Journal of Language Testing & Assessment (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/langta.2025.080107
ISSN 2616-2938 Vol. 8 Num. 1

Argument-based validation frameworks in language testing: What are we expecting?

Don Yao

College of Humanities and Foreign Languages, China Jiliang University, Hangzhou, Zhejiang, China
donnieyao@cjlu.edu.cn

Keywords: Argument-based approach, validation framework, critical review, merits and demerits, implications

Abstract: The emerging context of interest in language testing lies in the arena of argument-based approaches to validation especially over the last three decades. Among which, three argument-based validation frameworks are prevalent and have been adopted by numerous researchers in their scholarly research, i.e., (Kane, 1992; Kane,2006; Kane, 2012; Chapelle et al., 2008; Bachman & Palmer, 2010). The overarching purpose of the current paper is to critically review three contemporary validation frameworks and discuss the merits and demerits of each framework. Subsequently, classical argument-based validation research is reviewed to explore the relationship between validity evidence and test development and use and provide systematic and logical implications for further research. Results showed previous argument-based validation studies touched upon limited kinds of stakeholders, and only a partial validation framework was adopted to appraise certain inferences. Additionally, empirical research mainly focused on Kane's (1992, 2006, 2012) and Chapelle et al.'s (2008) frameworks. It is to be hoped that further research could take more kinds of stakeholders into consideration, and a more systematic and comprehensive validation study is suggested. Meantime, Bachman and Palmer's (2010) framework is also advocated because of its salient feasibility and practicability.

1. Introduction

Validity is always a key component of an assessment in the area of language testing [6, 7]. The traditional concepts of validity could be classified into four types, i.e., criterion, content, construct, and consequential validity [8]. Messick's (1989) unified validity model well combines different types of validity and it has been universally acknowledged and embraced by numerous researchers. This 2 x 2 matrix model divides validity into two intertwined facets: the *justification* of an assessment with evidential bias and consequential bias, and the *function* or *outcome* of an assessment with test interpretation and test use. The model makes fully use of test scores and clearly enumerates how the scores are interpreted and used. It also appreciates the consequences a test may cause by combining both value implications and social consequences. Nevertheless, the unified validity model is rich in conceptualization, but weak in practice and implementation [10]. To address this limitation, the argument-based approach to validation has been proposed and discussed over the last three decades

(e.g., [Cronbach, 1971; Cronbach, 1988; House, 1980; Messick, 1989; Kane, 1992]). Through using an argument-based approach to validation, test interpretation and use can be clearly stated and supported by evidence, and it offsets the drawbacks of Messick's with providing limited practical guidance of implementation. The present paper then critically reviews three leading validation frameworks in the field of language assessment, i.e., Kane's (1992, 2006, 2012) interpretative argument framework, Chapelle et al.'s (2008) interpretation and use argument framework, and Bachman and Palmer's (2010) assessment use argument framework and discuss the merits and demerits of each framework. Subsequently, empirical argument-based validation research is reviewed to explore the relationship between validity evidence and test development and use and provide some systematic and logical implications for further research.

2. Contemporary Argument-based Validation Frameworks

2.1. [1, 2, 3]'s Interpretative Argument Framework

Kane (1992) outlined a systematic argument-based validation framework, i.e., the interpretative argument (IA) framework. This framework is closely associated with the score interpretation with different kinds of inferences and assumptions. Evidence found to evaluate the argument may support or against the interpretation. The framework was further developed by Kane (2006) into a two-facet argument-based validation framework, including the interpretive argument (IA) that focuses on the proposed interpretations and uses of test scores, and the validity argument (VA) that emphasizes the overall evaluation of the proposed interpretations and uses. He then developed four kinds of inferences based on test performance, interpretations, and further decisions (see Figure 1).

Scoring is test takers' performance in light of the assumptions of scoring procedures, scoring consistency, and scoring fairness. *Generalization* grounds the foundation on assumptions of the representativeness of the test scores from the observed scores to the universe scores [14]. *Extrapolation* is represented from the universe scores to target scores to examine test takers' performance in the target language use (TLU) domain. *Utilization* is mainly about the decisions a test may account for and the values or consequences a test may bring about.

