Language Testing in Asia

RESEARCH

Open Access

Reliability of measuring constructs in applied linguistics research: a comparative study of domestic and international graduate theses

Kioumars Razavipour* and Behnaz Raji

*Correspondence:
razavipur57@gmail.com;
k.razavipour@scu.ac.ir

Department of English Language and Literature, College of Letters and Humanities, Shahid Chamran University of Ahvaz, Ahvaz, Iran

Abstract

The credibility of conclusions arrived at in quantitative research depends, to a large extent, on the quality of data collection instruments used to quantify language and non-language constructs. Despite this, research into data collection instruments used in Applied Linguistics and particularly in the thesis genre remains limited. This study examined the reported reliability of 211 quantitative instruments used in two samples of domestic and international theses in Applied Linguistics. The following qualities in measuring instruments were used to code the data: the instrument origin, instrument reliability, reliability facets examined, reliability computation procedures utilized, and the source of reliability reported (i.e., primary or cited). It was found that information about instrument origin was provided in the majority of cases. However, for 93 instruments, no reliability index was reported and this held true for the measurement of both language and non-language constructs. Further, the most frequently examined facet of reliability was internal consistency estimated via Cronbach's alpha. In most cases, primary reliability for the actual data was reported. Finally, reliability was more frequently reported in the domestic corpus than in the international corpus. Findings are discussed in light of discursive and sociomaterial considerations and a few implications are suggested.

Keywords: Reliability, Consistency, Data elicitation instrument, Thesis

Introduction

In educational measurement literature and in language testing, confidence in measurements depends on their consistency and validity. For an instrument to be valid, it has to be consistent (though the term consistency is more precise compared to reliability, in this paper, we used them interchangeably). That said, whereas in educational measurement and in language testing, much attention has been paid to investigating reliability and validity of tests used for selection and achievement purposes, the quality of measuring instruments used for research purposes in Applied Linguistics and language teaching remains underexplored. Such studies are warranted on the grounds that they carry immediate implications for practitioners, policy makers, and researchers. For the practitioners who rely on research findings to improve their language teaching practices, it is imperative that such research is based on sound measurements of constructs. Additionally, in action research, the effectiveness of educational interventions can only be examined through sound measurements of key variables. Sound measurements are also crucial for education policy makers who rely on research findings to choose, adapt, and implement language education policies. If the research informing policies is founded on inconsistent measurements, they are likely to derail proper policy making with grave consequences for language teachers, learners, and the wider society. Finally, proper measurements are of utmost importance for the progress of research and the production of knowledge in the field of Applied Linguistics and language teaching. Threats to the consistency and validity of measurements in research would potentially derail future research that depends on incremental accumulation of research evidence and findings. Given the mutual exchange of ideas and insights between Applied Linguistics and language testing (see Bachman & Cohen, 1998 and Winke & Brunfaut, 2021), the quality of research in different areas of AL influence research directions and decisions in language testing.

Despite the noted implications that reliable assessments hold for policy and practice, whether and the extent to which Applied Linguistics researchers examine or maximize the consistency of their measuring instruments remains underexplored. More specifically, the current literature on research instrument quality in AL is mostly focused on the published research papers. Indeed, we are aware of no published work on the reliability of measuring instruments in theses or dissertations in AL. We believe that as a distinct genre which operates under different sociomaterial circumstances and is written for a different audience, the thesis genre warrants closer scrutiny in terms of measurement quality because of the consequences and implications that the quality of this genre has for the academia and the wider society. This study intends to narrow the noted gap by investigating the reliability with which variables are measured in a corpus of theses and dissertations in Applied Linguistics across several academic settings. In the remaining of this paper, we first examine research quality in quantitative research in Applied Linguistics. We then zero in on issues of instrument validity and reliability within current theories of validity, particularly those of Messick and Kane.

Research quality and measurement

The fact that a good deal of Applied Linguistics research depends on the production and collection of quantitative data makes the quality of measuring instruments of crucial importance (Loewen & Gass, 2009). Unreliable data generates misleading statistical analyses, which, in turn, weakens or defeats the entire argument of quantitative and mixed methods studies. Subsequently, the quality of measuring instruments affects the internal validity of research studies (Plonsky & Derrick, 2016), which in turn compromises the credibility of research findings.

In the social sciences and Applied Linguistics, concern with reliability and validity of measuring instruments is a perennial problem that can "neither be avoided nor resolved" (Lather, 1993, p. 674) because unlike metric systems in physics, which are almost of universal value and credibility, measuring instruments in AL do not satisfy the principle of measurement invariance (Markus & Borsboom, 2013). That is, the properties of measuring instruments are dependent upon the properties of the object of measurement (i.e., research participants, context of use, etc.). Hence, every time, a test or a questionnaire is used in a research study, its reliability and validity should be examined.

Given the centrality of measurement invariance, Douglas (2014) uses the "rubber ruler" metaphor to refer to this property of measuring instruments in AL research. As a rubber ruler may stretch or shrink depending on temperature, the interval between units of measurement fluctuate with changes in temperature. Therefore, the quality of measuring instruments (MIs) in AL research is often subject to contextual fluctuations. For this reason, examining and maximizing the reliability of measuring instruments is crucial. The following quote from Kerlinger (1986 cited in Thompson, 1988) captures the significant of instrument reliability in quantitative research.

> *Since unreliable measurement is measurement overloaded with error, the determination of relations becomes a difficult and tenuous business. Is an obtained coefficient of determination between two variables low because one or both measures are unreliable? Is an analysis of variance F ratio not significant because the hypothesized relation does not exist or because the measure of the dependent variable is unreliable? ...High reliability is no guarantee of good scientific results but there can be no good scientific results without reliability. (p. 415)*

The above quote goes back to almost half a century ago, yet problems with MIs continue to persist in Applied Linguistics and SLA (Purpura et al., 2015).

