to elucidate the contribution of pilot studies in developing reliable, valid, and contextually appropriate measurement instruments. Additionally, addressing this gap is critical for developing a sound theoretical and methodological framework for quantitative language education research.

3. Methodology

3.1. Research Design

The present research employed a concurrent parallel mixed-methods design located within a pragmatic research paradigm to examine the role of pilot studies in the standardisation of quantitative measurement instruments in language education research. Documentary and qualitative data were collected concurrently from experienced researchers (two independent but complementary data sets) and was integrated at the time of interpretation for maximal understanding of the phenomenon being studied. A concurrent parallel design was selected because documentary evidence and researchers' experiences represented complementary but independent sources of evidence that could be compared simultaneously. Collecting both datasets concurrently reduced the influence of temporal variation and facilitated direct triangulation during interpretation. Additionally, adopting a pragmatic view allowed for methodological decisions to be guided by their ability to address the

research objectives, rather than by a single epistemological position. The mixed-method research approach was a suitable method for this study because documentary analysis provided empirical inputs about the development and improvement of measurement instruments. On the other hand, the qualitative data gained from interviews explained the decision-making process followed by researchers, challenges faced during the process, and their experiences prior to pilot testing. By enriching the data through the multiple methods chosen, this study was able to ensure the credibility, validity, and interpretive richness of findings.

3.2. Data Sources

The study employed two types of data: documentary data and qualitative data. The documentary data consisted of six quantitative studies on language education at the doctoral level conducted between 2010 and 2020. These studies reported on the pilot testing of measuring instruments and showed how pilot testing impacted the development of the instruments. Systematic searches of

Google Scholar, and institutional repositories were performed to identify potential studies.

The use of documentary analysis was applied in the analysis of six doctoral preliminary studies with purposive sampling. The analysis of doctoral dissertations was necessary because of the fact that they contain full methodological documentation which is usually unavailable in journal articles. These dissertations contain particulars concerning the development of research instruments, pilot testing, revision and reliability analysis. The selected reports provided detailed accounts of how an instrument was developed, piloted, and assessed for reliability and validity to facilitate the standardisation of quantitative measurement instruments. Only reports from language education and applied linguistics that included a documented pilot testing phase were included; however, qualitative studies and reports without sufficient information on instrument refinement were excluded.

Through use of semi-structured interviews and focused group discussions, qualitative data was obtained from 25 researchers with proven background and expertise in quantitative methods and pilot study design and instrument development. These participants had demonstrated success in publishing in language education or related fields. The qualitative data provided insight into the researchers' decision-making processes, methodological issues, and best practices in pilot-testing measurement instruments.

3.3. Sample Selection

This study employed purposive sampling to identify appropriate sources of analysis (i.e., documents) and participants (i.e., researchers). For the documentary analysis, six doctoral dissertations were selected according to three criteria: (1) those that utilized quantitative measures as part of language education research; (2) those that contained a documented pilot study; and (3) those that provided evidence of modifications to the instrument as a result of the pilot test. Studies that lacked sufficient detail about the pilot

testing process, or that were entirely qualitative, were not included. The six dissertations provided a comprehensive accounts of the development, pilot testing, and evidence of the reliability and validity of the measurement instruments used, making them suitable data sources for examining the standardisation of measurement instruments. The final sample for the qualitative component of the study comprised 25 researchers with substantial experience in conducting quantitative language education research and pilot testing. They were purposively selected to provide informed insights into instrument development and methodological decision-making. By triangulating documentary evidence and expert opinions, the study captured both empirical and experiential perspectives on how pilot studies contribute to the development of robust quantitative measurement instruments.

3.4. Data Collection Procedures

To conduct the analysis of the documentary data, a structured framework for extracting data from these documents was designed to capture information on key

characteristics including; the type of measurement instruments used and the purpose (i.e., measuring the same thing for multiple construct measures); pilot sample characteristics; reported reliability and validity data (i.e., reported Cronbach's alpha and test-retest reliability); and justification for any revisions made after pilot testing. This framework created consistency and comparability across all the documentary data collected to ensure a transparent analysis of the collected documents.

