Argument-Based Validity in Classroom and Program Contexts: Applications and Considerations

Justin Cubilo
cubiloju@hawaii.edu
University of Hawaii at Manoa

Central to determining the quality of any measure of learner ability is the determination of whether such measures provide a valid assessment of the abilities under question. The notion of what validity is and how to assess the validity of a given measure has undergone several changes over the past half century. Early conceptualizations of validity focused on the notions of criterion, content, and construct validity as more or less separate models. However, it has been recognized that criterion and content validity, while useful, are limited in what they can provide as supporting evidence for establishing validity since when they are used individually they only address a smaller portion of what needs to be considered for assessing the validity of a measure. This led some theorists such as Loevinger (1957) to suggest that criterion and content validities were simply parts of validation which fell under the umbrella of construct validation. Based on this view of validation, Messick (1989) proposed a unified model of validity, which included empirical methods for construct validation and consequences for test interpretation and use. At this time, Messick (p.13) defined validity as:

> An integrated evaluative judgment 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. [italics in original]

Thus, with his definition Messick removed the test itself from being the focus of validation and instead placed the focus on the score interpretation and use. This would ideally be accomplished through the construction of a logic-based validity argument by gathering the necessary evidence for and against the proposed interpretation or use of the test score and the inferences that are associated with these interpretations. Kane (2006) outlines such an argument-based approach, which is described below.

An Argument-Based Approach to Validity

According to Kane (2006), validation consists of two types of arguments, an interpretive argument and a validity argument. The interpretive argument is built upon a number of inferences and assumptions that are meant to justify score interpretation and use whereas the validity argument evaluates the interpretive argument in terms of how reasonable and coherent it is as well as how plausible the assumptions are (Cronbach, 1988). Development of such arguments requires the use of a clear structure on which the argument may be based. For this reason, those who work on developing interpretive and validity arguments (Kane, 2001; Mislevy, Steinberg, & Almond, 2003) base their arguments on Toulmin's (1958, 2003) framework for creating informal arguments, which essentially requires that a chain of reasoning be established that is able to build a case towards a final conclusion, which in this case would be to determine the plausibility and reasonableness of score interpretations and uses.

As is shown in Figure 1, Toulmin's (2003) argument structure is built on several components, which include the grounds, claim, warrant, backing, and rebuttal. As it relates to test score interpretation and use, the claim of an argument is the conclusion one draws about an individual based on test performance

whereas the grounds serve as the data or observations upon which the claim is based upon. For example, one may make the claim that an individual learning English has inadequate listening comprehension abilities for studying at an English medium university based on the grounds that they received a low score on a multiple-choice listening comprehension test consisting of a series of lectures utilizing academic vocabulary and structures. However, the inference linking the grounds to the claim is not given and therefore justification is needed in the form of a warrant (or assumption). The warrant in Toulmin's model is considered to be a rule, principle, or inference-license that is meant to provide justification for the inference connecting the grounds to the claim. Warrants in turn need backing (or evidence) which comes in the form of theories, research, data, and experience.

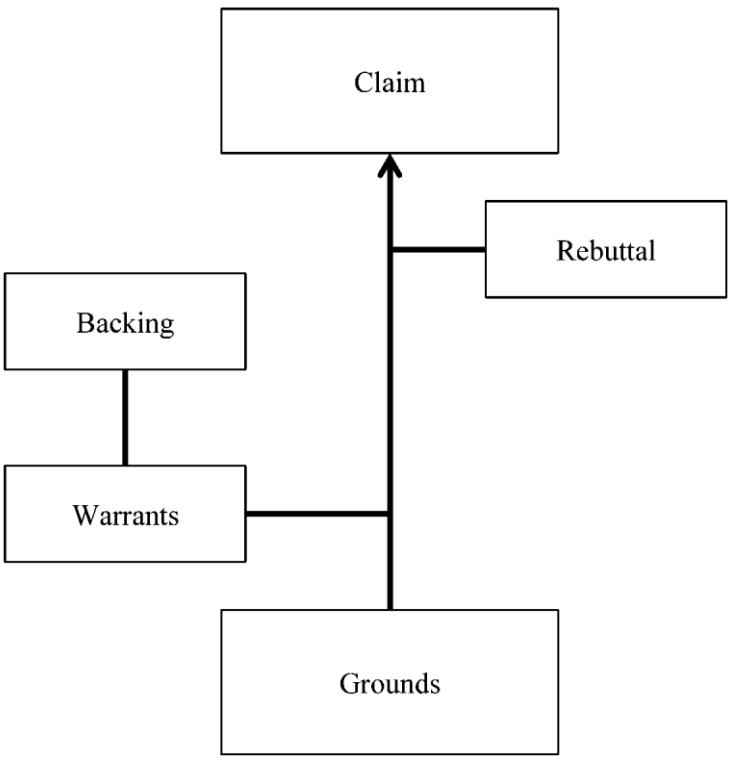

Figure 1. Model of Toulmin's (1958; 2003) argument structure.