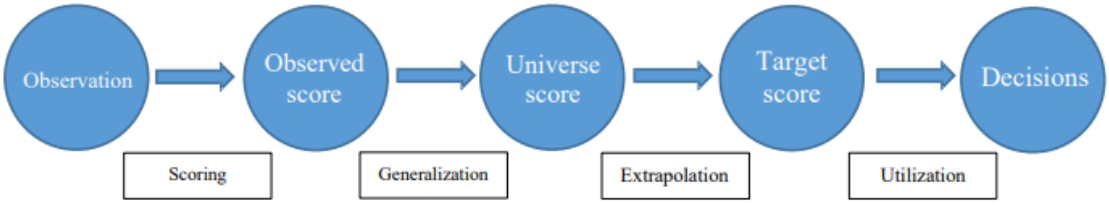

Figure 1: [2]'s Interpretative Argument Framework

Kane's [1, 2] IA framework did promote the development of language assessment concerning both score interpretation and score use, and it was in line with Messick's 2 x 2 unified validity model by taking test use and consequence into account. Kane (2012) then developed the interpretive argument into interpretation and use argument. The major revision was that the term decision was changed into decision rules and types of consequences were further deconstructed into intended outcomes, adverse impact, and systemic effect.

However, Kane's IA framework failed to provide a specific methodology to investigate and interpret the test use and consequence. Even Kane (2012) admitted that one of the fallacies of the framework was not automatic and algorithmic enough. Also, the test construct defined in the framework is solely based on test takers' observed performance in a prescribed domain, but it neglects the definition on the theoretical level [15]. Hence, the process of defining the construct is omitted in

this framework which necessitates a more comprehensive and practical framework for further research.

2.2. [4]'s Interpretation and Use Argument Framework

Remedying the imperfections of Kane's IA framework, Chapelle et al. conceptualized the interpretation and use argument (IUA) framework covering the domain description as part of construct definition (see Figure 2). They also added an explanation inference in the framework to better illustrate the relationship between test scores and test takers' performance. This framework has been universally embraced by a bunch of researchers and linguists. It offers a comprehensive and practical way for both test development and test validation starting with domain analysis, i.e., *domain description*, and being ended in test use, i.e., *utilization*, with different kinds of decisions and consequences.

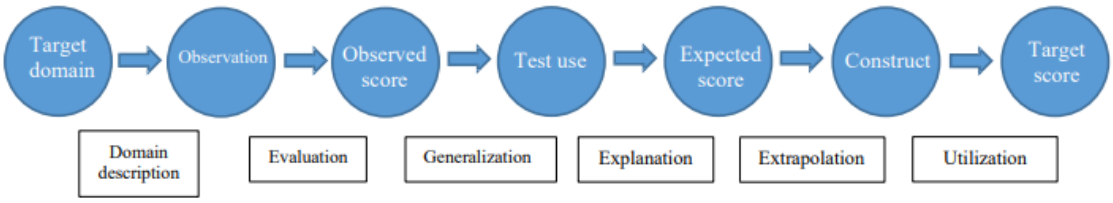

Figure 2: [4]'s Interpretation and Use Argument Framework

However, the term utilization contains both decisions and consequences. Chapelle et al.'s IUA framework did not distinguish the difference between these two components. Hence, researchers adopting this framework need to examine by themselves what kind of decisions an assessment may account for or what kind of consequences the assessment may bring about. In other words, test use is mentioned but not strongly stressed in the framework.

2.3. [5]'s Assessment Use Argument Framework

Addressing the limitations of Chapelle et al.'s IUA framework, Bachman and Palmer put forward the assessment use argument (AUA) framework laying emphasis on test use. This framework was initiated by Bachman (2005) containing two arguments, i.e., assessment *validity* argument and *utilization* argument [6, 16]. The assessment validity argument is mainly about test scores and test takers' performance, whilst the utilization argument is primarily about decision makings and consequences. Due to the imperfections of the previous framework, Bachman and Palmer then proposed the AUA framework with four specific components, containing assessment records, interpretations, decisions, and consequences (see Figure 3). This framework is rather concise and useful and well distinguishes the difference between decisions and consequences with the claims that decisions of an assessment should take into consideration existing community values and legal requirements and should be equitable for all the stakeholders [5], and consequences of an assessment should be beneficial to stakeholders [5]. It is also in accordance with and highlights Messick's value implications and social consequences. Meanwhile, the concepts of generalization and extrapolation in Chapelle et al.'s framework are both included in the interpretations inference.

