Validity and Reliability Reports in Applied Linguistics Research Articles: The case of tests and questionnaire

Khalil Tazik, Assistant Professor, School of Medicine, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran
khaliltazik@gmail.com

Abstract
This study intended to determine the way validity and reliability i.e., psychometric properties were reported in the Applied Linguistics research articles. The study also focused on the measurement methods applied to determine the validity and reliability of the scores derived from the tests and questionnaires in the empirical studies. The corpus of the study included 331 empirical studies derived from 733 research articles (RAs) published between 2005 and 2018 in three prominent Applied Linguistics journals – Applied Linguistics, Modern Language Journal, and TESOL Quarterly, The selected papers used test and/or questionnaire for data collection. Our analysis indicated that 77(20.98%) of the studies did not report validity and reliability measures, 82(22.35%) reported only reliability measures, 26(7.08%) reported only validity measures, and 182(49.59%) reported both the validity and reliability measures for the instruments. It was also found that content validity assessed through the pilot study had the highest frequency among validity evidences while internal consistency, mostly identified by Cronbach's alpha, was the most frequent reliability evidence.

Keywords: Validity, reliability, applied linguistics, test, questionnaire

Introducation

The common belief among all the researchers is that research findings must be legitimized and rigorous. Hence, one thing that scientific "research cannot afford is to be haphazard or lacking rigor" (Dörnyie, 2007, p. 48). Researchers should always come back and assess the quality and legitimacy of their findings. Quality of the data obtained from measurement instruments and the quality of the decisions and interpretations inferred from the data are consequential (Chan, 2014). Validity and reliability measures are the two well-known criteria for indicating the quality of the research instruments. Reviewing literature reveals that all the leading figures in the field of Applied Linguistics emphasize that the validity and reliability of the scores and the interpretations derived from the tests and questionnaires need to be checked (e.g., Cronbach, 1951; Hughes & Porter, 1983; Messick, 1989; Bachman, 1990, 2004; McNamara & Roever, 2006; Gillespie, 2012; Wools, Eggen, & Beguin, 2016, to name but a few). Perry (2008) asserts that "The strong consensus in the measurement community is that the level of confidence we can put into the findings of any given research is directly proportional to the degree to which data-gathering procedures are reliable and valid" (p. 130). Based on Chan (2014), "validity and validation are the most fundamental issues in the development, evaluation, and use of measurement instruments" (p. 9). Gethmann et al. (2015) noted that the validity of research mainly relies on "its compliance to credibility rules established within the science system" (p. 5). They believe that the acceptance of research findings rests on the well-established quality criteria in the field.

Reliability

Reliability by Dörnyie (2007) is defined as "the extent to which our measurement

instruments and procedures produce consistent results in a given population in different circumstances" (p. 50). For Perry (2008), reliability, as a quality measure, "has to do with the *consistency* of the data results" (p. 130). Variations in test method, raters, test takers, and the test itself (Bachman, 1990) can cause inconsistencies and produces unreliable results. Researchers, who use tests and questionnaires for data collection or cooperate with more than one rater or observer in their studies, expect consistent results regardless of test items or number of raters and observers (Perry, 2008). It is important to know that reliability is a psychometric property of test scores obtained from a test administered to a specific group of respondents rather than a property of the instrument. Onwuegbuzie and Leech (2005) reported that "the vast majority of quantitative researchers do not provide reliability estimates for their own data" (p. 378). Dörnyie (2007) relates this lack of reliability report to the wrong assumption that reliability is a feature of tests and research instruments rather than the scores themselves. Accordingly, Bachman (2004) emphasized that researchers must report the reliability estimates of the test scores for the instruments that they use. Perry (2008) notes that "Without knowing the reliability of a test, there is no way to know how consistent the results are" (p. 133). For reporting the reliability of test scores, Bachman (1990) introduces two approaches to estimate internal consistency: (1) an estimate based on the correlation between two sets of scores and (2) estimates based on ratios of the variances of halves or items of the test to total test score variance. In these cases, as Perry (2008) stated, correlation coefficient as a number that indicates the degree of relations between variables is the most common indicator for reporting reliability. Reliability coefficient is used in different forms across studies. For instance, in cases of inter-rater or intra-rater reliability, test-retest, and parallel form correlation coefficient is used, and for internal consistency of items, correlation, Spearman-Brown, Cronbach alpha (coefficient alpha), Kuder-Richrdson 20 & 21 are used. Cronbach alpha and Kuder-Richrdson 20 & 21 are the most commonly used statistical methods for internal consistency which are typically reported by the researchers (Perry, 2008). The score reliability is so important since it directly affects the test results and interpretations (Vacha-Haase, Ness, Nilsson, & Reetz, 1999). However, as Vacha-Hasse, et al. (1999) considered, no specific model for reporting validity and reliability was proposed across different editions of APA Publication Manual. It only encourages the authors to provide enough information for the readers regarding the validity and reliability of their findings. Perry (2008) warns those researchers who refuse to report the reliability measures for some assumptions such as the test is standardized or the test was a sub-part of a standardized test. He emphasizes that test items behave differently over different contexts; therefore, for any occasions reliability reports should be given and this cannot be taken for granted. He also noted that without reliability knowledge of a set of scores, the differences in findings might be attributed to different sources of error.