In language teaching research, concern with how researchers handle quantitative data has recently increased. As such, several studies have addressed the quality of quantitative analyses (Khany & Tazik, 2019; Lindstromberg, 2016; Plonsky et al., 2015), researchers' statistical literacy (Gonulal, 2019; Gonulal et al., 2017), and quality of instrument reporting (Derrick, 2016; Douglas, 2001; Plonsky & Derrick, 2016). Douglas (2001) states that researchers in SLA often do not examine indexes of performance consistency for the MIs they use.

Recently, inquiry into the quality of research studies has spurred interest in the evaluation of MIs, in particular their reliability and performance consistency (Derrick, 2016; Plonsky & Derrick, 2016) in published research articles. A common theme in both of the noted studies is that the current practices in measuring instruments' reliability reporting are less than satisfactory. That is, inadequate attention is often given to the reliability of MIs in Applied Linguistics research. The current slim literature on research instrument quality is largely about the research article (RA) genre in. As such, we are aware of no published research on how the reliability of quantitative instruments is handled and reported in the thesis genre in Applied Linguistics research and almost exclusively the academic north of the globe (Ryen & Gobo, 2011). Given the culture and context-bound nature of research methogology and hence assessment methods (Chen, 2016; Ryen & Gobo, 2011; Stone & Zumbo, 2016), studying MIs in other contexts is warranted. In addition, theses are not subject to the same space limitations that the research paper is; thus, one would expect detailed accounts of data elicitation instruments in a thesis. For the noted reasons, this study examines the quailty of data elicitation instruments in a sample of theses in Applied Linguistics. We hope that findings would encourage graduate students and early career researchers to exercise more care and seek more rigor in their choice of MIs and the inferences they make of them, which would enhance the credibilty of research findings. In the remaining of this paper, we will first briefly discuss validity in Applied Linguistics and language testing. We do so to situate issues of reliability and consistency in the broader context of validity, which is the ultimate criterion of data and inference quality. We will then present our own study along with a discussion of findings and implications it might carry for research in Applied Linguistics.

Quality of measurements: validity and reliability

In psychometrics and educational measurement as well as in Applied Linguistics research, quality of measuring instruments is often captured by the term validity. In more traditional yet still quite common definitions, validity refers to the extent to which a measuring instrument measures what it is purported to measure and reliability is about how consistently it does so (Kruglanski, 2013). From this perspective, reliability is considered a necessary but insufficient precondition for validity, that is, an instrument can be reliable without being valid (Grabowski & Oh, 2018), which implies that an instrument may demonstrate consistency in the kind of data it yields without essentially tapping what it is purported to tap. In recent conceptualizations of validity, however, reliability is integrated within the domain of validity (Kane, 2006; Newton & Shaw, 2014; Purpura et al., 2015; Weir, 2005). Largely thanks to Messick's legacy, validity is defined as an overall evaluative judgment of the degree to which empirical evidence and theoretical rationale justifies the inferences an actions that are made based on test scores (Messick, 1989). Viewed from this holistic approach to validity, reliability is considered one source of validity evidence that should be used to support the inferences that are to be made of test scores. Whereas this conceptualization of validity as argument is increasingly being embraced in educational measurement and language testing, it has yet to permeate the broad literature on Applied Linguistics research in general and TEFL in specific (Purpura et al., 2015). In fact, some scholars believe that lack of knowledge about how to effectively measure L2 proficiency is the main reason for the failure of the field of SLA to make real progress in explaining development and growth in an L2 (Ellis, 2005, cited in Chapelle, 2021).

While we are mindful of the importance of validity, in this paper, we focus exclusively on reliability for two reasons. First, we believe that despite the theoretical unification of aspects of validity evidence (Bachman & Palmer, 2010; Chapelle, 2021; Kane, 2013), reliability still serves as a good heuristic to examine measurement quality. This is evident even in Kane's argument-based validity. In going from data to claims, the first argument that must be supported in argument-based validation is evaluation, which refers to how verbal or non-verbal data elicited via a quantitative measure is converted to a quantity and unless this argument is adequately supported, the rest of the validity chain cannot be sustained. Secondly, despite the noted theoretical shift, scholars continue to make the distinction between validity and reliability, perhaps because for the practitioners both Messick's unified approach and Kane's argument based validation are difficult to translate into the practice of evaluating their measuring instruments. For the noted reasons, we thought that imposing a theoretical framework of validity that is incompatible with current practices may not be helpful.

Reliability of data collected via quantitative data collection instruments

Concern with the quality of measurements in Applied Linguistics research is not new. More than two decades ago, Bachman and Cohen edited a book volume on how insights from SLA and Language Testing can assist in improving the measurement practices in the two fields. More recently, several studies have investigated reliability and consistency of quantitative instruments across disciplines (Plonsky, 2013; Plonsky & Derrick, 2016; Vacha-Haase et al., 1999). Al-Hoorie and Vitta (2019) investigated the psychometric issues of validity and reliability, inferential testing, and assumption checking in 150 papers sampled from 30 Applied Linguistics journal. Concerning reliability, they found that "almost one in every four articles would have a reliability issue" (p. 8).

Taken together, the common theme in most studies is that the current treatment of quantitative measures and instruments is far from ideal (Larson-Hall & Plonsky, 2015). That said, the findings of past studies are mixed, ranging from six percent of studies reporting reliability to 64% (Plonsky & Derrick, 2016). This loose treatment of quantitative data collection tools seems to be common in other social science disciplines such as psychology (Meier & Davis, 1990; Vacha-Haase et al., 1999).

