Issues of Validity and Reliability in Foreign and Second Language Proficiency: A Review

Eunjeong Park[1]

[1]*(Department of English Education, College of Education/ Sunchon National University, South Korea)*

ABSTRACT: *This review paper examined prior literature on the issues of validity and reliability for assessing foreign and second language proficiency. This article discusses the issues of validity and reliability in the context of foreign and second language assessment. Despite to a lack of legitimate studies on foreign and second language proficiency connected to validity and reliability issues, several relevant studies were collected, reviewed, and synthesized in this paper.*

KEYWORDS – *Validity, validation, reliability, foreign/second language proficiency*

I. INTRODUCTION

With an inundated number of foreign and second language learners of English, language proficiency assessment has recently aroused a huge attention in the field of language education. Among many issues involved in language assessment, validity and reliability have been of great concern to language teachers and educators. Current inquiries into the issues of validity and reliability in second language proficiency manifest the field of language assessment with numerous outlooks and the use of cutting-edge research methodologies in encountering them [1,2].

First, validity has been considered as the most essential quality component in testing and assessment, concerning the extent to which meaningful inferences can be drawn from test scores [3]. Messick (1989) has conceptualized validity as the interpretation of test scores as opposed to its traditional conception as a property of a test [4]. Messick (1989) regards validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" [4, p. 13]. Validity requires a validation process by which a stakeholder organizes evidence to support the inferences of decisions made based on test scores [5].

Reliability is the second quality component as a prerequisite to validity in testing and assessment. Tests must provide consistent and reliable information about test-takers' language proficiency and performance. Reliability relies on the simulation of the test tasks and the consistency of the ratings [2]. The administration of language tests may vary in different contexts at different times; thus, it may cause inconsistency of ratings. Furthermore, language assessment requires human or mechanical raters' judgments. The reliability issue seems more complicated when it comes to subjective human judgements, which may lead to disagreement among the raters [6].

Due to their complex and technical nature, attention to the issues of validity and reliability in language proficiency assessment should continue in order to develop and advance the field of foreign and second language assessment. This review paper highlights critical issues of validity and reliability in selected prior literature by discussing how the professional and scholars encountered and resolved them and presents implications for future research on foreign and second language assessment with a focus on validity and reliability.

II. MAJOR AREAS AND COMPONENTS OF VALIDITY

Quite a few studies examined validity of foreign language proficiency tests. In particular, various types of validity were used for validation processes. Raymond and Roberts' (1983) study examined the foreign language attitude scale (FLAS) by comparing the validity coefficients of the FLAS with those of ability predictors. Stepwise regression analyses were employed to see if the FLAS predicts foreign language grades [7]. The findings suggest that the moderate relationship between the FLAS and foreign language proficiency may be used to enhance the validity in the prediction of success of foreign language learning. The major issue was a scale validation by conducting an item analysis, and predictive validity of the attitude scale was introduced in this study.

The studies of MacIntyre and Charos (1996) and Kucuk and Walters (2009) focused on predictive validity. MacIntyre and Charos' (1996) study established the significance of affective variables (i.e., attitudes, motivation, perceived competence, and anxiety) to predict success in second language learning and communication [8]. A path analysis was employed to examine the relations among the affective variables and the impact on the frequency of second language communication. The results showed that attitudes and anxiety set the psychological context for second language communication. MacIntyre and Charos' (1996) study is significant because they revealed the ratings on a 7-point Likert-type scale have acceptable convergent and predictive validity [8]. Kucuk and Walters' (2009) research focused on predictive and face validity of tests administered in an EFL university setting [9]. This study addressed the question of how well face validity reflects more objective measures of the quality of a test, such as predictive validity and reliability. The study also discussed ways of evaluating achievement tests—what measures can and should be taken—to ensure that achievement tests pursue the right purposes. Predictive validity was estimated by correlating the scores of the midterm achievement tests with those of the final exam and the final exam scores with grades from GE classes. Face validity was determined by asking both instructors and students about their perceptions of how well the contents of the courses were signified on the achievement tests. Despite the under-representation of the contents on the tests, the analyses concluded that the instructors and students regarded the tests as possessing a high degree of face validity. Kucuk and Walters concluded that the assessment of face validity fairly accurately reflects more objective measures of test quality.