Validity

Dörnyie (2007) described the reliability as fairly straightforward concept in quantitative research; however, regarding the validity concept, he found two parallel systems in the quantitative research: construct validity (measurement validity) and its components and internal/external validity dichotomy (research validity). Dörnyie (2007) added reliability as the third part of the discussion of quantitative quality standards. Research validity which focuses on external and internal validity concerns the soundness of the whole research process. Internal validity addresses the connection between the factors measured and the study outcomes. External validity addresses the generalizability of research findings beyond the research settings. Measurement validity focuses on "the meaningfulness and appropriateness of the interpretation of

the various test scores or other assessment procedure outcomes" (p. 50). Measurement validity can be seen as a unitary concept (Messick, 1989) given in terms of construct validity or it can be branched into content, criterion, and construct validity. In any case, the focus is on the validity of the interpretation of the test scores not the validity of the scores themselves (Bachman, 1990). Chinni and Hubley (2014) noted that "Validation practices can be thought of as the tools that researchers use to build their argument and justification for the test score inference or explanation" (p. 36). Zumbo (1998) stated that without validity our inferences from the research instruments are meaningless.

Review of Literature

Messick (1989) viewed validity as a unitary concept evaluated through the integration of empirical and theoretical evidence in order to support the adequacy of inferences and interpretations derived from the test scores. This view of validity implies that one source of evidence is not sufficient for supporting validity claims (Zumbo & Chan, 2014). Moreover, Zumbo and Chan (2014) noted that validity is not an adds-on concept, but it is viewed as "an ongoing process in which various sources of validity evidence are accumulated and synthesized to support the construct validity of the interpretation and use of instruments" (p. 4). Additionally, Shear and Zumbo (2014) noted that "validity is a matter of degree rather than all or none" (p. 95). Chan (2014) viewed construct validity as the focus of validity in contemporary views of validity. He also noted that to support the construct validity of inferences, interpretations, and uses of instrument scores various sources of validity need to be accumulated and synthesized.

Construct validity is referred to the correspondence between what is measured and what the study intends to measure. According to Bijlsma-Frankema and Rousseau (2012), construct validity indicates "whether the postulated relations between the constructs provide a valid representation of the empirical world" (p. 260). He stated that this type of validity is an essential requirement for scientific constructs and theory development. Researcher triangulation, panels of researchers, a group of field experts, and technical approaches such as tests of convergent and discriminant validity with the use of factor analysis or multicollinearity statistics are some practical methods for assessing construct validity. Chan (2014, p. 12), based on an extensive review of literature and according to the Messick's (1989) unitary concept of validity, introduced five validity sources for validating inferences and uses of the scores obtained from research instruments:

1.test content evidence obtained through assessing the relationship between the intended construct and the content of an instrument
2.response processes evidence which can be obtained by the think aloud strategies
3.internal structure evidence which indicates the extent to which the items are related to each other and can be obtained through statistical techniques like item response theory
4.relations with other variables evidence which can be gathered through convergent, discriminant, criterion-related validity, and finally
5.evidence based on consequences which can be assessed based on the unintended use of the instrument and the degree to which it affects the inferred interpretations

Chan (2014) also overviewed and presented some approaches and research designs which are used in gathering evidence for validity: "factor analysis, item-test correlations, measurement invariance, differential item functioning, multitrait-multimethod design, item response theory, and experimental and quasi-experimental designs" (p. 19). He emphasized that mere existence of evidence does not guarantee the validity of inferences; the quality of evidence is also significant. Reviewing previous studies on test validation practices, Chinni and Hubley (2014, reported that

(1) the frequency of reporting validity and reliability have increased over time (2) the researchers failed to regard characteristics of selected sample during reliability and validity reports according to the previous research (3) Cronbach's alpha by far was more frequent than other reliability estimates (4) validity evidence was limited to some forms and often reported poorly (5) some "validity evidence such as response processes and consequences" (p. 37) were rarely reported, and finally and importantly (6) there was a mismatch between validation theory and practices.