The qualitative data were collected through 18 face-to-face semi-structured individual interviews and two face-to-face focus group discussions, conducted according to participant availability. The individual interviews lasted between 45 and 60 minutes, while the focus groups lasted between 60 and 90 minutes. The interviews were conducted using a semi-structured protocol, which focused on the experiences of the participants in designing and implementing a pilot study, what they believed pilot testing helped to provide for the refinement of a measurement instrument, challenges that were encountered during the pilot testing phase, and what they

considered best practices for standardising measurement instruments. All sessions were audio-recorded with participant permission and transcribed verbatim.

3.5. Data Analysis

Quantitative and qualitative data were analyzed separately and brought together afterwards with a convergent mixed methods design. In terms of quantitative data, the descriptive analysis consisted of examining documentation for various indicators of performance (i.e., Cronbach's alpha coefficients, test–retest reliability coefficients, factor loadings, and instrument modifications). In each case, instruments from the pre-pilot study phase and pilot test phase were compared so that improvements in assessment quality associated with the pilot testing could be established.

Qualitative data that were collected in interviews and focus groups were analyzed using thematic analysis following the stages of familiarization with the data, initial coding, theme generation, theme review, and theme definition. Thematic analysis was selected because

it enables the systematic identification of recurring methodological practices, researcher experiences, and patterns across qualitative datasets while remaining compatible with the pragmatic mixed-methods design. This process allowed for (1) the identification of recurring themes with regards to the instrument optimisation, the quality of measurement, feasibility to administer instruments, and pilot testing as continuous improvement process, and (2) the integration of both strands of data during the interpretation of the analysis to identify areas of convergence, complementarity, and divergence between the two datasets. Documentary data provided empirical support for instrument development and standardisation, while the researchers' perspectives provided a contextual understanding of the methodological decisions, challenges, and experiences associated with pilot testing.

Through the integration of these two types of data, the findings are strengthened in terms of both the depth and credibility of the findings by having both empirical and experiential evidence to support the contribution of pilot

studies to the quality of quantitative measurement instruments in applied linguistics research. Through the process of methodological triangulation, a combination of documentary analysis, individual interviews, and focus group discussions was used for the purpose of ensuring credibility of the research results. Furthermore, an audit trail was kept during the entire research work in order to keep a record of the analytical decisions and coding methods used.

4. Results and Discussion

The aim of this study was to examine the role of pilot studies in the process of standardising quantitative measuring instruments applied in language education research. The investigation relied on documentary analysis of six selected PhD dissertations (Samaraweera, 2019; Masilamani, 2019; Purvis, 2014; Sadeghi, 2013; Gautam 2017; Srisang, 2017) and qualitative data collected from 25 researchers through interviewing and conducting focus groups. According to the results obtained, pilot studies improve psychological

characteristics of instruments. Being not only a step in the process of standardising, pilot studies allow improving the quality of measures in terms of item clarity, cultural relevance of the tool, methods of conducting the measurements. In general, both quantitative and qualitative results may be seen as strong evidence that pilot studies are an important process of creating standardised measuring instruments for language education research. The combination of qualitative and quantitative findings of the research confirms that pilot studies are of vital importance for the development of standardized and context appropriate measurement instruments for language education research. Prior research has placed importance on the role of pilot testing as a method of detecting ambiguities in the instruments items; however, the current research has established that ambiguity is not just a linguistic matter but also involves context, especially in a multilingual language education setting.

Analysis of the six doctoral studies revealed that pilot testing resulted in some significant changes across them

all. Common revisions included rewording ambiguous statements, removing redundant or poorly performing items, revising response scales, and clarifying administration instructions. Four of the six studies modified or removed between 10% and 25% of their measurement items based on pilot feedback.