Marian et al. (2007) study indicated the purpose of study is to develop a reliable and valid questionnaire of bilingual language status with predictable relationships between self-reported and behavioral measures [10]. Two studies were conducted for establishing internal and criterion-referenced validity. Factor analyses revealed consistent factors across both studies and suggested that the Language Experience and Proficiency Questionnaire (LEAP-Q) was internally valid. Multiple regression and correlation analyses were employed for the criterion-based validity. The results suggested that self-reports were reliable indicators of language performance. Self-reported reading proficiency was a predictor of first language performance; self-reported speaking proficiency was a predictor of second language performance. The study concluded that the LEAP-Q is a valid, reliable, and efficient tool for assessing the language profiles of multilingual adult learners.

Yu's (2013) study examined the interrelationships of five constructs: integrative motivation, perceived communication competence in second language (L2), sociocultural adaptation, academic adaptation, and persistence of international students at an Australian university [11]. Yu employed structural equation modelling to examine the interrelationships of five constructs: integrative motivation, perceived L2 communication competence, sociocultural adaptation, academic adaptation, and persistence of international students. Liu's (2007) study focused on pragmatic knowledge by developing a multiple-choice discourse completion test (MDCT) [12]. Liu conducted a situation pilot study to test whether each of the 15 situations elicited the speech act of apology and to get preliminary data for constructing the options for each MDCT item. First, to obtain construct validity, the raters were asked to highlight the inadequate parts and write the reasons for the inappropriateness. Verbal protocols and Rasch analysis were used to probe the thinking processes of the test-takers drawing on construct-relevant knowledge. The verbal protocols revealed that the students could identify the correct speech act and display awareness of the variables (severity, status, and familiarity). Content validity was carefully observed, based on prior research [13,14].

III. MAJOR AREAS AND COMPONENTS OF RELIABILITY

Reliability is another critical quality component in foreign and second language assessment and testing. Several studies focused on reliability issues in light of foreign and second language assessment. Clément's (1986) study investigated the relationship between the language status and individual differences in attitudes and motivation relevant to proficiency and acculturation in a second language [15]. The findings showed that the level of acculturation was a function of proficiency in the second language and an interactive function of language status and frequency of contact. Correlational analyses revealed that proficiency and acculturation were most strongly associated with self-confidence. Neither attitudes nor motivation had an impact on language outcome and language status. The study included the estimates of internal consistency reliability (Cronbach alpha) for supporting the reliability of the composite scores. Dao, Lee, and Chang's (2007) study examined the relationship between acculturation, perceived English fluency, social support, and depression among Taiwanese graduate students [16]. Ordinary Least Squares analyses were conducted for the study and the results indicated that the perception of English fluency was important in predicting depression among males and females. Four different instruments were employed to investigate the relationship between the four variables (i.e., acculturation, perceived English fluency, social support, and depression) through multiple regression analyses. Internal consistency reliability coefficients were reported for the instrument reliability. Liu's (2007) study also stood out in terms of reliability [12]. For reliability, a situation likelihood investigation was conducted to ensure the authenticity of the situations. The equivalence of the two versions was checked by means of back translation–the situations were translated into Chinese first, and back to English by another translator. A comparison of the original and the back-translated version showed consistency in the description of the situations. Internal consistency reliability was computed by means of Cronbach's alpha ($\alpha = .83$). A metapragmatic judgment test confirmed degrees of equivalence between two languages at the sociocultural and pragma-linguistic levels.