McNamara and Roever (2006) stated that validity arguments are context-sensitive and pointed out that researchers need to revalidate the test score inferences in any new context of use. Similarly, Bachman (2004) emphasized that validity is context-specific and need to be reevaluated for the use in any particular situation. Likewise, Thompson (1994) explained that administering a similar instrument to two different groups will yield scores with different validity and reliability scores. Therefore, researchers should report the validity and reliability of the data extracted from the research instruments rather than the validity and reliability of the instruments themselves (Barry et al., 2014). Consequently, qualitative and quantitative evidence are required for supporting the validity of test score interpretations, conclusions, and inferences (Bachman, 2004). To attain this purpose, the researchers need to be straightforward, clear, and in depth in the methodology section of their article. In fact, the quality of research methods, instruments, and designs determines the quality of research findings (Barry et al., 2014). Therefore, this is important for the researchers to carefully include a detailed description of what they have done for satisfying the quality concerns of their research. However, Chinni and Hubley (2014) argued that a strong tendency was observed among the researchers to present some statistical findings without detailed description of their relevance to the validity. Therefore, it is quite necessary for the researchers to have a clear sense of having appropriate evaluation of different kinds of validity evidence. As argued earlier, researchers are both concerned about the measurement validity and research validity of their study. The quality concerns of each are reviewed in literature. In this study, the main focus is on the measurement validity and the way this issue has been reported in the Applied Linguistics research articles. In fact, the question posed in this study is"How did researchers report the validity and reliability concerns in Applied Linguistics RAs?"

Generally, this study aims to shed some lights on the issues and challenges involved in reporting reliability and validity measures in Applied Linguistics RAs. Drawing on this review, the researchers present the trends in reporting psychometric properties of the scores derived from the tests or questionnaires. The study also reflects on the statistical and qualitative methods used to assess different kinds of validity and reliability. The study is limited to the empirical studies in which tests and questionnaires were used as the main instruments for data collection. The main reasons for selecting these instruments are (1) their popularity in empirical studies for both large- and small-scale uses (2) their replicability across studies and (3) agreed-upon methods for assessing reliability and validity measures which provides researchers with insights into the underlying constructs of these instruments. These advantages, especially the third one, propose that the empirical studies with tests and questionnaires provide qualified results which are explicitly validated through rigorous methods.

Our findings can be regarded as important for arguing the implementation of theoretical requirements in the RAs. The journal editors need to be aware of the pitfalls of the papers since Dörnyie (2007) emphasized that falling in research pitfalls will dictate irreparable flaws in our findings. Therefore, to be cost-effective and deep in our research, we need to be aware of and/or follower of theory-driven techniques of validation studies, and in clear terms, we do what should be done. Related studies have been done in neighbouring disciplines like counseling psychology and health educaton and behavior. Meier and Davis (1990) examined the trends of psychometric

reports in counseling psychology RAs. Their findings indicated that most of the studies relied on cited estimates and did not report the psychometric estimates for their own instrument scores. Similar findings were reported by Barry et al. (2014).

Method

Corpus

The corpus of this study consisted of 331 RAs published between 2005 and 2015 in three journals: TESOL Quarterly, Applied Linguistics, and Modern Language Journal. All the empirical studies which used test and/or questionnaire were included and coded for analysis, and the non-empirical studies or those empirical studies which used instruments other than tests and/or questionnaires were excluded. Four analysts cooperated in the coding process: the researchers and three colleagues. Each analyst evaluated the coded paper independently and to check the correspondence between the evaluations, a separate session was held. All the discrapancies were identified and resolved after long discyssions. In fact, the author was in charge of making all the decisions.

Analysis

To code the validity and reliability reports, the researcher provided a coding sheet. The validity and reliability evidences presented in each paper were given in the coding sheet and named after that paper. The main sources of validity for this study were the same as what Zumbo et al. (2014) included in their study: "face, content, construct, predictive, concurrent, convergent, discriminant, response processes, consequences, reliability, internal structure, etc" (p. 69). The validity sources were coded in line with what explicitly or implicitly reported in related studies. For instance, if a paper reported that factor analysis was used to check the validity of the questionnaire, we implied that the construct validity was assessed. The authors' justifications for the use of particular kinds of validity were also included in the coding sheet. For reliability, internal consistency, parallel form, test-retest, and inter-rater evidences were coded. The articles were classified as those with (1) only validity reported, (2) only reliability reported (3) reported validity and reliability evidence for the analyzed data, (4) citing the studies reported the validity and reliability of the instruments and (5) no mention of the reliability and validity.