The primary drivers of these changes were participant misunderstanding, cognitive overload, and cultural or linguistic inappropriateness. These findings suggest that when functioning within a multilingual diverse context, attention was placed on being lexically clear and socioculturally sensitive so that the definitions of the constructs could be understood similarly throughout the different participant cohorts. The results provided further support for existing methodological guidance that defines piloting as providing a diagnostic phase to identify sources of measurement error before large-scale data collection. Notably, continuous instrument development plays a vital role in the development and standardisation of quantitative measurement instruments for language education research in terms of enhancing the

psychometric quality and contextual appropriateness of instruments. The present study extends existing methodological literature by demonstrating that item refinement should be viewed as a continuous standardisation process rather than a single stage of measurement instruments editing.

The results of this research suggest that each revision process generated improved internal consistency. Both internally and externally consistent scores were found across all 6 original studies through analyzing quantitative data using Cronbach's alpha coefficients, which in turn increased from a range of moderate scores (approximately $\alpha = 0.62{-}0.70$) to an acceptable or strong level (approximately $\alpha > 0.75{-}0.88$) after item refinement. An illustration of this is found in Samaraweera (2019), where the reliability measure of the instrument increased from an $\alpha = 0.62$ to an $\alpha = 0.85$ after the removal of low-loading items and clarification of ambiguous wording. Additionally, this finding is consistent with previous research that identifies pilot testing as a mechanism for improving internal

consistency and scale reliability (DeVellis, 2017; Boateng et al., 2018). Whereas DeVellis (2017) conceptualises reliability improvement primarily as an outcome of scale refinement, the present findings demonstrate that reliability gains emerged through an iterative combination of participant feedback, psychometric analysis, and contextual adaptation. In addition, the observed test–retest reliability coefficients (for the two studies) increased, suggesting a measurement of improved temporal stability. Furthermore, where applicable, clearer scoring rubrics contributed to greater inter-rater reliability.

These findings support the argument of pilot testing aiding in measurement variability through identifying problematic items, vague phrasing, and unstable constructs prior to the full implementation of the research. These reliability gains confirm that pilot testing serves the purpose of psychometrically stabilising measurements; therefore, reinforcing the argument that consistent and reliable measurement scores may be

achieved through the utilisation of instruments that have undergone pilot testing procedures.

During the pilot testing phase, validity was enhanced in several ways. Researchers conducted exploratory factor analyses, allowing them to locate both low-loading and cross-loading items for replacement or revision, thereby enhancing the coherence of the constructs and aligning with existing theoretical frameworks. These findings support recommendations by Dörnyei (2010) & Boateng et al. (2018), who emphasise pilot testing as a means of strengthening construct and content validity. Additionally, through the use of expert reviewers and participant feedback, content validity was also improved by verifying that the developed items of the study adequately represented the constructs they were intended to measure. Pilot testing played a particularly important role in bilingual research areas for enriching linguistic relevance and, thereby reducing construct-irrelevant variances and possible biases. Finally, enhancing discrimination indices confirmed that developed items were able to distinguish between high performers and

low performers. These refinements to the pilot study demonstrate that pilot studies increase both conceptual clarity and empirical accuracy, thereby directly contributing to the standardisation of instruments.

Pilot studies served not only to improve statistical characteristics via psychometric measurement improvements but also as “field laboratories," as cited by Rose et al. (2016), to refine and streamline administrative processes and logistics. This approach supports the conceptualisation of pilot testing as feasibility assessments, as proposed by van Teijlingen & Hundley (2001) and Thabane et al. (2010). Multiple feasibility measures were evident through the pilot studies. Ceiling and floor effects were reduced significantly through the removal of extreme response items and adjustments to the time limits to provide sufficient variability for correlational and inferential analyses. The reduction in administration time as a result of encountering fewer redundancies and streamlined instructions averaged 15%–30% less time to administer post-pilot studies, compared to pre-pilot studies. Reduced administration

time is not merely an operational improvement but also contributes to measurement quality by decreasing participant fatigue, maintaining concentration, and reducing careless responding.