Compared to research articles, much less work has been done on how the quality of MIs is addressed in other research genres such as theses and dissertations. Evaluating the research methodology of dissertations and published papers, Thompson (1988) identified seven methodological errors, one of which was the use of instruments with inadequate psychometric integrity. Likewise, Wilder and Sudweeks (2003) examined 106 dissertations that had used Behavioral Assessment System for Children and found that only nine studies did report reliability for the subpopulation they had studied and the majority of the studies only cited reliability from the test manual. Such practices in treating reliability likely arise from the misconception that reliability or consistency is an attribute of a measurement tool. However, given that reliability, in its basic definition, is the proportion of observed score variance in the data to the true variance, it follows that observed variance depends on the data collection occasion, context, and participants; change the context of use, and both observed variance and true variance change. That said, perhaps because of discursive habits, reliability is often invoked as an instrument property not the property of the data that is gathered via the instrument.

In sum, the above brief review points to a gap in research into the reliability of MIs in Applied Linguistics research. The current study intends to narrow this gap in the literature in the hope that it will raise further awareness of the detriments of poor research instruments. Our review of the literature showed that writers of RAs sometimes fail to provide full details regarding their MIs (Derrick, 2016), a practice which has repercussions for future research. Given the differences between the RA and thesis genre noted above, it is important to see how the quality of measuring instruments quality is addressed in the theses. The literature also suggests that reliability is underreported in RAs. In addition to addressing this in the thesis genre, in this study, we also delve further into the facets of reliability that are given attention. Given that in the discourse around reliability in Applied Linguistics, reliability is often attributed to the instrument not to the data, we further inquire into the extent to which this discourse affects the way researchers report the reliability of their data or choose to rely on reliability evidence reported in the literature. In addition, to our knowledge, extant literature has not touched upon possible relationship between reliability reporting behavior and the nature of constructs measured, a further issue we address in this study. Finally, given the situated nature of knowledge and research, it is important to know how the quality of quantitative research instruments is treated across contexts. The above objectives are translated into the following research questions.

1. How frequently are the origins of research instruments reported?
2. How frequently is the reliability reported? And when it is reported, what reliability facets are addressed and what estimation procedures are used for computing it?
3. What is the source of reliability (i.e., primary, cited, or both) that is reported?
4. Does the reliability reporting practices differ across construct types measured (language vs. non-language constructs) and across geographical regions?

We believe that these questions are important because the insights gained can contribute to our collective assessment literacy (Harding & Kremmel, 2021), which "has the capacity to reverse the deterioration of confidence in academic standards" (Medland, 2019, p. 565), for research that relies on instruments of suspicious consistency add noise to the body of scholarship and can mislead and misinform future research.

Methods

To answer the research questions, a corpus of 100 theses and dissertations from 40 universities in 16 countries across the world was collected. Roughly half of the theses were chosen from Iran, and the other half were selected from 39 universities based mostly in American and European countries. The theses from universities in the USA had the highest frequency (15) followed by those in the Netherlands (6), Canada (5), and England (4). Given that at the time of data collection, we knew of no comprehensive repository of theses accommodating theses from all universities across the globe, a random sample of theses could not be secured. Therefore, we do not claim that the corpus of theses examined in this study are representative of the universe of theses across the globe; yet, they are diverse enough to provide us with relevant insights.

For international theses, the most popular database is the ProQuest (https://pqdtopen.proquest.com). Yet, its search mechanism does not allow the user to search the theses by country and once the theses are searched using key words, the search results yielded are mostly those written in North American universities, specially the USA. To diversity the corpus and make it more representative of theses done in other universities of the world, we searched the following website: http://www.dart-europe.eu, which gives the user the option of limiting the search to a given country. All the international theses collected were then saved as PDF files.

Our only inclusion criterion was whether a thesis had made use of quantitative measures such as language tests, surveys, questionnaires, rating scales, and the like. To make inclusion decisions, the abstract and the Methods section of each thesis were carefully examined. In order to determine whether and how reliability was treated in each thesis in the domestic corpus, the abstract, the Methods chapter, and in some cases, the Results and Findings chapter were closely examined. As for the international theses, the entire Methods chapter was checked. In case we could not find information about the reliability in the noted sections, we used the search option in Acrobat Reader using the following search terms: reliability, consistency, agreement, alpha, Cronbach, valid, and KR (i.e., KR-20 and KR-21).

Our unit of analysis was the measuring instrument and not the thesis. Two hundred and eleven MIs including 110 language tests, 82 questionnaires, 9 rating scales, 8 coding schemes, and two tests of content area (e.g. math) had been used in the corpus of theses we examined. The most frequently tested aspects of language were overall language proficiency (22), vocabulary (13), writing (12), and reading comprehension (11). Regarding the questionnaires, the most frequently measured constructs were learning strategies (8), motivation (4), and teacher beliefs (4).

The coding process was mainly informed by the research questions, which were about the origin, reliability type, reliability source, and reliability estimation methods. In addition, coding schemes used in similar studies such as Plonsky and Derrick (2016) and Derrick (2016) were reviewed. Thus, coding began with the major categories highlighted in research questions. We coded the MIs used in the first 30 theses and after a thesis was coded, if a new category was found, the coding scheme was further refined to accommodate new categories. Therefore, though we started with a set of categories a priori, the actual coding was rather emergent, cyclic, and iterative. Once we settled on the final coding scheme, the entire corpus was coded once again from scratch. To minimize the subjectivity that inhere in coding, a sample of the theses was coded by the second author. The Kappa agreement rate was 96% and in a few cases of disagreement, the differences were resolved through discussion between the authors. Table 1 shows the final coding system used.

Finally, to analyze data generated using our coding scheme, we mainly used descriptive statistics such as raw frequencies, percentages, and graphic representation of data using bar graphs. In cases where we needed to compare the domestic with international theses, we used Pearson chi-square test of independence, as a non-parametric analytic procedure (see Pallant, 2010, p. 113). The above-mentioned analytic procedures were deemed appropriate because of the nominal and discrete nature of the data we worked with in this study.