Results

To start with, the number of research articles (RAs) analyzed in this study and the journal from which the papers were derived are given in the following table.

Table 1. *Number (N) and Type of Research Articles Reviewed in the Study*

Journals	Non-Empirical	Empirical			Total (%)
		Qualitative	Quantitative	Mixed-methods	
Applied Linguistics	42(16.47)	148(58.04)	55(21.57)	10(3.92)	255(100)
Modern Language Journal	43(18.07)	104(43.70)	80(33.61)	11(4.62)	238(100)
TESOL Quarterly	40(16.66)	111(46.25)	74(30.84)	15(6.25)	240(100)
Total	125(17.05)	363(49.52)	209(28.52)	36(4.91)	733(100)

According to Table 1, 733 RAs were reviewed in this study among which 125 (17.05%) were non-empirical, 363(49.52%) qualitative, 209 (28.52%) quantitative, and 36(4.91%) mixed-methods. Since the study focused on the quantitative studies which used questionnaire and test for data collection, the quantitative studies as well as the quantitative methods used in the mixed-methods and those qualitative studies which used questionnaire or test were adopted as the corpus of our study. Therefore, the overall studies reviewed were 209 quantitative, 36 mixed-methods, and 56 qualitative RAs. Of these 331 RAs, 140 RAs used questionnaires, 155 RAs used tests, and 36 RAs used both test and questionnaire as the research instruments for data collection. The distribution of validity and reliability reports in the reviewed studies are given in Table 2.

Table 2. *Frequency of Validity and Reliability Reports*

	Questionnaire (N=176)		Test(N=191)		Total
	F	%	F	%	N=367
No report	36	20.45	41	21.46	77(20.98%)
Only reliability	44	25	38	19.90	82(22.35%)
Only validity	11	6.25	15	7.85	26(7.08%)
Citing validity and reliability	85	48.30	97	50.79	182(49.59%)

As Table 2 shows, our analysis indicated that 77(20.98%) of the studies did not report validity and reliability measures, 82(22.35%) reported only reliability measures, 26(7.08%) reported only validity measures, and 182(49.59%) reported both the validity and reliability measures for the instruments. Among the studies with reliability and/or validity reports, 8 studies presented two sources for reliability (e.g., inter-rater and internal consistency), 16 studies reported two sources for validity (e.g., content and construct), and the remaining 266 articles reported one source for each measure (e.g., internal consistency for reliability and content for validity). Table 3 presents the sources of validity identified in the reviewed articles. It is worth noting that to make the data distinguishable, the types of validity and reliability were separately given for the questionnaires and tests.

According to Table 3, of the 367 papers reviewed in this study, 208 papers reported validity evidence (see Tables 2 and 3). Content validity, citing validity reports of previous studies, construct validity, concurrent and face validity were the observed validity evidence for both questionnaires and tests. Theoretical discussion and predictive validity, however, were only used as validity evidence for tests. Generally, content validity showed the highest frequency and observed in 91(43.75%) papers. About 55 (26.44%) papers relied on the validity reports of previous studies as validity evidence for the instruments. Construct validity evidence was reported in 24(11.54%) of the papers. This trend of reporting was similar for the questionnaires and tests. In providing two sources for validity, content and construct validity (5.21%) for questionnaires, and content and face/construct (3.57% / 2.68%) for tests were more observed. One implication is that the validity types reported for the questionnaires and tests were somehow similar. The differences just observed in the methods employed for assessing the validity types.

As Table 3 shows, different methods were used to study different kinds of validity. The methods employed to study content validity were pilot study, expert judgment, and literature review among which pilot study (N=73, 35.09%) was the most frequent method. For construct validity, factor analysis, correlation coefficient, and explaining the process of test construction were use among which factor analysis (N= 21, 10.09%) was more observed. Correlation

coefficient was used for measuring concurrent and predictive validity, and expert judgment and interview were used for face and consequential validity, respectively. Among all the methods, pilot study was the most frequent method used in 76 (36.54%) papers. However, in about 43 (56.58%) of these papers, little information for the process of pilot study was presented.

Regarding reliability reports, 264 articles provided reliability estimates among which internal consistency (45.45%), mostly identified by Cronbach's alpha (77.50%), was the most frequent one. This dominance of internal consistency estimates implied that the classical true score approaches to reliability are more favorable to the researchers. Table 4 depicts the types of reliability reports and their associated methods found in reviewed studies.