Four themes emerged from the thematic analysis of interviews and focus group discussions regarding the perceived value of, and challenges associated with, pilot testing. Researchers reported using pilot testing to refine instruments before full implementation through progressive modification that allows them to identify items lacking clarity, weaknesses in procedures, and inconsistencies in the structure of their instruments. The researchers' reflections suggest that experienced investigators conceptualise pilot testing as an evidence-generating process rather than a procedural requirement. This indicates that methodological expertise influences the depth with which pilot evidence is incorporated into instrument refinement. These perceptions align with previous methodological literature that views pilot testing as a repeated evaluation of scale enhancement (Rose et al., 2016; Kim, 2011). Furthermore, pilot testing allowed

researchers to identify measurement errors early by revealing disconnects in understanding, mismatches in cultural context, and inconsistencies in scoring that would have resulted in unreliable data. Additionally, researchers indicated ongoing tensions between ensuring methodological rigor and practical issues that limited pilot testing opportunities (limited time, lack of funding, difficulty in recruiting a representative sample). Although researchers recognised the value of producing methodological soundness through pilot testing, these types limited the extent and depth of piloting testing. As a result, researchers' individual reflections demonstrate the significance of these themes; a researcher stated that pilot feedback was “critical for making several revisions prior to commencing the full study", and another researcher stated that revisions had to be made to clarify items that were misinterpreted during piloting. These qualitative findings support the quantitative findings by demonstrating that iterative pilot feedback contributes to improving both the psychometric robustness and contextual appropriateness of measurement instruments.

Pilot studies are found to be crucial in the process of standardisation of measurement in quantitative research. The findings from qualitative and quantitative data suggest that pilot testing leads to higher reliability coefficients, better construct validity, and greater feasibility rates. These findings also show that the process of standardisation is not a one-time process, nor is it strictly technical. Rather, it is a dynamic process involving constant adjustment and modification. The findings suggest an extension of the existing discussions on the issue of instrument development since they allowed to reveal the mechanisms through which pilot testing contributes to obtaining psychometric quality of the chosen instruments and their contextual adaptation to specific language education contexts.

Although participants strongly endorsed pilot testing, they consistently reported that limited funding and restricted access to representative samples constrained the iterative refinement process. These findings indicate that institutional conditions influence the extent to which methodological recommendations can be implemented in

practice. Foremost among these are limitations in time, financial resources, and institutional support, which often restrict the scope and depth of piloting activities. For postgraduate researchers in particular, the requirement for additional rounds of data collection, instrument revision, and statistical analysis may be perceived as burdensome within tight academic timelines, increasing the risk that pilot studies are conducted superficially rather than systematically. Issues of sample representativeness further complicate implementation, as pilot participants ideally should reflect the characteristics of the target population. However, restricted access to appropriate participants frequently necessitates reliance on convenience sampling, thereby limiting the generalisability and contextual sensitivity of pilot findings. In language education research, this challenge is intensified by learner heterogeneity across proficiency levels, sociolinguistic backgrounds, and institutional contexts. Statistical complexity constitutes an additional barrier, as advanced psychometric procedures such as exploratory factor analysis and reliability estimation require both conceptual understanding and technical

proficiency. Without adequate analytical expertise or mentorship, researchers may misinterpret pilot data or implement inappropriate revisions, undermining the purpose of piloting. Finally, reflexivity must be carefully managed, since preliminary patterns identified during the pilot phase may inadvertently influence researchers' expectations and subsequent analytical decisions. Transparent documentation and critical methodological reflection are therefore essential to ensure that pilot findings inform refinement without compromising objectivity.

The significance of the findings from this research has far-reaching consequences for language education research. Language learning processes experience some form of cultural grounding, thus requiring the measuring instruments to make sure that they are psychometrically sound, linguistically clear, and relevant to the learner's culture. The process of pilot testing has been found to be a systematic and evidence-based methodology of evaluating the clarity of research items for research projects of great magnitude. The system of piloting

serves to provide means of perfecting the measuring instruments and making the internal consistency of research instruments more effective through a series of repetitions. This is an important issue in sociolinguistically diverse multilingual higher education contexts, where significant differences may exist among learners within the same institution. Overall, the findings show that pilot studies must be seen as an important part of the methodology of developing valid, reliable, and contextually sensitive tools for conducting language education research.