Table 1 The final coding scheme

Categories	Values
Country	Open
University	Open
Research variables	Open
Instruments type	One to five (language test, questionnaire, rating scale, coding scheme, tests of content areas)
instrument origin	Seven (adopted, adapted, designed, translated, adapted and translated, compiled and adapted, origin not given)
Reliability type used	One to nine (internal consistency, inter-rater reliability, intra-rater reliability, both inter and intra reliability, internal consistency and inter-rater reliability, test–retest, inter-coder reliability, not specified, no reliability reported)
Reliability source	One to five (primary, cited, both primary and cited, not specified, not reported)

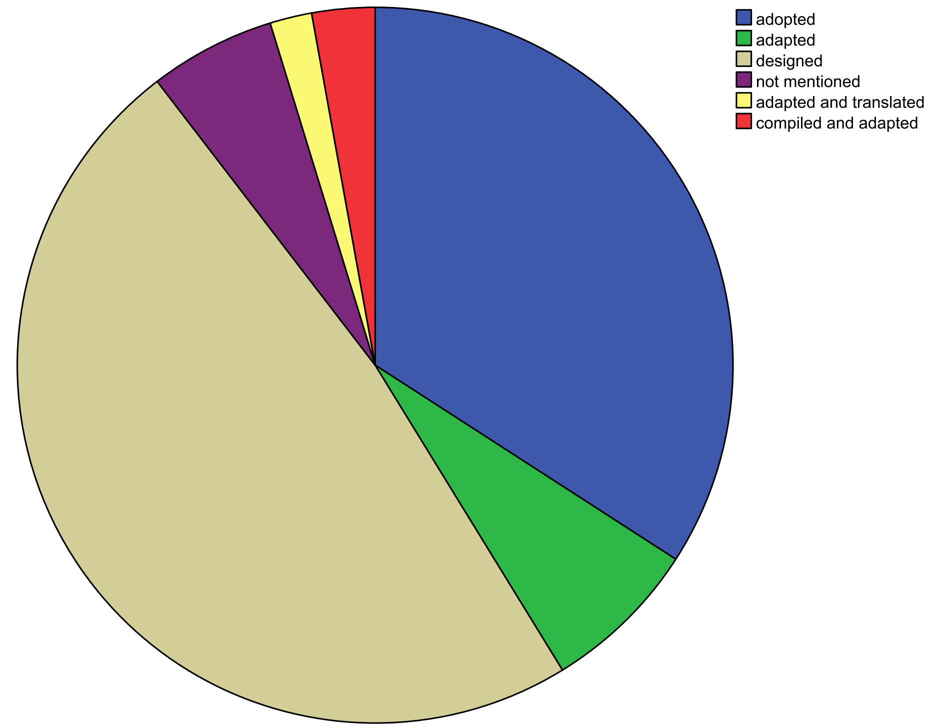

Fig. 1 Frequency of reporting instrument origin

Results

In this section, we first report the findings on the origin of MIs. Next, findings with regard to facets of reliability reported. This is then followed by reporting the results related to reliability estimation procedures used in the corpus. The source of reliability estimate along with reliability reporting across construct types comes next. Finally, findings pertaining to reliability across the domestic and international corpus of theses are reported.

Our first research question was about the origin of measuring instruments used. That is, we looked for information about whether a measurement tool used had been adapted or adopted from a previous work, designed by the researcher, adapted and translated, compiled from various measures and then adapted to the study context, or if the origin of the MI was not specified in the theses. As Fig. 1 displays, in 12 cases, the authors failed to give information regarding the origin of their MIs. Roughly half of the MIs had been designed by researchers, and a third of them had been adopted from previous studies. In the remaining cases, they had been either adapted ($n=15$), their origin was not reported ($n=12$), they were compiled and then adapted (6), or adapted and then translated ($n=4$).

The second research question of the study concerned facets of reliability (Grebowsky, 2018) that were addressed and the estimation procedures used for computing reliability. According to Fig. 2, for 93 MIs, the authors did not provide any information about the reliability of the instruments they used. In cases where reliability was reported, internal consistency was the most commonly used reliability facet ($n=75$), followed by inter-rater reliability ($n=8$), inter-rater and internal consistency ($n=7$), and the test–retest

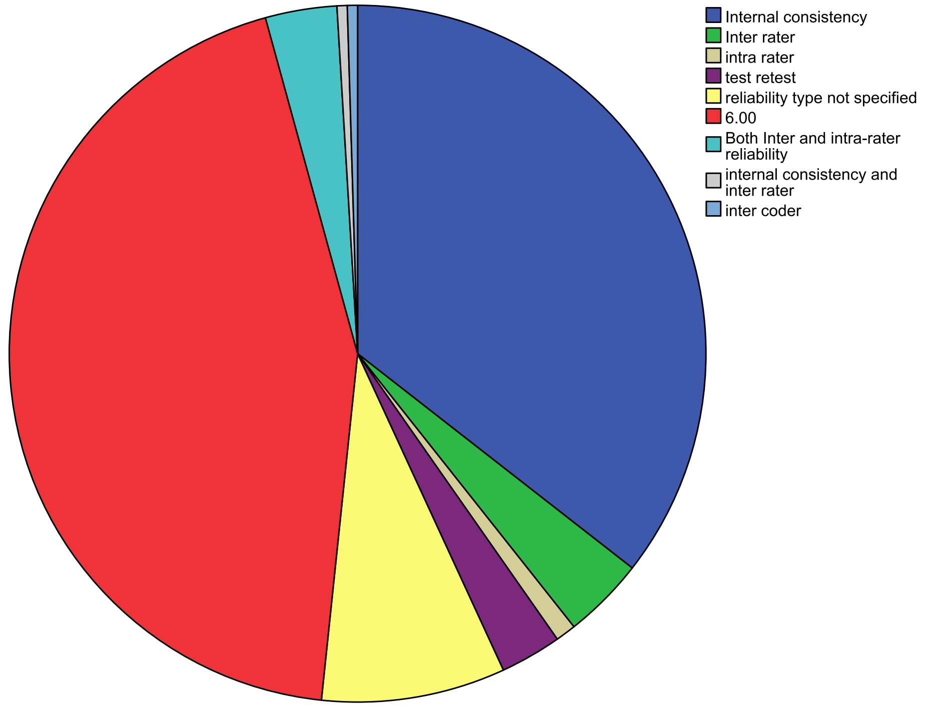

Fig. 2 Frequency of reporting each reliability facet

method ($n = 6$). On the other hand, for 18 instruments, no information was provided about the reliability facet that had been reportedly used. That is, the thesis writers did not specify the facet of reliability they had examined.