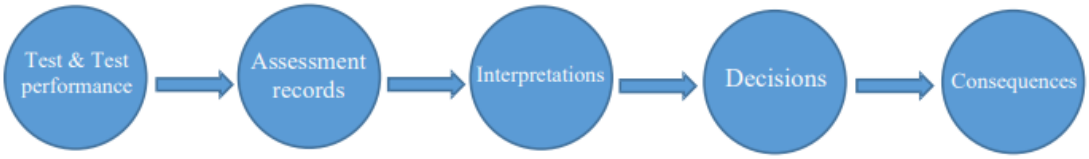

Figure 3: [5]'s Assessment Use Argument Framework

Since most of the currently used argument-based validation frameworks were derived from Toulmin's argument model [17], Bachman (2004, 2006) that the model could also be used to practical reasoning as a basis for articulating an assessment use argument [18, 19, 5]. A fundamental argument structure in Toulmin's argument model is composed of five elements: claim, warrant, rebuttal, backing, and data (see Figure 4). Referring to the evaluation of language assessment, the Toulmin's model of argument can be fully covered. The *claim* refers to the interpretations that an assessment ought to achieve, and it is a conclusion of the argument. The *warrant* is the rationale or language assessment theories that can support the claim. The *backing* can be embodied by theories or results from empirical studies or research that support the warrant. The *rebuttal*, also known as the counterclaim, is the challenge to the claim with similar supporting evidence. The *data* are usually test takers' test performance or other quantitative or qualitative data.

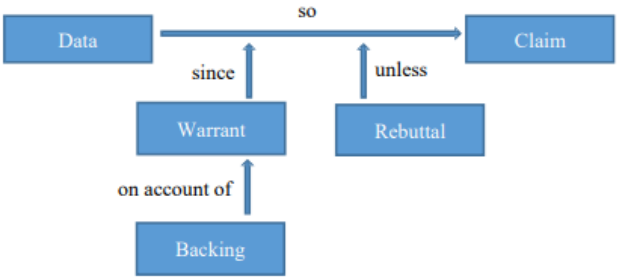

Figure 4: [17]'s Model of Argument

Regardless of the comprehensive and sufficient explanations that the AUA framework has offered, some concerns and ambiguities are still observed. To start with, some newly invented terminologies are adopted in the framework such as meaningfulness (i.e., validity) and impartiality (i.e., fairness). It may cause unnecessary confusion, especially to young researchers or novice scholars who have just stepped into the field of language assessment. Furthermore, the AUA framework provides an exhaustive list of warrants to support the claims by using a checklist approach [20, 15]. However, the checklist approach is not easy to be operationalized in test validation with various warrants for different claims. Even Bachman and Palmer acknowledged that not all the warrants or rebuttals should be necessarily required in the framework and researchers could consider the warrants or rebuttals that they need in their own research. Additionally, in terms of consequences that an assessment may bring about, the framework only mentions beneficial and detrimental consequences, but overlooks the compounded consequences [21]. Finally, the term impartiality in the interpretations inference is deconstructed into several sub-components, but the term *offensiveness* remains vague and incomplete with the statement "the assessment tasks might be offensive (topically, culturally, and linguistically inappropriate) to some test takers" ([5], p. 163). This kind of offensiveness stands for the terminology "bias" in the field of language assessment. And the perspective of demographic information (e.g., gender, age, and region) is not clearly clarified in the framework.

2.4. [22]'s Test Fairness Framework

To address the last concern in AUA framework, the concept in Kunnan's (2004) test fairness framework (TFF) was employed Kunnan (2004) [22]. Test fairness, as an essential quality of an assessment (Kunnan, 2017; Wallace, 2018; Wallace & Qin, 2021; Ho et al., 2021), refers to the equitable treatment to all the test takers based on the process of test, access, absence of bias, and test score interpretations for intended purposes [8, 23, 24, 25, 26]. In other words, to achieve fairness, an assessment ought to be fair in the testing process and selection. Also, an assessment should avoid bias in test content, language or response format. And an assessment ought to be fair in test takers' outcomes of learning and opportunity-to-learn ([OTL], [8]). The TFF, hence, encompasses five

primary qualities comprising validity, absence of bias, access, administration, and consequences (see Figure 5). This framework not only links the validity to consequences but also ensures the fairness in an assessment [27]. In terms of absence of bias, the claim in the framework refers to a test ought to avoid content or language bias and group membership bias (e.g., gender, race and ethnicity, religion, and age) and it should have a standard-setting [28]. Kunnan's introduction to the term bias in the TFF provides a strong and thorough supplement to the deficiency in the AUA framework. Therefore, a more systematic and logical argument-based validation study could be conducted [29].