5. Conclusion

This study demonstrates the necessity for pilot studies to enhance both the accuracy of measuring educational performance and the contextual appropriateness of educational measurement in language education research. Through conducting pilot studies, researchers are able to improve psychometric quality for the instruments developed for use in the field of language education by providing greater internal consistency for the instrument, providing greater test-retest reliability, and providing

greater construct validity, thereby enhancing the confidence in the data obtained from the research. In addition, pilot studies provide researchers with an opportunity to better adapt the instruments developed to be used in educational settings in different cultures and languages so that researchers can better target the population that they are collecting data from and effectively remove biases that could potentially impact the data collected. Additionally, pilot studies enable researchers to enhance their data collection through improved method administration, improving participant engagement in the study by establishing an appropriate timeframe for the completion of tasks with respect to data collection, therefore improving the quality of the data collected. Researchers incorporating pilot studies into their overall methodologies are in a position to produce sound, replicable, and credible research conducted among linguistically and culturally diverse populations. Pilot testing of instruments is not a procedural step but rather an essential phase of instrument design that provides empirical data and feedback on participant performance, which can assist in refining the design of

the instrument by informing systematic development. In this way, methodological rigor can be improved; measurement error can be reduced, and thus further enhance the credibility and replicability of quantitative research conducted in multilingual and culturally diverse language education settings.

While this research makes important contributions to both researcher decision-making and the understanding of research methodologies, it is subject to a number of limitations. The analysis of documentary evidence was limited to six doctoral dissertations in language education, and the qualitative component was developed from a sample of purposefully recruited experienced researchers. Therefore, while both of these data sources provided valuable methodological understanding, the nature of these data sources may limit the ability to generalise the results of this research to other educational disciplines. Future research should include other educational disciplines, a broader sample of researchers, and additional documentary sources to enhance the ability to generalise the results.

The results indicated numerous avenues for future research. Research that compares the pilot study practices of multilingual and monolingual educational contexts and of high- and low-resource institutions would inform the reader of the best practices for the respective contexts. Conducting longitudinal studies that monitor the reliability of pilot-tested instruments over time and across different data collection timeframes (i.e., the current study and subsequent studies) may provide researchers with vital information regarding the reliability of pilot-tested instruments, and the research credibility of long-term pilot-tested instruments. Finally, the development of a structure or standardised guidelines to govern the use of pilot testing in language education may enhance the transparency and standardisation of pilot testing in language education methods and may lead to consistency in methodology and data collection. Additionally, future research should explore the potential cost-effectiveness of alternative pilot testing designs to create scalable models that can accommodate both post-graduate research and research conducted in low-resourced areas.

References:

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Brown, J. D. (2014). *Testing in language programs: A comprehensive guide to English language assessment* (2nd ed.). McGraw-Hill.

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Burns, N., & Grove, S. K. (2010). *Understanding nursing research: Building an evidence-based practice* (5th ed.). Saunders Elsevier.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Drennan, J. (2003). Cognitive interviewing: verbal data in the design and pretesting of questionnaires. Journal of advanced nursing, 42(1), 57-63.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE.

Dörnyei, Z. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). Routledge.

Garner, H., Revell, A., & McKenzie, K. (2012). The development and piloting of a self-report measure of social problem solving for people with intellectual disabilities. *Journal of Applied Research in Intellectual Disabilities, 25*(5), 409–421. https://doi.org/10.1111/j.1468-3148.2012.00683.x

Gautam, S. (2017). *Predictors of Punjabi, Hindi and English reading comprehension among multilingual children in the Punjab region of India* (Doctoral dissertation, University of Canterbury, New Zealand).

Kim, Y. (2011). The pilot study in qualitative inquiry: Identifying issues and learning lessons for culturally competent research. *Qualitative Social Work, 10*(2), 190–206. https://doi.org/10.1177/1473325010362001

Lee, H., Whitehead, A. L., Jacques, R. M., & Julious, S. A. (2015). The statistical interpretation of pilot trials: Should significance thresholds be reconsidered? *BMC*

Medical Research Methodology, 15, Article 52. https://doi.org/10.1186/s12874-015-0040-6

Masilamani, M. B. (2019). *Influence of morphological, orthographic and phonological awareness in writing skills among bilingual Malay-English speakers: A study of adult (pre-university) students in Peninsular Malaysia* (Doctoral dissertation).