As to reliability estimation procedures used in the corpus, Fig. 3 shows that Cronbach alpha stands out with a frequency of 65, followed by Pearson correlation ($n = 7$). The two Kuder-Richardson formulas with a frequency of six and five, respectively, come next. Other less frequently used reliability estimation procedures are Spearman, Kappa, Pearson chi-square, Cohen *K*, and paired sample *t*-test. It bears noting that in 19 cases, the reliability estimation procedure was not specified. In other words, the thesis writers did not specify how they had arrived at the reliability coefficient they reported.

The third research question was about the source of the reported reliability estimate. We sought to know whether and the extent to which researchers report the reliability of their own data (i.e., primary reliability), report a reliability index from a previous study (i.e., cited reliability), or report both primary and cited reliability. The results showed that in the majority of cases ($n = 96$), primary reliability was reported. In four cases, both primary and cited reliabilities were reported and for 10 MIs, a reliability estimate from another study was reported (i.e., cited reliability).

Our fourth research question concerned whether the type of construct measured by MIs (i.e., language vs. non-language constructs) moderates the frequency with which reliability is reported (see Figs. 3 and 4).

To know if there is any association between construct type measured and the extent to which reliability is reported, Pearson chi-square test of independence (see Pallant,

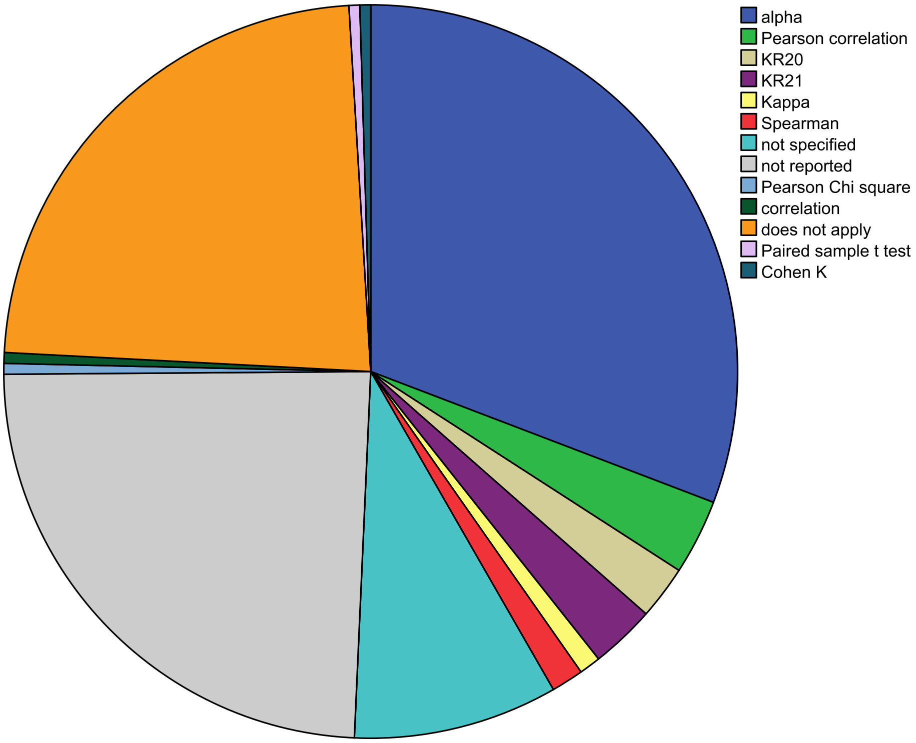

Fig. 3 Methods of estimating reliability

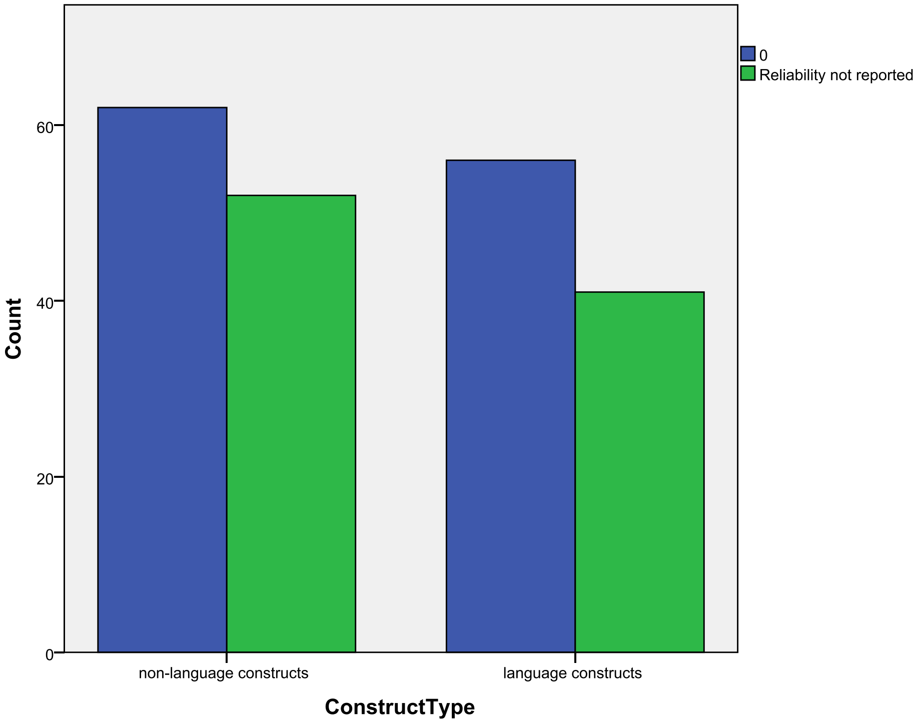

Fig. 4 Frequency of reporting reliability for language and non-language measures

Table 2 frequency of reliability reporting in domestic and international theses

	Reliability reported	Reliability not reported	Total
Domestic	73 (62%)	44 (38%)	117
International	45 (48%)	49 (52%)	94
Total	118	93	211

2010 p. 113) was run. It was found that reliability reporting did not significantly vary across construct types, $X2$ (1, $N = 211$) = 0.23, $p = 0.62$.

