CONSTRUCT VALIDITY IN FORMATIVE ASSESSMENT: PURPOSE AND PRACTICES

By

SAMANTHA RIX

University of Northern Iowa.

ABSTRACT

This paper examines the utilization of construct validity in formative assessment for classroom-based purposes. Construct validity pertains to the notion that interpretations are made by educators who analyze test scores during formative assessment. The purpose of this paper is to note the challenges that educators face when interpreting these scores and states that how the scores are used, instead of the actual score itself, is integral for conveying the two parts of the General Validity Theory; actions and structures. The paper continues by intertwining the idea that validation can eventuate only if evidence is being supported by different sources. Finally, the paper concludes by providing examples of how educators can prevent misinterpretation and deter bias whilst using the validity argument when conducting formative assessments.

Keywords: Construct Validity, Formative Assessment, Assessment of Formative Testing Scores, Validity, Classroom Assessment, Teacher Judgments.

INTRODUCTION

Teacher judgments have always been the foundation for assessing the quality of student work. More often than not, teachers use their own personal knowledge of the subject area, their understanding and feelings toward the student, and the years of experience in teaching to assess students' academic capabilities. Today, teaching methodology has taken a leap toward integration, by which the fields of psychology, education, and the sciences have coalesced to better fit the needs and strengths of all learners. Technology in the 21st century has also taken a leap, and with it the availability of standardised testing. While standardised testing can be a convenient tool for teachers as it is less time consuming, it doesn't always fit the needs of students. For teachers, standardised testing can also have an ambiguous element for, once the test has been given, many teachers are hesitant as how to interpret these results. This paper will delve into the idea of teacher hesitancy and the judgments they bring with them about the test and their students. As well as looking specifically at formative assessment, a common testing form in many schools, construct validity can be elucidated to help educators interpret the results.

Validity

The origin of the word validity comes from Latin, validus, meaning strong. Validity requires strong evidence in order to be able to support a claim or an idea. For the purpose of this paper, validity will be linked with the idea that well-founded evidence will be interpreted through test scores. Kane (2001) stated that “the process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations (p. 328, as cited by the 1999 Standards for Educational and Psychological Testing). Kane (1992) stressed that these scores also need to have a level of appropriateness, wherein the process itself and the conclusions of the analysis correlate with other factors, such as student variables and teacher judgments.

Consequently, the General Validity Theory implies that the actual interpretation of the test scores and how it is utilised is more important than the actual grade or numerical value the test displayed (Nichols, Meyers & Burling, 2009). The framework of this theory has two parts: actions and structures (Nichols, Meyers & Burling, 2009). The action part conveys the instructional practices of the teacher, for example, how lesson plans are presented and how subsequent testing should cover “a student's current understanding of the subject matter” as presented by the teacher (Nichols, Meyers & Burling, 2009, p. 15). The structures portion implies that the material is taught in a manner that is conducive for student learning (Nichols, Meyers & Burling, 2009). For example, teaching should cover different learning styles of students, such as visual, auditory, spacial, analytic, and those that learn in “fits” and those that seem to learn in more of a “steady, stream-like” manner.

Construct Validity

One form of validity that is relevant for teachers in order to assess formative testing scores is construct validity. Bachman & Palmer (1996) expressed that “construct validity pertains to the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores” (p. 20). Therefore, it is important to justify the reasoning behind why a certain method or system was chosen, not just surmise that it works. A construct is a specific ability that students need to master, for example, the conditional tense in English. So construct validity needs to occur when a particular construct is assessed by the instructor who decides what the interpretation of that score from the assessment actually entails.

Formative Assessment

Formative assessment is one form of testing that can be used to help students achieve success with the constructs they are learning. O'Malley & Valdez Pierce (1996) stated that formative assessment is an “ongoing diagnostic assessment providing information to guide instruction” (p. 238). Formative assessment can also be a way to show the gap between actual performance and desired levels of fluency. Llosa (2011) constructively expressed that if teachers can effectively assess students using some form of validity, it could ultimately lead to success in closing the gap between the students' interlanguage and actual fluency. Nichols, Meyers & Burling (2009) affirmed that formative assessment “is an implied claim of validity” (p. 14). If teachers can evaluate formative assessment by applying a form of validity correctly, the teachers could modify their own teaching methods and practices to help students succeed at language learning tasks in ways that pique curiosity and interest in the material (Llosa, 2011).