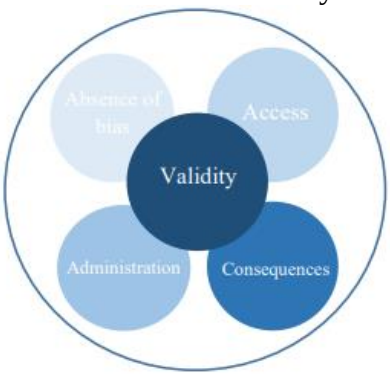

Figure 5: [22]'s Test Fairness Framework

2.5. Classical Argument-based Validation Research

The argument-based approach is one of the most functional and influential approaches to validation research in that it offers a systematic way of practicing and implementing the research and connecting validity evidence and test development and use [15]. Chapelle and Voss (2021) reviewed two major journals (*Language Testing* and *Language Assessment Quarterly*) in the field of language assessment and reported that there was a trend for scholars adopting the argument-based approaches, especially during the period of 2006 and 2011 [30]. They reviewed four papers using an argument-based approach for both high-stakes assessment [31, 32] and low-stakes assessment [33, 34]. In light of Chapelle and Voss's review, Im et al. (2019) further reviewed empirical argument-based validation studies from three more journals (*Assessing Writing*, *Language Testing in Asia*, and *Papers in Language Testing and Assessment*) and doctoral dissertations via ProQuest database published from 1992 to 2016 [15]. Altogether a total of 33 journal articles and dissertations were found. The current paper reviews six pieces of representative argument validation research in a chronological way (see Table 1).

Table 1: Classical Argument-based Validation Research

Author & Year	Test	Framework
Lim (2009) [35]	MELAB	Chapelle et al.'s IUA
Enright & Quinlan (2010) [31]	TOEFL iBT®	Chapelle et al.'s IUA
Berstein et al. (2010) [32]	TOEFL iBT®	Kane's IA
Chapelle et al. (2010) [36]	Proficiency assessment (ISU)	Chapelle et al.'s IUA
Liu (2013) [37]	CET-4	Bachman & Palmer's AUA
Tominaga (2014) [38]	OPI	Kane's IA

Lim (2009) investigated the effects of questions in the writing assessment in Michigan English Language Assessment Battery (MELAB) and raters' perceptions on test takers' responses using Chapelle et al.'s IUA framework [35]. Five inferences were used including evaluation, generalization, explanation, extrapolation, and utilization. Item difficulty and rater bias were examined through Multi-facet Rasch analysis, together with the score consistency. Also, an analysis of variance

(ANOVA) was adopted to examine the mean difference among different types of writing questions. The major conclusion was that raters and writing questions had an undue effect on score validity. Score interpretations and uses were considered valid.

Drawing on Chapelle et al.'s IUA framework, Enright and Quinlan (2010) discussed how the evidence in the evaluation of the human and machine (e-rater) scoring of an independent writing task from the TOEFL iBT® was related to four components: evaluation, generalization, extrapolation, and utilization [31]. The components were evidenced by some empirical studies between machine scoring using e-rater and human scoring in terms of reliability, generalizability, and consequences. After synthesizing all the literary works, they concluded that complementary methods of scoring were more useful for the English as a Foreign Language (EFL) writing assessment.

Bernstein et al. (2010) examined the validity of the automated spoken language tests: Versant automated tests, the speaking section of the TOEFL iBT®, and the speaking tasks within in the Pearson Test of English (PTE) [32]. Evidence was gathered from the inferences of evaluation, generalization, explanation, and extrapolation with respect to Kane's IA framework. Score accuracy (evaluation), score consistency (generalization), the correlation between scores on automated spoken language tests and other communicative tests (explanation), and TLU domain (extrapolation) were all examined as evidence to support the claim of each inference. The overall result was that the construct underlying two speaking assessments had a strong correlation and a stable relationship.

Chapelle, Chapelle et al. (2010) adopted Chapelle et al.'s IUA framework to validate test items in a computer-based proficiency assessment used by the Iowa State University (ISU) [36]. Inferences such as domain description, evaluation, generalization, explanation, and extrapolation were included in the research. Results showed that the test items matched the target second language acquisition (SLA) research domain, and it was plausible to deliver the computer-based test and scoring.