Finally, we sought to know whether reliability reporting practices vary across the domestic and international corpus. Table 2 gives the frequency of reporting reliability in the domestic and international theses.

As Table 2 displays, reliability seemed to be more frequently reported in the domestic corpus of theses. To see if the apparent difference in frequency is significant, another Pearson chi-square test for independence was conducted, which showed that the difference is significant, X^2 (1, $N = 211$) = 4.59, $p = 0.02$.

Conclusions and discussion

The credibility of knowledge and of research findings continues to spark debate, confusion, and controversy. Hence, across research paradigms, the question of whether and how truth is to be established has been addressed differently. In Applied Linguistics, the question of truth and credibility is often addressed using the notion of research validity, which can be threatened or compromised by different sources including inconsistencies in evidence arising from temporal, spatial, social sources. The issue of consistency is treated by examining reliability. It is assumed that when consistency is not established, claims of truth or validity cannot be made (Chapelle, 2020). In this study, we examined whether and the extent to which the reliability of measuring instruments used in measuring variables in research is addressed. More specifically, we probed into reliability reporting practices in a corpus of domestic and international theses in Applied Linguistics.

Overall, our findings in this study indicate that in a considerable number of cases, the researchers failed to examine the reliability of their research instruments and this held constant across language and non-language measuring instruments, which echo the findings of similar studies on published papers such as Plonsky and Gass (2011), Plonsky and Derrick (2016), and Purpura et al. (2015). It was also found that reliability was often treated in a ritualistic manner where, by default, researchers opt for examining the internal consistency facet of their instruments without providing a logic to choosing this facet at the elimination of other reliability facets. This finding accords with those of several studies across a number of fields (Douglas, 2001; Dunn et al., 2014; Hogan et al., 2000; Plonsky & Derrick, 2016). Finally, it was observed that in domestic corpus of theses, reliability if frequently reported than in the international corpus. In the remainder of this section, we try to explain the observed findings drawing on a socio-material frame of thought (see Canagarajah, 2018; and Coole & Frost, 2010) and sociology of knowledge (Dant, 2013).

More specifically, our finding that compared to the research articles, reliability is more frequently reported in theses might have to do with space issues as a dimension of material considerations or disciplinary conventions (Harding & Kremmel, 2021). Likewise, the dominant tendency to choose Cronbach's alpha as an index of reliability must be due to logistic and practicality concerns, as alpha is the default reliability facet in most statistical packages. Socio-material considerations are also at play when researchers often treat reliability in a post hoc manner after they have already conducted their main study. In such cases, if reliability of the data turns out to be low, researchers would prefer to skip reporting reliability (Grabowski & Oh, 2018) rather than starting over, modifying instruments, and collecting new data.

Other aspects of the findings can be accounted for by drawing on sociology of knowledge, particularly by invoking issues of genre and conventions within Applied Linguistics as discourse communities. For instance, contrary to our expectations, we found more frequent reporting of reliability in the domestic corpus. We tend to think that this might have to do with a certain discourse around reliability that is dominant in the Iranian Applied Linguistics community, where common sense meaning of reliability and its psychometric meaning are possibly conflated. As Ennis (1999) notes, reliable data does not mean good data, nor does it mean data we can rely on. These are common sense meanings of the term reliability. In contrast, in the educational measurement and psychometric discourse community, reliable data only mean data that is consistent across some test method facets. When researchers take reliable data to mean good data, they would give it more value and try to report it more frequently as a perceived index of research rigor.

Another observation that can be made sense of by invoking discursive realities has to do with the origin of MIs, which in many cases were designed by researchers. Measurement in the social sciences continues to be a source of controversy (Lather, 1993). There are some who believe that all measurements in psychometrics and education are flawed because they conflate statistical analysis with measurement, for the very objects of measurement fail to satisfy the ontological conditions of quantification (see Michell, 1999, 2008). Lather even go so far as to say that validity as a mechanism "to discipline the disciplines" is in fact the problem not the solution. Yet, despite all the complexities around measurement, it is not uncommon in Applied Linguistics to observe simplistic approaches to measuring instruments where any set of assembled items is taken to serve as a measuring instrument. It is for this reason that language testing scholars believe that designing a measuring instrument demands expertise and assessment literacy (Harding & Kremmel, 2021; Phakiti, 2021; Purpura et al, 2015), which is often in short supply in the academic south of the world (Oakland, 2009).

A further discursive myth regarding reliability that is somewhat common in Applied Linguistics community is that reliability is a characteristic of the measuring instrument (Grabowski & Oh, 2018; Larson-Hall & Plonsky, 2015; Vacha-Haase, 1998). This myth explains our finding that in many cases, some thesis writers rely on a reported reliability in the literature rather than examining the reliability of their own data. As Rowley (1976) states "It needs to be established that an instrument itself is neither reliable nor unreliable…A single instrument can produce scores which are reliable, and other scores which are unreliable" (p. 53).