In relation to the example provided above, the warrant justifying the inference between the grounds and the claim would be that performance on the listening comprehension tasks reflect relevant and necessary language abilities needed in an academic context. This warrant would then be supported by backing that might say that individuals with low-level listening ability generally have difficulty understanding academic words, making inferences or predictions from what a speaker has said, or poor knowledge of signal words and phrases meant to hint at main ideas or important points and that such deficiencies lead to poor performance in an academic English-speaking context. Finally, while warrants and backing justify the inferential link between the grounds and claim, rebuttal data can serve to weaken the initial argument by providing evidence or possible explanation which may call into question the warrant. Going back to the previous example, a possible rebuttal may be that several of the topics presented in the lectures may have been too technical or abstract, the vocabulary may have consisted primarily of less commonly or

frequently used academic vocabulary, or even that the audio quality was poor. Such data would serve to weaken the inference connecting the grounds and claim and would either have to be investigated further or accepted by the test developer with the knowledge that it places a limit on the argument. Thus, these components are all connected with each other and are essential for establishing an inferential connection between the claims and grounds.

In order to establish a connection between the claims and grounds, Kane (1992) stated that multiple inferences of different types must be used in a chain to connect observations and conclusions. Therefore, Kane, Crooks, and Cohen (1999) developed a three-bridge model for the three types of inferential bridges they thought were essential for linking arguments together in order to move from observation (i.e., the grounds) to score interpretation (i.e., the claim). Each inference is in turn based on a series of assumptions, each of which requires support. These three inferences were identified as evaluation, generalization, and extrapolation inferences. The evaluation inference refers to the score that is assigned to an individual's performance on a measure with the underlying assumption that appropriate criteria are used to score the performance, that they have been applied as planned, and that the conditions under which the performance took place match the intended score interpretation (Kane, 2002b; Kane, 2013; Kane et al., 1999). Following the evaluation inference, the generalization inference refers to the use of an observed score as a way of estimating future performance or scores of a test taker if given parallel tasks or test forms. Finally, following generalization is the extrapolation inference that refers to predictions of how the expected score is to be interpreted as an indication of performance and scores that the individual would receive in the target domain. An important assumption of extrapolation is that test tasks are authentic relative to tasks test takers would be expected to perform in the target domain.

In applying the bridge model to language testing, Chapelle, Enright, and Jamieson (2008) describe three further inferences in their validity argument for the TOEFL iBT that can be used to strengthen the connection between the grounds and claim and these are labeled as the explanation, domain description, and utilization inferences. The explanation inference describes the relationship between the observed test performance and a theoretical construct (e.g., a construct of second language listening). The domain description inference refers to a detailed description of the target domain and is meant to provide a link between performances in the target domain and observed performance on the test. Finally, the utilization inference provides the link between the target score that has been obtained for the test taker and the decisions that will be made about the test taker in relation to policy. Taken together, these six inferences along with their assumptions and support, which is obtained through a variety of methods, are able to provide a chain of arguments that can support the link between the grounds and claims of the overall validity argument. The types of evidence that can be collected as support for the assumptions in each of these inferences is manifold.

Applying the Argument-Based Validity Framework

Kane's framework (and its expansion by Chapelle) is a useful tool for considering the interpretations and uses of test scores. However, since it is slightly abstract in nature, it can be difficult for instructors and administrators to fully realize how to apply it to in their specific situations. In essence, to fully comprehend how this framework can be utilized within a particular situation, being able to see how evidence can be acquired to provide support for the different inferences within the model is necessary. This will ensure that teachers' and administrators' score interpretations and uses can be fully supported in their particular contexts. Below I discuss how teachers can gather evidence for some of the more relevant inferences for the classroom context.