Table 3. *Sources Reported for Validity and Their Related Statistics Types*

Questionnaire One source reported (N= 90)	Test One source reported (N= 102)	Total N= 208
Content (43, 44.79%) pilot study: 27 Experts' judgment: 14 Literature review: 2	Content (48, 42.86%) pilot study: 37 Experts' judgment: 11	91(43.75%)
Cited validity reports (29, 30.21%)	Cited validity reports (26, 23.21%)	55(26.44%)
Construct (12, 12.5%) Factor analysis: 9 Correlation coefficient: 3	Construct (12, 10.71%) Factor analysis: 4 Correlation coefficient: 3 Process of test construction: 5	24(11.54%)
Concurrent (3, 3.12%) Correlation coefficient: 3	Concurrent (4, 3.57%) Correlation coefficient: 3	7(3.36%)
Face (3, 3.12%) Expert judgment: 3	Face (3, 2.68%) Expert judgment: 3	6(2.88%)
	Theoretical discussion (5, 4.46%)	5(2.40%)
	Consequential (2, 1.78%)	2(0.96%)
	Interview: 2	
	Predictive (2, 1.78%) Correlation coefficient: 2	2(0.96%)
Two sources reported (N= 6)	Two sources reported (N= 10)	
Content and Construct (5, 5.21%) Pilot study and factor analysis: 3 Expert judgment and factor analysis: 2	Content and Construct (3, 2.68%) Pilot study and factor analysis: 3	8(3.84%)
Content and face (1, 1.04%) Pilot study and expert judgment: 1	Content and face (4, 3.57%) Pilot study and expert judgment: 4	5(2.40%)
	Construct and predictive (1, .89%) Correlation coefficient: 1	1(0.48%)
	Construct and consequential (1, .89%)	1(0.48%)
	Correlation and interview: 1	
	Content and concurrent (1, .89%)	1(0.48%)

	Pilot study and correlation coefficient: 1

Table 4. *Sources Reported for Reliability and Their Related Statistics Types*

Questionnaire	Test	Total
One source reported (N= 127)	One source reported (N= 129)	N= 264
Internal consistency (66, 51.16%) Cronbach's alpha: 54 Correlation Coefficient: 5 KR-21: 7	Internal consistency (54, 40%) Cronbach's alpha: 39 Correlation Coefficient: 3 KR-21: 6 Guttman's split-half: 2 Spearman's Brown: 2 G-study: 1 Rasch model: 1	120(45.45%)
Inter-rater (29, 22.48%) Correlation Coefficient: 17 Cohen's Kappa: 6 Pearson Product Moment: 2 Kappa coefficient: 2 Cronbach's alpha: 2	Inter-rater (51, 37.78%) Correlation Coefficient: 35 Cohen's Kappa: 8 Pearson Product Moment: 2 Kappa coefficient: 2 Spearman's rho: 2 Cronbach's alpha: 2	80(30.30%)
Cited reliability (32, 24.80%)	Cited reliability (19, 14.07%)	51(19.32%)
	Test-retest (4, 2.96%) Correlation coefficient: 4	4(1.51%)
	Parallel-form (1, .74%) Correlation coefficient: 1	1(0.38%)
Two sources reported (N= 2)	Two sources reported (N= 6)	
Internal consistency and inter-rater (2, 1.55%)	Internal consistency and inter-rater (3, 2.22%)	5(1.89%)
Cronbach's alpha and correlation coefficient: 2	Cronbach's alpha and correlation coefficient: 3	
	Internal consistency and test-retest (2, 1.48%)	2(0.75%)
	Cronbach's alpha and correlation coefficient: 2	
	Test-retest and inter-rater (1, .74%)	1(0.38%)
	Correlation coefficient and correlation coefficient: 2	

Referring to Table 4, it is shown that inter-rater reliability was used in 80(30.30%) papers as regarded as the second frequently observed reliability type. The most frequent method used to assess this kind of reliability was correlation coefficient (65%). In about 51(19.32%) papers,

reliability measures found in previous studies were reported as the reliability measures for the new context of use. The high frequency of internal consistency, inter-rater, and citing previous studies were observed in reporting reliability for questionnaires and tests similarly. Test-retest and parallel reliabilities assessed by correlation coefficient were just used in reporting reliability of the tests. In cases of two-source reliability reports, internal consistency and inter-rater were more observed in the papers (N= 5, 1.89%).