Shepard (2009) noted an important caveat to formative assessment: “just because it's labelled formative assessment doesn't make it so” (p.33). In order to be considered formative, the assessment must be able to examine the whole test along with specific components, be a form on on-going analysis, and be used in a manner that could be beneficial for student development (Shepard, 2009).

A study done by Llosa (2007) gave evidence as to how teachers and administrators could implement the idea of construct validity in formative assessment by examining two standardised tests in California, the ELD classroom assessment, which is a “standards-based classroom assessment of English proficiency used in a large urban school district in California,” against the California English Language Development Test, the CELDT (p. 493). Both of these tests measure English language proficiency in the three categories of reading, writing, and listening (Llosa, 2007). Districts in California were chosen based upon whether the county had 50% or more of its students considered to be English language learners (ELLs). The results showed that the test was very accurate for grades two through four, which displays who the test is best for in order to make assessments and to set goals (Llosa, 2007). There were some concerns though, such as the results for grade two came back almost a year later, which is not helpful for teachers to manipulate their curriculum in order to meet the specific needs of the students in their classrooms. Overall, though, the validity between the two types of tests, the formative ELD proficiency test, and the summative CELDT, showed relative consistency in scoring results. Llosa (2007) concluded by stating that is an “important piece in the validity argument for the use of classroom-based assessment of students' language proficiency” (p. 512).

How Validation Eventuates

In order for teachers and administrators to be sure that construct validity is occurring with their formative assessment forms, they need to be sure that “…the integration of different kinds of evidence is arriving from different sources” (Llosa, 2008, p. 33). The validity argument stated that there must be links “from the test performance to a test score to an interpretation (Llosa, 2008, p. 34). The interpretation can then be used to ascertain that the test is examining the constructs that are being taught.

Llosa (2008) gave some practical steps for teachers to build their own validity argument. I) it is important to make a "claim" (Llosa, 2008, p. 34). The teacher needs to have genuine knowledge of what the students know and are able to perform. ii) gather the "data," and then the teacher or administrator needs to collect the scores from the assessment and write down the manner in which they presented the assessment (Llosa, 2008, p.34). iii) decide upon a "warrant;" the person(s) who are assessing the data needs to justify their reasoning in interpreting the data (Llosa, 2008, p. 34). iv) make sure there is "backing" (Llosa, 2008, p. 34).

Ample evidence along with possible refutations are important in order to clarify the data interpretation. Llosa (2008) declared that "backing can be obtained from the test design and development process and from evidence collected as part of research studies and the validation process" (p. 34).

Ergo, the process of construct validation in formative assessment is a mixture of the interpretation of scores based on an argument and the evidence behind it (Kane, 1992). Kane (2001) explained that validity "aims for a coherent analysis of all the evidence for and against the proposed interpretation and, to the extent possible, the evidence relevant to plausible alternate interpretations" (p. 329). This can also include teacher and administrator judgments and assumptions about both the students and the assessment. Nichols, Meyers & Burling (2009) concurred with this analysis of how construct validation should be utilised, and that the interpretation can help or hinder student development and achievement with future language tasks.

What the Interpretations Involve

Once teachers and administrators have steps to validate their formative assessments and have come to conclusions about how the scores from the test can assist students in their own language learning development, they must understand what those interpretations involve. First noted by Messick in the late 1980s, Kane (1992) stated that these interpretations involve "meaning or explanation" or how the score is explained and the implications of the test itself (p.527). Kane (2001) further noted that the interpretations will involve "an extended analysis of inferences and assumptions and will involve both a rationale for the proposed interpretation and a consideration of possible competing interpretations" (p. 328). Also, the culmination of the data gathering and the inferences can be interpreted with the caveat that the interpretation can be skewed by variables such as student behaviour or the test not examining the constructs learned in class. For example, Kane (1992) found one formative reading test that based its score on "the ability to comprehend a variety of written materials in a variety of contexts, even though the test may consist of discrete, multiple-choice items administered in one testing session" (p. 529). Therefore, it is highly salient that when explanations are made, these variables are made clear so they do not distort and lead to misinterpretations of the data.