One of the first things that instructors or administrators must do in creating a valid test for their classroom or for placement purposes is to adequately define the domain that they are attempting to assess. In order

to accomplish this, instructors and administrators can do several things. The first think that should be examined are the curricular and course objectives and student learning outcomes. These should be used to guide discussion regarding the content of the assessment and for determining the appropriate question formats for adequately assessing these outcomes and objectives. For instance, if students are expected to display appropriate pragmatic knowledge in making refusals at a certain level in a program or by the end of a course, a test should include a component that is meant to assess this ability and a role play or some other speaking activity may be more appropriate for assessing such knowledge rather than a multiple choice test. Additionally, it is important that the underlying trait, or construct, is appropriately defined so that educators can be absolutely sure that they are assessing what they wish to assess. Having a clearly defined construct will essentially provide stakeholders with a clear idea of what characteristics should be incorporated into an assessment meant to measure a given skill area.

Once constructs and outcomes and objectives have been clearly defined, teachers and administrators can proceed to develop appropriate tasks that are meant to target these aspects. This type of consideration is placed under the evaluation inference within an argument-based validity framework. This is essentially the time where a teacher can pilot the items of the test to gather evidence related to how they are working and to see how administrative conditions are affecting performance (Enright et al., 2008). In this way, teachers can revise their items or tasks by examining item facility and b-index or item discrimination values to see how items differentiate between learners on certain objectives. Furthermore, test adminstration characteristics can be investigated at this time to determine if such characteristics significantly affect performance in a positive or negative way. Examples of such characteristics would be to examine how the presence or absence of extra planning time for responses on speaking or writing tests affect overall performance or whether notetaking on a listening test significantly helps or hinders performance. This would also be an excellent time to get feedback from students who will be taking the test as they can tell you which of the administration conditions they prefer and how they relate to their affective state as this is something that could affect the overall relation between test performance and the construct that has been defined.

The generalization and explanation inferences come next in the argument-based framework, and it is here that some statistical evidence is required. Generalization in essence refers to the reliability of the assessment and whether the student would perform comparably well on future administrations of similar tests. This can be determined by calculating split-half reliability or the K-R20 or K-R21 values (where the test has been administered only once), by investigating test-retest reliability where a test is given twice and the results are correlated with each other, or by parallel forms reliability in which the test is correlated with another equivalent form targeting similar material (for more on the calculation of these coefficients, see Brown, 2005). Such information will provide the test designer with information on the amount of construct-irrelevant error that is present within the test so that they can determine how to proceed (often by either increasing the number of items found on a test or ensuring that test items are not ambiguous). Furthermore, teachers who are using tests as measures to determine who has mastered and who has not mastered content can evaluate the consistency of such determinations by using test dependability measures. One possibility for calculated this dependability index is to calculate the phi (lambda) coefficient which will provide the developer with information related to the dependability of a given cut score, taking into account the fact that some people pass a test to a greater extent than others. A discussion of how to calculate the actual coefficient is beyond the scope of this paper, but the reader is directed to Brown (2005) for his discussion of this topic.

Beyond the generalization inference, the explanation inference provides evidence to show that performance on a test is in line with the construct previously designed by the test developer. For instance, if the test that a teacher is developing is meant to assess achievement in meeting learning outcomes in an intermediate language skills classroom, the teacher can assess whether the test does indeed do this by

having beginning, intermediate, and advanced learners take the test in order to investigate differential item functioning. If the test and items function in accordance with what would be expected in relation to learning outcomes and objectives (i.e., intermediate students scoring significantly higher than beginning students and advanced students showing greater mastery of the outcomes and objectives than both intermediate and advanced groups) and the construct, then the teacher or administrator would have evidence to support the explanation inference. Furthermore, if the school has access to other measures of a similar skill (e.g., listening, speaking, writing, etc.) that they can have their students take, they can take results from these measures and correlate them with the measure they are developing in order to assess the test's convergent validity. This is effectively the correlation between two measures of the same or similar construct that use different methods (e.g., multiple choice and short answer questions or direct and semi-direct speaking assessments) (Crocker & Algina, 2008). Having a high correlation would show that a similar construct is being assessed and would lend credence to the support of the explanation inference. For a school or program environment, these two types of evidence would be good starting points for providing sufficient evidence for the explanation inference.

The final portion of constructing the validity argument requires providing support for inferences that focus on connecting test performance to performance and effects of score use outside of the actual test. The extrapolation inference is the first of these and can be supported through correlation studies (Kane, 2013). Whereas correlations in the explanation inference are used to provide evidence for the relationship between scores and the construct, correlations in the extrapolation inference are used to make connections to performance in the target domain and these correlations can be done with similar measures. They can also be correlated with course performance to ensure that there is a strong relationship between test performance and performance in the language or content classroom, which would indicate good fit for the test in relation to learning outcomes (which, by extension, would ideally be related to performance on real world tasks). For instance, if a teacher of English academic listening were seeking to obtain extrapolation for their test, they might correlate performance on their test with performance in lecture-style content courses and that performance in these content courses differs with each level of listening ability based on how many learning outcomes have been mastered by the student as displayed by the test score.