Discussion

The study was inspired by the emphasis given to the validity and reliability evidences for the instruments used in Applied Linguistic RAs. However, results of this study indicated that about one-fifth (20.98%) of the papers failed to report reliability and validity of the instruments, 22.35% failed to report validity, and 7.08% did not report reliability measure for the instruments. The lack of reliability and validity reports in the RAs reviewed indicated that the authors followed the conventions of publishing in journals and ignored the theoretical requirements of reporting quality concerns for the research findings. Vache-Haasse et al. (1999) related this shortcoming to the inability of authors and readers in determining "intelligently the extent to which score measurement error affects the results and their interpretations" (p. 340). The lack of reliability and validity report makes it possible to conclude that researchers might be measuring wrong constructs and making erroneous conclusions. The lack or/and insufficiency of validity and reliability explanations were also reported in some other studies, e.g., in the Journal of Counseling Psychology reported by Meier and Davis (1990) and in the field of health education by Barry et al. (2014). Barry et al. (2014) assert that "by not ensuring the instruments employed in a given study were able to produce accurate and consistent scores, researchers cannot be certain they actually measured the behaviors and/or constructs reported" (p. 16). This is true for the cases of translation or changes in test or questionnaire items. In the cases of translation, it was found that except three articles, the researchers did not provide any explanation about the translators and their qualifications. In cases of item changes very little information was provided. Failure to provide in-depth explanations would endanger the quality of Applied Linguists' findings. Some researchers may claim that they trim their papers because of space limitation. However, it is suggested that the detailed processes for validity and reliability should be submitted to the editors in the complementary data.

According to the analysis, it was revealed that the theoretical strengths of the validity and reliability are weakly practiced in the Applied Linguistics literature. Generally, our review indicated some serious limitations:

Construct Validity: Our findings indicated that the researchers have not indorsed in the unitary concept of validity and mostly reported different kinds of validity rather than construct validity. It was also found that the reported information for evaluating the construct validity of the instruments were insufficient, cursory, and mostly summarized in reliability statistics. In cases that construct validity was concerned, the exploratory and confirmatory factor analysis were used. As a matter of fact, what is directed in the theory of construct validity has been observed to be less visible in the real practice in empirical studies. This trend was reported by Bijlsma-Frankema and Rousseau (2012). They noted that even in highly ranked publications, considerable attention is given to the technical methods used to legitimize the measures, while these methods have little to do with the construct validity. In other words, the researchers tend to report construct validity based on conventional practices and fail to follow the actual theoretical requirements. One of the consequences of failing to report the validation process of research

methods is the lack of public discussion of construct validity in theory-directed way. It also suffers the thorough relations between concepts in theory building (Bijlsma-Frankema and Rousseau, 2012).

Lack of Integration of validity and reliability reports: The current state of Applied Linguistics reflects its multidimensionality and diversity over different contexts and concepts. One of the consequences of this diversity is the appearance of different areas of study with distinct data collection instruments. This has led to the fragmentation of the use of research instruments in the field as well as the diversity in validity and reliability reports. It is observed that content validity measured through pilot study and internal consistency measured through Cronbach's alpha were the most frequent validity and reliability types reported in about half of the studies. In other studies no clear trend for reporting validity and reliability was observed. For instance, in a study by Vandergrift (2005) published in Applied Linguistics journal, 3 questionnaires were used: the first one had no report of validity but internal consistency through Cronbach's alpha for reliability, validity of the second questionnaire was previously reported, and the validity of the third one was assessed by Cronbach's alpha, "A motivation questionnaire, the Language Learning Orientations Scale (see Appendix A), previously validated by Noels et al. (2000), consists of twenty randomly ordered statements designed to assess AM, the three types of EM, and the three types of IM" (p. 76)….. A listening comprehension test, developed from previously elaborated tests for core French students (Lapkin 1994; Wesche, Peters, and MacFarlane 1994), was validated with another class for the purpose of this study with an acceptable Cronbach alpha of .83 (p. 76).

Lack of Validity Concerns for the Use of Instruments for a Range of Targets: Most of the instruments used in the field of Applied Linguistics were developed and validated to assess particular variables in specific contexts. Yet, in 26.44% papers it was assumed that the instruments are validated for every related target and can be used over different contexts, without any validity check for the new context and target (Examples, 1 & 2).

Example 1. "Previous validation studies have suggested that the scale is both valid and reliable (Cornwell and McKay 2000)" (Pae, 2012, Applied Linguistics, p. 238).

Example 2. "The test has shown very high internal consistency (above .90) in several studies (Aida, 1994; Horwitz, 1986; Liu & Jackson, 2008; Rodriguez & Abreu, 2003) …… The test–retest reliability of the questionnaire was .83 in Horwitz (1986) and .80 in Aida's (1994) study. Its validity, too, has been supported or partially supported by research (e.g., Aida, 1994; Argaman & Abu–Rabia, 2002, 2002; Horwitz, 1986; Liu & Jackson, 2008; MacIntyre & Gardner, 1989; Rodriguez & Abreu, 2003)" (Shao, Ji, & Yu, 2013, Modern Language Journal).