Importance of Validation

As stated earlier, validation helps guide educators and administrators in how to incorporate and interpret the scores that their students receive. Bachman (2000) proposed that validation is crucial in order to "analyse critically the ways in which tests are used, to assure that such uses are fair and ethical" (p. 17). For example, an exam that does not test what is being taught and then gives unfair grades to students provides washback that is not advantageous for future linguistic development due to the lack of correlation between classroom objectives and analysis of the material.

Another issue that can occur is bias stemming from cultural or dialectal issues (Beswick, Willms & Sloat, 2005; Llosa, 2007). When construct validity is used, many presumptions are adopted that may or not be correct, no matter how much empirical data is applied and bias for or against certain students can occur. For example, students who speak Vernacular Black English (VBE) are often considered illiterate or uneducated because of the manner in which they speak. Also, students who speak other languages at home and have cultures that are vastly different from the person who is interpreting the results may be perceived by the assessor as impolite or brusque. Beswick, Willms & Sloat (2005) continued by giving an example where gender played a crucial possible bias. A study conducted in Connecticut found that four times as many boys as girls were considered to have reading difficulties in a particular school, but when the boys were assessed on an individual level, the number of girls and boys who were considered to have any type of reading difficulty came out equal.

Another study by Beswick, Willms & Sloat (2005) involved comparing teachers rating of 205 kindergarteners' emergent literacy skills in nine different schools with a series of validity tests. The study found that there was a large gap between the teacher's ratings and the students' actual performance on a standardised test "with prior evidence of construct validity (Beswick, Willms & Sloat, 2005, p. 130). The results stated that the teachers were much more critical of the ability of these students in comparison "derived from direct assessment" (Beswick, Willms & Sloat, 2005, p.130). They believed the discrepancy lied in characteristics such as "child, family, and behavioural factors" (Beswick, Willms &Sloat, 2005, p. 130). More specifically, categories such as "students [who] are repeating kindergarten are male, have mothers with low education, and exhibit behavioural difficulties in the classroom" were found to be more negatively biased against (Beswick, Willms & Sloat, 2005, p. 130). While some of these factors could indeed hinder reading development for many of these students, it is highly unlikely that it is the case for all of the students.

Finally, variability in teacher's interpretations can lead to misdiagnosis of strengths and weaknesses of students' language abilities (Llosa, 2007). Llosa (2011) noticed when she conducted a study in California that some teachers would merely "tick the boxes" by using the standard assessment given to them by the school, while others would rely much more on professional experience (p. 372). She also noted that there was quite a range in those who had been trained to assess formative language tests and those who had not, and also stated that those who had received any form of formal training, did so "several years ago" (Llosa, 2011, p. 372).

Many of these teachers, especially novice ones, exclaimed that they often went with the standard form given to them because they were unsure of what the definition of progress meant for emergent bi and multilinguals in their classroom (Llosa, 2008). Therefore, the teachers own background can affect on their judgement and analysis of the data (Llosa, 2007; Llosa, 2011). Brindley (2001) stated that this was not an uncommon concern and that "…unless greater attention is given to providing adequate time allocation and appropriate forms of professional development, the many potential benefits of involving teachers in assessment will not be realised" (p. 403).

Preventing Misinterpretation

There are many ways to deter bias and other issues that may come across when validating formative assessments for interpretation and analysis. Advances in internet related technologies (IRT) have made it feasible to individualise tests for students' needs (Bachman, 2000). Bachman (2000) also stated that individualised tests can be made en masse if they are salient enough. For example, the TOEFL, an English language exam that many have to take in order pass into regular academic classes at the university level, could be individualised so that the test reflects better how English was learnt. Bachman (2000) calls this idea "adaptive language testing" (p. 9).

As Brindley (2001) mentioned above, professional development days is another method that could actively aid teachers in how to use construct validity when examining formative assessment. Focus questions could be ones such as what is progress in language development in our school district, how do we implement changes in our classroom to better serve our students to pass the formative evaluations given to them, and how do we write a rubric to help both our novice and more advanced teachers take these ideas into their own classrooms?