Relatedly, some measuring conventions and reliability practices seem to have become dogmatized, at least in some communities of social science and Applied Linguistics. One such dogma is the status that Cronbach alpha has come to enjoy. Some methodologists maintain that repeated use of alpha has become dogmatized, routinized, and ingrained in the culture of research in social sciences and humanities (Dunn et al., 2014), and despite the heavy scrutiny that alpha has recently come under, recommendations from statistics experts have yet to penetrate research in social science, psychology, and Applied Linguistics research (McNeish, 2018). Alpha, like many other statistics, makes certain assumptions about the data, which are often ignored by researchers (Dunn et al., 2014; McNeish, 2018). In addition, these assumptions have been demonstrated to be unrealistic and difficult to meet (Dunn et al., 2014). For the noted flaws in alpha, scholars have called for more robust ways of assessing reliability such as exploratory and confirmatory factor analysis. Yet, there seems to be a prevailing reluctance on the part of most researchers to go beyond Cronbach alpha perhaps because of the technical knowledge that is necessary for proper use, implementation, and interpretation of exploratory and confirmatory factor analysis. A further limitation that should be taken into consideration with regard to alpha is that alpha is essentially a parametric statistic assuming continuous data and non-skewed distributions (Grabowski & Oh, 2018). However, in much Applied Linguistics research, the kind of score interpretations made of quantitative data are of criterion-referenced nature with positively or negatively skewed distributions, which would require specific reliability estimation that are different from those commonly used for norm-referenced interpretations (Bachman, 2004; Brown, 2005; Brown & Hudson, 2002).

Implications

In this study, we claimed that sociomaterial and discursive considerations account for current practices and approaches to measuring instruments and their reliability in theses written in Applied Linguistics. As noted above, some of the pitfalls in measuring language and non-language constructs stem from rigid disciplinarity that characterizes current higher education structure. This insulation of disciplines results in our becoming unaware of insights and progress that is made in neighboring disciplines. As Long and Richards (1998, p. 27) maintain, “advances in language testing” remain “a closed book” for some, if not many, Applied Linguistics researchers (Chapelle, 2021). Perhaps, this is partly due to further compartmentalization that has transpired in Applied Linguistics as a result of which the sub-disciplines of the field are hardly aware of each other’s advances (Cook, 2015).

Therefore, more inter and cross-discipline dialogue and research holds the potential to deepen our understanding of sound measurement of constructs in Applied Linguistics. Some scholars go even further to suggest that Applied Linguistics must be seen as epistemic assemblage, which would strip the established sub-disciplines of Applied Linguistics of their ontological status as disciplines (Pennycook, 2018). Accordingly, to increase research rigor, we would like to call further cross-fertilization among SLA, language teaching, language testing, and even the broader field of measurement in social and physical sciences.

One curious observation we made in this study was that, in some cases, high alpha indexes were reported for proficiency tests that had been used to ensure the homogeneity of a sample of participants, often with the conclusion that the sample turned out to be homogenous. Given that parametric assumptions of alpha are violated with a homogenous sample of participants, high alpha values are almost impossible to obtain. How such high alpha coefficients have been produced remains an open question. The implication that awareness of such malpractices carries is that Cronbach's alpha and other reliability estimation procedures make assumptions about the data. Unless there is evidence that such assumptions have been met, one is not justified in using the chosen reliability estimation methods (Grabowski & Oh, 2018). Therefore, to foster research rigor, a ritualistic reporting of a high alpha coefficient is not adequate. Rather, both common sense and expertise in language assessment must be drawn upon to judge MI quality.

The other implication is that investigating and maximizing reliability must not be guided solely by practical considerations and statistical analysis. Instead, theoretical and substantive considerations should inform the process. As every research context is likely to be different, it falls on the researcher to predict and explain all the possible internal and external factors bearing on the consistency of the data collected via quantitative instruments (Grabowski & Oh, 2018). It is this context-bound nature of reliability that makes it difficult to prescribe any rule that would work across contexts for all instruments.

We would like to support the call for more rigor and conservatism in designing, adopting, and adapting measurement instruments in Applied Linguistics research. Graduate students and early career professors should not shy away from deep reflections on and involvement in the foundations of research design and data collection methods. The critique made of research in education four decades ago Pedhazur (1992 p. 368) still holds true.

There is a curious mythology about understanding and mastery of the technical aspects of research. Statistics is often called "mere statistics," and many behavioral researchers say they will use a statistician and a computer expert to analyze their data.

An artificial dichotomy between problem conception and data analysis is set up.

To think that a separate group of experts are responsible for the design and development of proper measurements and to think that the job of the research practitioner is to merely use those instruments is to perpetuate the noted artificial dichotomy between research practice and theoretical conceptions.

In sum, measurement is a tricky business even physics. In the social sciences where we work with humans, language, and discourse within complex socio-political structures, isolating, defining, and measuring constructs is very complicated. If this statement sounds radical, it is only because we in Applied Linguistics are insulated from serious debates about the ontology and epistemology of measurement (see Michell, 1999; Markus & Borsboom, 2013; Chapelle, 2020). Furthermore, the massification of higher education and the publish or perish regime in the academia has generated a mindset which takes a superficial and simplistic approach to testing complex social constructs. To improve on this situation, the fast food approach to research production (Pourmozafari, 2020) should be discouraged and countered.

Abbreviations

MI: Measurement instrument; EFL: English as a Foreign Language; RA: Research article; KR: Kuder-Richardson; SLA: Second language acquisition.

Acknowledgements

We appreciate the Editors and Reviewers of Language Testing in Asia for their timely review and feedback.

Authors' contributions

This paper is partly based on Ms. Raji's M.A thesis, which was supervised by Kioumars Razavipour. The paper however was solely written by Kioumars Razavipour. The author(s) read and approved the final manuscript.

Funding

We received no funding for conducting this study.