Researchers used this justification as a reason for the use of the instrument, without any validity check for its use in their study. Gillespie (2012) emphasized that "publication does necessarily mean it is well validated" (p. 182). Therefore, any instrument that needs to be used in our study should be assessed for its reliability and validity. Bachman (2004) emphasized that language tests are context-specific, and they should be checked for their validity to be used in new contexts. Gillespie (2012) pointed out that "even when instruments are designed to assess a range of targets, often items are worded in a way that makes translation to other targets difficult" (p. 179). Therefore, one valid item for assessing a research variable might not be meaningful and valid for assessing another related variable. Therefore, one should be cautious about the context-specificity, replicability, and generalizability of research instruments before using them as valid instruments for data collection. Another point noted by Gillespie (2012) was the use of

instruments across national cultural contexts. Our analysis did not show any concern about the use of different instruments over different cultural contexts. As a conclusion, neglecting these issues in the process of data collection decreases the validity of research findings. Through the development of well-validated research instruments, more replication studies with higher confidence in their results could be conducted.

Priority of reliability reports over validity

While it is emphasized that validity of test scores and the produced data are of great importance in examining the quality of findings, our findings indicated that most researchers tended to report reliability over validity. This propensity was observed in studies that assessed internal consistency through Cronbach's alpha (Example 3).

Example 3. "Parental encouragement [four items out of which three were originally developed by Gardner (1985) and one additional item from Dornyei et al. (2006)]: the extent to which parents support their children in studying English. Example: my parents really encourage me to study English. (Cronbach alpha= 0.83.)" (Kormos, Kiddle, & Csizer, 2011, Applied Linguistics, p. 502).

Of course, assessing and reporting consistency of findings are important; however, as Barry et al. (2014) argued, "although reliability statistics are important and give readers insight into the consistency of the scales used, it does little to quell concerns associated with the accuracy of the findings" (p. 16). Therefore, it is quintessential that Applied Linguistics researchers report both reliability and validity for assuring the quality of research findings. If these properties are neglected, the time, fund, and the decisions made according to the research findings would be wasted and inaccurate.

The gap between the conceptualization of validity and reliability and their measurements: It seems that validity theory does not determine the validity practice in academic journal of Applied Linguistics, but the genre of validity reports in the journals determines the theory and practices of validity. This is a concern because reporting any specific validity evidence must be in line with the purposes of measurement. Moreover, validity is an adds-on concept for which different kinds of evidence need to be accumulated. However, it was found that such a view which was originally presented by Messick (1989) was not penetrated in the validity practices. Moreover, results of our analysis indicated that the frequency of reliability and validity reports in the reviewed papers lagged behind the theoretical expectations. The main concern is that the conventional trend in practicing validity has shaped the theoretical conception of validity.

Conclusuion

In conclusion, regarding the research instruments used in the quantitative studies, it is discussed that the instruments must satisfy the psychometric properties (i.e., reliability and validity). If the researchers ignore these properties, they may make erroneous conclusions. What we observed in the analysis is the lack of consistent and common validity and reliability reports among the researchers. The actual theory-driven practices in academic studies can be useful for enhancing "the chance of agreement on meanings of constructs and their valid measurements" which can result in "enhancing the comparability of findings and the confidence in their generality" (Bijlsma-Frankema & Rousseau, 2012, p. 271). Therefore, efforts to evaluate the research instruments for their validity and reliability and making it a common discussion and practice among the scholars can be very helpful in enhancing the quality of our studies. The

cooperation of theorists and the practitioners are also helpful for strengthening the validity of score interpretations and inferences. Finally, it is recommended that researchers have to be transparent and complete in their validity reports and practices. They have to follow the validity guidelines and standards for enhancing the quality of their findings. Moreover, as Chan et al. (2014) points out, journal editors can also have an important role in this regard; in fact, the editors are "in the best position to promote the use of guidelines for the reporting of validity evidence" (p. 82). They also recommended including some validity courses in the graduate and post-graduate curriculum.

References

Ashleigh, M. J. & Meyer, E. (2012). Deepening the understanding of trust: combining repertory grid and narrative to explore the uniqueness of trust. In F. Lyon, G. Mollering, & M. N. K. Saunders (Eds.), *Handbook of Research Methods in Trust* (pp. 138-148). UK: Edward Elgar Publishing Limited.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing* (7th Ed.). Oxford: Oxford University Press.

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

Bachman, R. (2012). Utilising repertory grids in macro-level comparative studies. In F. Lyon, G. Mollering, & M. N. K. Saunders (Eds.), *Handbook of Research Methods in Trust* (pp. 130-137). UK: Edward Elgar Publishing Limited.

Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: a review of seven journals. *Health Education & Behavior, 41*(1), 12-18.

Bijlsma-Frankema, K. M. & Rousseau, D. M. (2012). It takes a community to make a difference: evaluating quality procedures and practices in trust research. In F. Lyon, G. Mollering, & M. N. K. Saunders (Eds.), *Handbook of Research Methods in Trust* (pp. 259-276). UK: Edward Elgar Publishing Limited.

Chan, E. K. H. (2014). Standards and guidelines for validation practices: development and evaluation of measurement instruments. In B. D. Zumbo and E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 9-24). Springer International Publishing.

Chan, E. K. H., Munro, D. W., Huang, A. H. S., Zumbo, B. D., Vojdanijahromi, R., Ark, N. (2014). Validation practices in counseling: major journals, mattering instruments, and the Kuder occupational interest review. In B. D. Zumbo and E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 67-90). Springer International Publishing.

Chinni, M. L., & Hubley, A. M. (2014). A research synthesis of validation practices used to evaluate the satisfaction with life scale. In B. D. Zumbo and E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 35-66). Springer International Publishing.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16, 292-334.

Dörnyie, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*. Oxford University Press.

Gethmann, C. F., Carrier, M., Hanekamp, G., Kaiser, M., Kamp, G., Lingner, S., Quante, M., & Thiele, F. (2015). *Interdisciplinary Research and Trans-disciplinary Validity Claims*,

Ethics of Science and Technology Assessment 43, Springer.

Gillespie, N. (2012). Measuring trust in organizational contexts: an overview of survey-based measures. In F. Lyon, G. Mollering, & M. N. K. Saunders (Eds.), *Handbook of Research Methods in Trust* (pp. 175-188). UK: Edward Elgar Publishing Limited.

Hesse-Biber, Sh. N. (2010). *Mixed Methods Research: Merging Theory with Practice*. NY: The Guilford Press.

Hughes, A., & Porter, D. (1983). *Current Development in Language Testing*. London: Academic Press.

Kormos, J., Kiddle, Th., & Csizer, K. (2011). Systems of goals, attitudes, and self-related beliefs in second-language-learning motivation. *Applied Linguistics, 32*(5), 495-516.

Lyon, F. (2012). Access and non- probability sampling in qualitative research on trust. In F. Lyon, G. Mollering, & M. N. K. Saunders (Eds.), *Handbook of Research Methods in Trust* (pp. 85-93). UK: Edward Elgar Publishing Limited.

Lyon, F., Mollering, G., & Saunders, M. N. K. (2012). *Handbook of Research Methods in Trust*. UK: Edward Elgar Publishing Limited.

McNamara, T., & Roever, C. **(**2006**).** *Language Testing: The Social Dimension*. Oxford, UK: Blackwell Publishing.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal* of *Counseling Psychology, 37,* 113-115.

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

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: the importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375-387.

Pae, T. (2012). Skill-based L2 anxieties revisited: their intra-relations and the inter-relations with general foreign language anxiety. *Applied Linguistics, 34*(2), 232-252.

Russell, J. D. (1996). An approach to organizational ethnographic research: strategy, methods and processes. In *Families and Social Capital ESRC Research Group Departmental Discussion Paper* (p. 36). London: South Bank University.

Scandura, T. A. & Williams E. A. (2000). Research methodology in management: current practices, trends, and implications of future research. *Academy of Management Journal, 43* (6), 1248–64.

Shao, K., Ji, Zh., & Yu, W. (2013). An exploration of Chinese EFL students' emotional intelligence and foreign language anxiety. *Modern Language Journal, 97*(4), 917-929.

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: a review of validity studies in educational and psychological measurement. In B. D. Zumbo and E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 91-112). Springer International Publishing.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837-847.

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

Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive

awareness and proficiency in L2 listening. *Applied Linguistics*, 26(1), 70-89.

Wools, S., Eggen, T., & Beguin, A. A. (2016). Constructing validity arguments for test combinations. *Studies in Educational Evaluation*, 48, 10-18.

Zumbo, B. D. (1998). Opening remarks to the special issue on Validity theory and the methods used in validation: Perspectives from the social and behavioral sciences. *Social Indicators Research*, 45, 1–3.

Zumbo, B. D., & Chan, E. K. H. (2014). *Validity and Validation in Social, Behavioral, and Health Sciences*. Springer International Publishing.

Zumbo, B. D., Chan, E. K. H., Chen, M. Y., Zhang, W., Darmawanti, I., & Mulyana, O. P. (2014). Reporting of measurement validity in articles published in Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement. In B. D. Zumbo and E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 27-34). Springer International Publishing.