Llosa (2011) addressed this issue by stating that "...less attention has been paid to the role of standards-based assessment in the classroom" and therefore conducted a study to see how a standards-based, formative assessment in an urban school in California correlated with the SBCA, a standards based national assessment used in California (p. 368). The formative assessment in the school had levels one through four with relatively specific criterian in each level and specified that students can only pass up to the next level, of which there were five, if they received a three or a four on the exam (Llosa, 2011). The most important conclusion of the study was that teachers "...did not interpret the standards consistently and, as a result, the extent to which a student was determined to master a standard was largely based on a particular teacher's interpretation of that standard (Llosa, 2011, p. 370). Both the issues of how to implement construct validity in the classroom and the findings by Llosa (2011) could be resolved by executing Brindley's (2001) idea of putting into action training of these topics during teacher in-service days.

Bachman (2000) proposed the idea of having professionally trained personnel come into the schools to do the testing and interpretation. This could decrease the possibility of personal bias toward the students who took the exam. These professionals could also train the teachers and administrators how to use validity when constructing hypothesis of language development in their students (Bachman, 2000).

Finally, age is an important factor to take into consideration when delivering a type of formative assessment. For example, a study done by Nichols, Meyers & Burling (2009) found that for children under the age of eight, any type of formal assessment would not be appropriate and that individualised testing would need to be done. This is an important admonition for student variables in language assessment, such as dialectical and cultural differences. In some instances, official formative assessment may need to be tailored in order for teachers to interpret the needs of the student.

Conclusion

Construct validity in formative assessment is a key way to help educators, administrators, and assessors use the data they have collected from scores to prompt them to more accurate forms of interpretation and washback. Assessing linguistic capabilities can be extremely difficult for teachers due to concerns such as what is considered progress for my student(s), how can I evaluate my student without bias, and how do we, as teachers, match up our assessments to make formative decisions for our students? The opportunity to incorporate some of these issues into staff development days, having professional people come in and do training, and to keep in mind the backgrounds of students are all ways that could aid those who do assessments and base interpretations off of these assessments. Finally, remember that all students are unique, and with the technologies of today and the integrated approach of teaching, any type of formative assessment might need to be tailored to help students grow and develop their own language repertoire.

References

[1]. **Bachman, L. (2000).** Modern language testing at the turn of the century: Assuring that what we count counts. *Language testing, 17(1), 1-42.*

[2]. **Bachman, L. & Palmer, A. (1996).** *Language testing in practice.* Oxford: Oxford University Press.

[3]. **Beswick, J., Willms, J. & Sloat, E. (2005).** *Education, 126(1), 116-137.*

[4]. **Brindley, G. (2001).** Outcomes-based assessment in practice: some examples and emerging insights. *Language testing, 18(4), 393-407.*

[5]. **Kane, M. (1992).** An argument-based approach to validity. *Psychological Bulletin, 112(3), 527-535.*

[6]. **Kane, M. (2001).** Current concerns in validity theory. *Journal of Educational Measurement 38(4), 319-342.*

[7]. **Llosa, L. (2007).** Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language testing, 24(4), 489-515.*

[8]. **Llosa, L. (2008).** Building and supporting a validity argument for a standards-based classroom assessment of English Proficiency Based on Teacher Judgments. *Educational Measurement: Issues and Practice.*

[9]. **Llosa, L. (2011).** Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language testing, 28(3), 367-382.*

[10]. **Nichols, P., Meyers, J. & Burling, K. (2009).** A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice.*

[11]. **O'Malley, J. & Valdez Pierce, L. (1996).** *Authentic assessment for English language learners.* United States of America: Addison-Wesley Publish Company.

[12]. **Shepard, L. (2009).** Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice.*

ABOUT THE AUTHOR

Samantha Rix is currently a graduate assistant at the University of Northern Iowa studying teaching english to students of other languages (TESOL). She focuses her studies on second language acquisition, and specifically on how learner traits and cognitive factors can facilitate or hinder a students' ability to learn a second or third language.