Ardasheva et al.'s (2012) study also indicated issues of reliability. The test for invariance across two samples supported the validity of the measure [17]. Factor Analysis (i.e., exploratory and confirmatory factor analyses) showed a three-factor solution with intrinsic motivation, introjected regulation, and external regulation as the best model, compared to the expected four-factor solution. For reliability, the study suggested establishing internal consistency reliability. The reliability statistics' range indicated moderate to high level of internal consistency and approached those reported in studies with similar age populations [18,19] (Ryan & Connell, 1989; Vandergrift, 2005). Lowinger et al.'s (2014) study examined the relationship between academic self-efficacy, acculturation difficulties, and language abilities on procrastination behaviors [20]. 264 Chinese international students participated in the study via convenient sampling. The results indicated that the impact of the independent variables (i.e., academic self-efficacy, acculturation difficulties, and language abilities) on procrastination behaviors varied by gender.

IV. DISCUSSION

1.1 Issues of Validity

The first issue is construct validity. MacIntyre and Charos' (1996) study mainly focused on an attempt to replicate relations in a model of language learning motivation and a model of willingness to communicate and to assess interrelations between those models [8]. That is why the study used the variables with higher reliability coefficients to conduct the path analysis. This study measured convergent and predictive validity as subcategories of construct validity. The researchers should demonstrate the evidence of both convergent and discriminant validity for construct validity. In other words, measures of constructs theoretically-related to each other should show convergence between similar constructs. Conversely, measures of constructs theoretically unrelated to each other should discriminate between dissimilar constructs. On top of the divergent direction, the convergent correlation coefficients should always be higher than the discriminant ones. In this sense, MacIntyre and Charos' (1996) study should have examined discriminant validity so that the results may reveal the reasonable measurement of construct validity [8]. Liu' (2007) study also underlined construct and content validity obtained by means of verbal protocols to examine the thinking processes [12]. Ericsson and Simon (1984) advocated particularly concurrent and think-aloud methodologies because verbal protocols do not change people's cognitive processes and are relatively complete accounts of cognitive processes [21]. Hughes (1989) explicated that construct validity can be investigated through two principal methods: think aloud and retrospection [22]. Wilson (1994) also strongly supported verbal protocols as the excellent methodology to study the contents of consciousness [23].

The second issue is about predictive validity. Raymond and Roberts' (1983) study maintained that the FLAS augmented the accuracy of the prediction of foreign language grades by entering regression equations after the dominant ability predictor and implying the unique prediction of some of the variance in foreign language proficiency in the attitude scale [7]. Unlike a moderate degree of predictive validity, this study identified a small degree of support for the construct validity of the FLAS. Some psychometrics have recommended multiple regressions as a means to calculate predictive validity [24]. However, a lot of counterarguments have also witnessed that multiple regressions may not be a reasonable indicator to measure predictive validity. For example, Olea and Ree (1994) acknowledged that multiple R and Wherry's adjusted multiple R grossly over-predict the validity of the set of predictors in a new sample and are not appropriate estimates for comparing the validity of factors and facets. Several formulas were developed to statistically estimate the cross-validated multiple correlation, and its efficiency was empirically demonstrated [25]. However, Olea and Ree argued that the use of the inadequate formula makes the validity of the facets overestimated [25]. For this reason, the alternative would be to use path analysis. Path analysis is used to describe the directed dependencies among a set of variables. Path analysis includes models equivalent to any form of multiple regression analysis, factor analysis, canonical correlation analysis, discriminant analysis, as well as families of models in the multivariate analysis of variance and covariance analyses [26]. Keith and Reynolds (1990) suggested path analysis as a means of assessing predictive bias [27]. A limitation of this approach is that no true ability measures exist. Therefore, the researchers should consider the plausibility and interpretability of predictive validity if it is measured in their research.

The third issue is about the distinction between validity and validation processes. Borsboom, Mellenbergh, and van Heerden (2004) assert that there should be a distinction between validity and validation made for scales [28]. Borsboom et al.'s (2004) argument is that validity is a property, representing an ideal situation with the ontological view; whereas, validation is an activity that researchers strive to find out whether a test contains the property of validity with the epistemological view [28]. Raymond and Roberts' (1983) study attempted to reveal the epistemological view of validity through the scale validation. Hence, the first issue seems to be convincing for validating the scales used for the study [7].