Based on Bachman and Palmer's AUA framework, Liu (2013) investigated the validity of the College English Test-Band 4 (CET-4) using a mixed-methods approach. For interpretations, quantitative approaches of statistical analyses of CET-4 scores were adopted. A total of 2,692 data points underwent descriptive analyses, correlations and factor analyses. The qualitative approach of textual analysis (i.e., content analysis) was examined to assess the construct of CET-4 and its content coverage and representativeness. As for decisions, questionnaires and interview protocols were adopted as principal instruments to investigate the impartiality of CET-4's decision makings. The document analysis was served as an essential element in the thesis to provide CET-4 syllabus for content analysis. Referring to consequences, two questionnaires (student and teacher questionnaires) and interviews were mainly adopted to examine in what way and to what extent CET-4 affects English teaching and learning practices. Descriptive statistics such as frequencies, means, standard deviations, kurtosis and skewness were calculated. Inferential statistics such as correlations and multiple regressions were examined to interpret the relationships between perceptions and test performance. Results showed the CET-4 test was acceptable in content validity. Decisions manifested a trend of large-scale tests as a catalyst for teaching and learning innovations. Test takers' perceptions of motivations, test-taking strategies had a salient influence on test performance.

Tominaga (2014) examined the scoring criteria of the Japanese Oral Proficiency Interview (OPI) based on the American Council on the Teaching of Foreign Languages (ACTFL) level descriptors using Kane's IA framework [38]. The qualitative approach, conversation analysis, was adopted to examine how test takers used turn-taking and answered tasks in the test to explore the appropriateness of the text type. The major result was that the test type criterion in OPI was not entirely aligned with ACTFL descriptors, especially for test takers who were at the low proficiency level. In other words, the text type criterion did not absolutely match test takers' actual performance, so did the ACTFL descriptors. This outcome called for further research to provide more evidence to revise the descriptors.

3. Discussion and Conclusion

Validity is always a key concept and stressed in the field of language assessment. And the validity of an assessment is usually evaluated through a validation framework that guides critical discussions or analyses [39]. The current paper reviewed the three most universally acknowledged frameworks in the field. It was found the AUA framework well justifies the test interpretation and test use and addresses the demerits of the IA framework with no specific methodologies, and the IUA framework fails to distinguish decisions and consequences. However, the lacuna remains that very little empirical research has adopted the AUA model for validation studies. Therefore, the use of the AUA framework for validation studies is advocated if the assessment lays particular emphasis on test use. Additionally, it is noteworthy that conducting the research using argument-based approaches to validation is endorsed in pluralism in the arena of language assessment [40]. However, it is inevitable to shun some limitations and corresponding implications are, therefore, presented for further research. To start with, previous research has focused on limited stakeholders such as test takers and raters (e.g., Enright & Quinlan, 2010; Lim, 2009), calling for further research to consider multiple stakeholders, e.g., test designers and test developers. Besides, former studies (e.g., Berstein et al., 2010; Chapelle et al., 2010) have mainly adopted a partial argument-based validation framework evaluating certain inferences and claims, necessitating a more comprehensive and systematic validation study. Finally, some research (e.g., Tominaga, 2014) has adopted only one method for analysis, demanding further research to use multiple methods to conduct the argument-based validation studies. This review paper, hence, sheds some light on argument-based validation research in the field of language assessment and provides some implications for further research. It is to be hoped that the further research could consider containing more kinds of stakeholders, encompassing more inferences in different frameworks, and adopting the most suitable and feasible framework.

Acknowledgements

Funding information: This paper is funded by Ministry of Education Vocational College Education Professional Teaching Steering Committee (No: 2025JGYB006).