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

O'Cathain, A., Hoddinott, P., Lewin, S., Thomas, K. J., Young, B., Adamson, J., Jansen, Y., & Morgan, M. (2015). Maximising the impact of pilot and feasibility studies in health and social research. *BMC Medical Research Methodology, 15*, Article 1. https://doi.org/10.1186/s12874-015-0104-7

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Purvis, C. J. (2014). *Determining and supporting the reading comprehension and metalinguistic abilities of undergraduate pre-service teachers* (Doctoral dissertation).

Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2016). Evaluation of a preliminary physical function item bank supported the expected advantages of PROMIS. *Journal of Clinical Epidemiology, 69*, 87–95. https://doi.org/10.1016/j.jclinepi.2015.06.025

Samaraweera, S. G. S. (2019). *Direct or indirect relationship between morphological awareness and reading comprehension: A study of adult Sinhala (L1) and English (L2) students* (Doctoral dissertation, University of Canterbury, New Zealand).

Sadeghi, A. (2013). *Towards a universal model of reading: Investigations into Persian monolingual and English-Persian bilingual speakers* (Doctoral dissertation, University of Canterbury, New Zealand).

Srisang, P. (2017). *Influence of inferential skills on the reading comprehension ability of adult Thai (L1) and English (L2) students* (Doctoral dissertation, University of Canterbury, Christchurch, New Zealand).

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., Robson, R., Thabane, M., Giangregorio, L., & Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology, 10*, Article 1. https://doi.org/10.1186/1471-2288-10-1

van Teijlingen, E., & Hundley, V. (2001). The importance of pilot studies. *Social Research Update, 35*, 1–4.

van Teijlingen, E., Rennie, A., Hundley, V., & Graham, W. (2001). The importance of conducting and reporting pilot studies. *Journal of Advanced Nursing, 34*(3), 289–295.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE.

Acknowledgments:

The author expresses gratitude to all the doctoral researchers whose theses served as documentary data for this research project. Their academic contributions formed a substantial basis of empirical data related to the design, piloting, and validation of measurement instruments within the field of language education research. In particular, I would like to recognise the contributions of C. J. Purvis, S. Gautam, M. B. Masilamani, A. Sadeghi, and P. Srisang. The results and data collected from each of their individual doctoral studies helped clarify the processes related to the refinement, reliability, and validity of measurement instruments and pilot study practices within a broad range of linguistic and education contexts. The data collected and reported by these doctoral researchers

would also have provided a substantial element of evidence in evaluating the use of piloting to potentially standardise the measurement instruments used in quantitative language education research.

Biographical Note

Dr. S.G.S. Samaraweera is a Senior Lecturer and the Head of the Department of English Language Teaching (DELT), Faculty of Humanities and Social Sciences, University of Ruhuna, Sri Lanka. With over sixteen years of experience in higher education, he has made significant contributions to English language teaching, teacher education, research, and academic leadership. He holds a Doctor of Philosophy (PhD) from the University of Canterbury, New Zealand, and has also earned a PhD, MPhil, and Master of Arts (MA) from the University of Kelaniya, Sri Lanka, in addition to a Bachelor of Arts from the University of Sri Jayewardenepura, Sri Lanka. He further enhanced his professional expertise by completing the Teacher Training Certificate Course in Developing and Teaching Academic Writing at Iowa

State University, USA. His research interests include Teaching English to Speakers of Other Languages (TESOL), Teaching English as a Second Language (TESL), language assessment, second language acquisition, academic writing, and English language pedagogy. He has published more than fifteen peer-reviewed research articles in these areas and continues to contribute to the advancement of English language education through research, curriculum development, postgraduate supervision, and scholarly collaboration. He is committed to fostering excellence in teaching, research, and community engagement while mentoring undergraduate and postgraduate students and supporting the professional development of language educators.