To make a concrete construct validation, Millon, et al.'s (2009) study suggests that test construction undertake three stages of validation: theoretical-substantive validity, internal-structural validity, and external-criterion validity [29]. The development is an iterative process, with each step reanalyzed each time items were added or eliminated. According to Millon, et al. (2009), theoretical-substantive validity is the first stage of a deductive approach by developing a large pool of items; the number of items can be reduced based on the degree to which they fit a theory of the proposed study [29]. The second validation stage includes internal-structural validity by assessing how well items are interrelated and the psychometric properties of the test are well determined. Millon, et al. indicated two measurement scales: internal consistency and test-retest reliability [29]. The final validation stage includes external-criterion validity: convergent and discriminative validity of the test. Positive predictive power is the likelihood of being right given a test positive, which ranged from .30 to .81.

1.1 Issues of Reliability

Regarding reliability issues, internal consistency was the major reliability index in most studies. Clément's (1986) study focused on internal consistency and inter-rater reliability, using both factor analysis and multiple regressions [15]. Psychometricians and statisticians have agreed upon a function of factor analysis. For example, Williams et al. (2012) argued that factor analysis is “a multivariate statistical procedure that has many uses…it provides ‘construct validity’ evidence of self-reporting scales” [30, p. 2]. Goodwin (1999) also claimed that factor analysis has played a crucial role in attempts to estimate “construct validity” [31, p. 86]. Dao, et al.’s (2007) study employed multiple regression analyses to examine relationships between acculturation, perceived English fluency, social support, and depression, but it could have used Guttman split-half coefficient [16]. Split-half is a measure of internal consistency that involves dividing a test into two equivalent halves [32]. Split-half reliability coefficients would support the reliability of each instrument used in the study. Another method would be average inter-item correlation and average item-total correlation. The average inter-item correlation uses all of the items on the instrument with the same construct; the average item-total correlation is a correlation between the item score and the overall score. An item-total correlation test is useful to check if any item in the instrument is inconsistent with the averaged behavior of the others.

V. CONCLUSION

Among the reviewed studies, correlational analysis was the most frequently-used method for reliability and multiple regression and factor analysis followed the next for validity. One third of the studies reported predictive validity, and many studies revealed convergent, content, and face validity. For reliability, half of the studies estimated internal consistency reliability through correlation, multiple regression, and factor analyses. Hence, it can be said that the researchers tend to report internal consistency reliability when they conduct their studies containing instruments of survey questionnaires and test items.

The most important thing from this review would be the argument about estimating predictive validity. Predictive validity can be obtained through multiple regressions. However, multiple regressions may not be appropriate to measure predictive validity because multiple R and Wherry's adjusted multiple R grossly over-predict the validity of the set of predictors in a new sample. The use of the inadequate formula in multiple regression analysis may overestimate the validity of the facets. The use of path analysis as the alternative of estimating predictive validity [27]. There is still a limitation of path analysis, like no true ability to measure due to latent variables. Hence, the researchers should take into consideration about the plausibility and interpretability of predictive validity if it is measured in their research.

REFERENCES

[1] A. Tasgin, and M.Korucuk, Development of foreign language lesson satisfaction scale (FLSS): Validity and reliability study. *Journal of Curriculum and Teaching, 7*(2), 2018, 66-77. https://doi.org/10.5430/jct.v7n2p66

[2] Y.-F.Liao, Issues of validity and reliability in second language performance assessment. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 4*(2), 2004, 1-4.

[3] L. Bachman, *Fundamental considerations in language testing* (Oxford: Oxford University Press, 1990).

[4] S. Messick, Validity, in R. L. Linn (Ed.), *Educational measurement*(New York: Macmillan, 1989).

[5] L. Crocker, and J. Algina, *Introduction to classical and modern test theory*(Belmont, CA: Wadsworth/Thomson Learning, 1986).

[6] T. McNamara, *Measuring second language performance* (London: Longman, 1996).