Availability of data and materials

Data can be supplied upon request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 6 October 2021 Accepted: 26 April 2022
Published online: 18 May 2022

References

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: a review of 30 journals' statistical quality and their CiteScore, SJR, SNIP JCR impact factors. *Language Teaching Research, 23*(6), 727–744.

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bachman, L. F., & Cohen, A. D. (Eds.). (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: a comprehensive guide to English language assessement*. New York: McGraw-Hill.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Canagarajah, S. (2018). Materializing 'competence': Perspectives from international STEM scholars. *The Modern Language Journal*, *102*(2), 268–291.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Los Angeles: Sage.

Chapelle, C. A. (2021). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11–20). New York: Routledge.

Chen, X. (2016). Challenges and strategies of teaching qualitative research in China. *Qualitative Inquiry, 22*(2), 72–86.

Cook, G. (2015). Birds out of dinosaurs: the death and life of applied linguistics. *Applied linguistics, 36*(4), 425–433.

Coole, D., & Frost, S. (2010). Introducing the new materialisms. New materialisms: ontology, agency, and politics. In D. Coole & S. Frost (Eds.), *New materialisms: Ontology, agency, and politics* (pp. 1–43).

Dant, T. (2013). *Knowledge, ideology & discourse: a sociological perspective*. London: Routledge.

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50*(1), 132–153.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing research: a conceptual gap. *Second Language Research, 17*(4), 442–456.

Douglas, D. (2014). *Understanding language testing*. London: Routledge.

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.

Ennis, R. H. (1999). Test reliability: a practical exemplification of ordinary language philosophy. *Philosophy of Education Yearbook*

Gonulal, T. (2019). Statistical knowledge and training in second language acquisition: the case of doctoral students. *ITL-International Journal of Applied Linguistics, 17*(1), 62–89.

Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *ITL-International Journal of Applied Linguistics, 168*(1), 4–32.

Grabowski, K. C., & Oh, S. (2018). Reliability analysis of instruments and data coding. In A. Phakit, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 541–565). London: Springer.

Harding, L., & Kremmel, B. (2021). SLA researcher assessment literacy. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing*. New York: Routledge.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: a note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523–531.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*. Westport, Conn: Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Khany, R., & Tazik, K. (2019). Levels of statistical use in applied linguistics research articles: from 1986 to 2015. *Journal of Quantitative Linguistics, 26*(1), 48–65. https://doi.org/10.1080/09296174.2017.1421498.

Kruglanski, A. W. (2013). *Lay epistemics and human knowledge: cognitive and motivational bases*. New York: Plenum Press.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: what gets reported and recommendations for the field. *Language Learning, 65*(S1), 127–159.

Lather, P. (1993). Fertile obsession: validity after poststructuralism. *The Sociological Quarterly, 34*(4), 673–693.

Lindstromberg, S. (2016). Inferential statistics in language teaching research: a review and ways forward. *Language Teaching Research, 20*(6), 741–768.

Loewen, S., & Gass, S. (2009). The use of statistics in L2 acquisition research. *Language Teaching, 42*(2), 181–196.

Long, M. H. & Richards, J. C. (1998). Series editors' preface. In Bachman, L. F., & Cohen, A. D. (Eds.). (1998). *Interfaces between second language acquisition and language testing research* (p. 27–28). Cambridge: Cambridge University Press.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: measurement, causation, and meaning*. New York: Routledge.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412.

Medland, E. (2019). 'I'm an assessment illiterate': towards a shared discourse of assessment literacy for external examiners. *Assessment and Evaluation in Higher Education, 44*(4), 565–580.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*(1), 113.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Michell, J. (1999). *Measurement in psychology: a critical history of a methodological concept*. Cambridge: Cambridge University Press.

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6,* 7–24.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. California: Sage.

Oakland, T. (2009). How universal are test development and use. In E. Grigorenko (Ed.), *Multicultural psychoeducational assessment* (pp. 1–40). New York: Springer.

Pallant, J. (2010). *SPSS Survival Manual (*4th ed). Open University Press: Maidenhead.

Pedhazur, E. J. (1992). In Memoriam—Fred N. Kerlinger (1910–1991). *Educational Researcher*, *21*(4), 45–45.

Pennycook, A. (2018). Applied linguistics as epistemic assemblage. *AILA Review, 31*(1), 113–134.

Phakiti, A. (2021). Likert-type Scale Construction. In P. Winke, & T. Brunfaut (eds).*The Routledge handbook of second language acquisition and language testing*

Plonsky, L. (2013). Study quality in SLA: an assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition, 35*(4), 655–687.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538–553.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: the case of interaction research. *Language Learning, 61*(2), 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x.

Plonsky, L., Egbert, J., & Laflair, G. T. (2015). Bootstrapping in applied linguistics: assessing its potential using shared data. *Applied Linguistics, 36*(5), 591–610.

Pourmozafari, D. (2020). *Personal communication*.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research 1. *Language Learning, 65*(S1), 37–75.

Rowley, G. L. (1976). Notes and comments: the reliability of observational measures. *American Educational Research Journal, 13*(1), 51–59.

Ryen, A., & Gobo, G. (2011). Editorial: managing the decline of globalized methodology. *International journal of Social Research Methodology, 14,* 411–415.

Stone, J., & Zumbo, B. D. (2016). Validity as a pragmatist project: A global concern with local application. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice* (pp. 555–573). Newcastle: Cambridge Scholars Publishing.

Thompson, B. (1988). *Common methodology mistakes in dissertations: improving dissertation quality*. Louisville, KY: Paper presented at the annual meeting of the Mid-South Educational Research Association.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6–20.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: a review of three journals. *The Journal of Experimental Education, 67*(4), 335–341.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Wilder, L. K., & Sudweeks, R. R. (2003). Reliability of ratings across studies of the BASC. *Education and Treatment of Children, 26*(4), 382–399.

Winke, P., & Brunfaut, T. (Eds.). (2021). *The Routledge handbook of second language acquisition and language testing*. New York: Routledge.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.