The utilization inference is the final inference in the argument-based validity framework and is often the inference that is addressed after a test has been developed and administered. This inference requires support for moving score interpretation to score use and requires examination of the consequences of the test and its effects on policy (Kane, 2002a; 2013). Bachman & Palmer (2010) outline a number of factors that are important in relation to score use in the decision-making process. Specifically, they mention that the consequences of the test should be beneficial for all stakeholders, reports should be clearly presented and easily interpretable, and the test has positive washback on instructional practice and learning. The utilization inference rests upon the assumptions that the consequences and decision-making process have been investigated in order to ensure that decisions and consequences equitable, scores are interpretable, and that instruction is positively affected by test use.

Teachers and administrators can do a number of things to gather evidence for this inference. First, washback studies can be conducted in which teaching is observed and learning is assessed. In this way it is possible to see whether course objectives are being targeted appropriately within the classroom and whether student learning as assessed by the new test is focusing on appropriate content and how this is related to topics covered in the classroom. Furthermore, stakeholder input from students and other teachers who may use the test would serve to be valuable in ensuring that the test is perceived as fitting with instruction and course objectives and that it is perceived as adequately assessing student mastery of specific learning outcomes. This type of feedback will serve to make for better score interpretations related to performance in the target domain. Finally, further investigations can be conducted in order to assess cutoff score determinations in order to make sure that such scores are appropriate for making decisions of

mastery versus non-mastery. This is especially important when achievement tests are used to determine if individuals have adequately learned the material to advance to a higher level in the language program. For a discussion of methods related to determining cut scores, the reader is directed to Brown (2005) and Fulcher (2010). All of this evidence will provide a clear and easy-to-follow blueprint for instructors to use so that they remember how to use their tests appropriately and how to interpret the scores.

Conclusion

Taken together, the evidence from each of the inferences mentioned above is put together into a single validity argument. The purpose of the validity argument is to determine whether the evidence that has been collected for each of the inferences is appropriate and actually supports the interpretations and uses of the tests either within a single classroom or program-wide. In order to condense the information, Table 1 summarizes the key points and sources of evidence for each inference.

Table 1
Summary of Inferences and Evidence Sources

Inference	Purpose	Evidence
Construct & Domain Definition	Describing and understanding the target domain (context) and skill to be measured to support intended interpretations	Literature Analysis Content Analysis (Examining program and class objectives and student learning outcomes)
Evaluation	Scores of observed performances are examined as measures of performance in the L2 ability. Highly relevant for determining score meaning	Item Analysis (Item Facility, B-Index, Item Discrimination) Stakeholder Opinions Examining effects of Administrative Conditions
Generalization	Ensuring that observed scores are consistent over future, parallel task versions so that expected scores can be estimated	Reliability Statistics (K-R20, K-R21, Split-Half Reliability, Phi (lambda))
Explanation	Determining that scores are related to the defined construct in a way that aligns with theory	Differential Item Functioning Studies Differential Group Studies Convergent validity
Extrapolation	Extension to performance outside of the test within the target domain	Correlation to performance in the target domain
Utilization	Moves from score interpretation to score use. Considers impact of test in relation to decision-making policies and curricular adaptation.	Stakeholder feedback related to perceptions of score interpretations Washback studies Cut Score Examination

It is recommended that inferences be addressed in the order that they are placed in the table as they will help to focus evidence for later inferences. While it is not always possible to gather evidence for all of these inferences within a given context, it is preferable to do so as this will only serve to provide stronger support for intended score interpretations and uses, which is what test developers, in both major testing companies and in classrooms, should be striving to do.

Bibliography

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). New York, NY: Routledge.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R. N., Mollaun, P., Nissan, S., . . . Schedl, M. (2008). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 97-143). New York, NY: Routledge.

Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. T. (2002a). Practice-based standard setting. *The Bar Examiner, 71*, 14-24.

Kane, M. T. (2002b). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 18*, 5-17.

Kane, M. T. (2006) *Validation*. In R. Brennan (Ed.), *Educational Measurement*, 4th ed. (pp. 17-64), Westport, CT: American Council on Education and Praeger.

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

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3*, 635-694.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, 3rd ed. (pp. 13-103). New York: Macmillan.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge, UK: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument: Updated edition.* Cambridge, UK: Cambridge University Press.