[7] M. R. Raymond, and D. M. Roberts, Development and validation of a foreign language attitude scale. *Educational and Psychological Measurement, 43*, 1983, 1239-1246.

[8] P. D. MacIntyre, and C. Charos,Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology, 15*(1), 1996, 3-26.

[9] F. Kucuk, and J. Walters, How good is your test? *ELT Journal, 63*(4), 2009, 332-341.

[10] V. Marian, H. K. Blumenfeld, and M.Kaushanskaya, The language experience and proficiency questionnaire: Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50,* 2007, 940-967.

[11] B. Yu, Asian international students at an Australian university: Mapping the paths between integrative motivation, competence in L2 communication, cross-cultural adaptation and persistence with structural equation modelling. *Journal of Multilingual and Multicultural Development, 34*(7), 2013, 727-742.

[12] J. Liu, Developing a pragmatics test for Chinese EFL learners. *Language Testing, 24*(3), 2007, 391-415.

[13] T. Hudson, E. Detmer, and J. D. Brown, *Developing prototypic measures of cross-cultural pragmatics* (Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa, 1995).

[14] S. O. Yamashita, *Six measures of JSL pragmatics*(Honolulu: Second Language Teaching & Curriculum Center of University of Hawaii at Manoa, 1996).

[15] R.Clément, Second language proficiency and acculturation: An investigation of the effects of language status and individual characteristics. *Journal of Language and Social Psychology, 5*(4), 1986, 271-290.

[16] T. K. Dao, D. Lee, and H. L. Chang, Acculturation level, perceived English fluency, perceived social support level, and depression among Taiwanese international students. *College Student Journal, 41*(2), 2007, 287-295.

[17] Y. Ardasheva, S. Tong, and T. R. Tretter, Validating the English Language Learner Motivation Scale: Pre-college to measure language learning motivational orientations among young ELLs. Learning and Individual Differences, 22(4), 2012, 473-483.

[18] R. M. Ryan and J. P. Connell, Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology, 57*, 1989, 749-761.

[19] L. Vandergrift, Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics, 26*, 2005, 70-89.

[20] R J. Lowinger, Z. He, M. Lin, M. Chang, The impact of academic self-efficacy, acculturation difficulties, and language abilities on procrastination behavior in Chinese international students. *College Student Journal, 48*(1), 2014, 141-152.

[21] K. A. Ericsson, andH. A. Simon, *Protocol analysis: Verbal reports as data* (Cambridge, MA: MIT Press, 1984).

[22] A. Hughes, *Testing for language teachers* (Cambridge: Cambridge University Press, 1989).

[23] T. D. Wilson, The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5*(5), 1994, 249-252.

[24] F. B. Bryant, Assessing the validity of measurement, in L. G. Grimm and P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics*(Washington, DC: American Psychological Association, 2000).

[25] M. M. Olea, M. J.Ree, Predicting pilot and navigator criteria: not much more than g. *Journal of Applied Psychology, 79*, 1994, 845-851.

[26] Y. Dodge, *The Oxford Dictionary of Statistical Terms*(Oxford: Oxford University Press, 2003).

[27] T. Z. Keith, C. R. Reynolds, Measurement and design issues in child assessment research, in C. R. Reynolds and R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (New York, NY: Guilford Press, 1990).

[28] D. Borsboom, G. J.Mellenbergh, and J. van Heerden, The concept of validity. *Psychological Review, 111*(4), 2004, 1061-1071.

[29] T. Millon, C.Millon, R. Davis, and S. Grossman, *MCMI-III Manual* (4th ed.) (Minneapolis, MN: Pearson Education, 2009).

[30] B. Williams, T. Brown, A.Onsman, Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*(3), 2012, 1-13.

[31] L. D. Goodwin, The role of factor analysis in the estimation of construct validity. *Measurement in Physical Education & Exercise Science, 3*(2), 1999, 85-100.

[32] L R. Gay, G. E. Mills, and P.Airasian,*Educational research: Competencies for analysis and applications* (Upper Saddle River, NJ: Pearson, 2006).