References

[1] Kane, M. (1992). An argument-based approach to validity. Psychological Bulletin, 112(3), 527-535.
[2] Kane, M. (2006). Validation. In R. L. Brennan (4th Ed.), Educational measurement (pp.17-64). American Council on Education and Praeger.
[3] Kane, M. (2012). Articulating a validity argument. In G. Fulcher, & F. Davison, The Routledge handbook of language testing. Routledge.
[4] Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). Building a validity argument for the test of English as a foreign language. Routledge.
[5] Bachman, L. F., & Palmer, A. S. (2010). Language assessment in the real world: Developing language assessments and justifying their use. Oxford University Press.
[6] Chapelle, C. A. (1999). Validity in language assessment. Annual Review of Applied Linguistics, 19, 254-272.
[7] Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), The companion to language assessment (pp. 1079-1097). Wiley.
[8] Kunnan, A. J. (2017). Evaluating language assessments. Routledge.
[9] Messick, S. (1989). Validity. In R. L. Linn (3rd Ed.), Educational measurement (pp. 13-103). Macmillan.
[10] Knüsel, B., Baumberger, C., Zumwald, M., Bresch, D. N., & Knutti, R. (2020). Argument-based assessment of predictive uncertainty of data-driven environmental models, Environmental Modelling & Software, 134, 104754.
[11] Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (2nd Ed.), Educational measurement (pp. 443-507). American Council on Education.
[12] Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 3–17). Lawrence Erlbaum Associates, Inc.

[13] House, E. R. (1980). Evaluating with validity. Sage Publications.
[14] Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. Educational Measurement: Issues and Practice, 18(2), 5-17.
[15] Im, G. H., Shin, D., & Cheng, L. Y. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. Language Testing in Asia, 9(14), 1-26.
[16] Bachman, L. F. (2005). Building and supporting a case for test use. Language Assessment Quart, 2(1), 1-34.
[17] Toulmin, S. (2003). The uses of argument (2nd Ed.). Cambridge University Press.
[18] Bachman, L. F. (2004). Linking observations to interpretations and uses in TESOL research. TESOL Quarterly, 38(4), 723-728.
[19] Bachman, L. F. (2006, April). Linking interpretation and use in educational assessments. National Council for Measurement in Education (NCME), San Francisco, U.S.
[20] Schmidgall, J. E. (2017). Articulating and evaluating validity arguments for the TOEIC tests. ETS Res Report.
[21] Yao, D., & Wallace, M. P. (2021). Language assessment for immigration: A review of validation research over the last two decades. Frontiers in Psychology, 12, 773132.
[22] Kunnan, A. J. (2004). Test fairness. In M. Milanovic, & C. Weir (Eds.), Europe language testing in a global context: Selected papers from the ALTE conference in Barcelona (pp.27-48). Cambridge University Press.
[23] Wallace, M. P. (2018). Fairness and justice in L2 classroom assessment: Perceptions from test takers. The Journal of Asia TEFL, 15(4), 1051-1064.
[24] Wallace, M. P., & Qin, Y. (2021). Language classroom assessment fairness: Perceptions from students. LEARN Journal: Language Education and Acquisition Research Network, 14(1), 492-521.
[25] Ho, A. O. K., Yao, D., & Kunnan, A. J. (2021). An analysis of Macau's Joint Admission Examination-English. Journal of Asia TEFL, 18(1), 208-222.
[26] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. AERA.
[27] Kunnan, A. J. (2010). Test fairness and Toulmin's structure. Language Testing, 27(2), 183-189.
[28] Yao, D., & Chen, K. (2020). Gender-related differential item functioning analysis on an ESL test. Journal of Language Testing & Assessment, 3, 5-19.
[29] Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. Front Psychol, 11(330), 1-14.
[30] Chapelle, C. A., & Voss, E. (2021). Validity argument in language testing: Case studies of validation research. Cambridge University Press.
[31] Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. Language Testing, 27(3), 317-334.
[32] Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. Language Testing, 27(3), 355-377.
[33] Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. Language Assessment Quarterly, 7(2), 137-159.
[34] Koizumi, R., Saka, H., Ido, T., Ota, H., Hayama, M., Sato, M., & Nemoto, A. (2011). Development and validation of a diagnostic grammar test for Japanese learners of English. Language Assessment Quart, 8(1), 53-72.
[35] Lim, G. S. (2009). Prompt and rater effects in second language writing performance assessment. Deep Blue.
[36] Chapelle, C. A., Chung, Y. R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. Language Testing, 27, 443-469.
[37] Liu, M. (2013). Applying an assessment use argument to investigate a college-level English test in universities in Xi'an. PolyU Electronic Theses.
[38] Tominaga, W. (2014). Validating the score inference of the Japanese OPI ratings: The use of extended turns, connective expressions, and discourse organizations. Semantic Scholar.
[39] Pochon-Berger, E., & Lenz, P. (2014). Language requirements and language testing for immigration and integration purposes. Report of the Research Center on Multilingualism, 2-40.
[40] Chapelle, C. A. (2020). Argument-based validation in testing and assessment. Sage Publications.