Considerations of Construct Validity in Language Testing Context

Suttinee Chuanchaisit[1]

Business English Department, School of Humanities,

University of the Thai Chamber of Commerce, Bangkok, Thailand

Abstract

This paper aims to describe clear concept of construct validity and its relevance, focusing on language testing context. The major contribution is to clarify the characteristics of construct validity and explain it in a straightforward way leading to the derivation of an acceptable and workable conceptual scheme with practical implications. Messick's (1989) theory of test validity, particularly the unified view of construct validity, is profoundly influential in part because it brings together disparate contributions into a unified framework for building validity arguments. Two major discussions were provided, the first being more straightforward and including a discourse on several ways of thinking about the ideas of validity, and notions of construct validity. Here, construct validity is seen as the overarching quality with all of the other measurement validity labels falling beneath it. The second, an elaboration on the expansion of the unitary concept of construct validity. Finally a synthesis of ideas is presented through a conceptual framework demonstrating the similarity between construct validation procedures taking place in operational and in a language testing contexts.

Keywords: Construct Validity; Validity; Language Testing

[1] Email: drsuttinee@gmail.com

บทคัดย่อ

บทความนี้นำเสนอสาระสำคัญเกี่ยวกับความตรงตามโครงสร้าง (Construct validity) ในบริบทของการวัดและประเมินผลทางภาษาโดยมีวัตถุประสงค์หลักเพื่อให้ความกระจ่างเรื่องลักษณะเฉพาะของความตรงตามโครงสร้าง โดยอธิบายให้เห็นภาพรวมของทฤษฎีความตรง (Theory of Validity) และจะนำเสนอต่อในเชิงลึกเรื่องความตรงตามโครงสร้าง ซึ่งจะนำไปสู่ความเข้าใจและการเรียบเรียงกรอบความคิดพร้อมทั้งการนำไปประยุกต์ใช้

ทฤษฎีความตรงในการวัดและประเมินผลทางภาษาของเมสสิค (Messick, 1989) มีส่วนสำคัญในการเรียบเรียงบทความนี้ โดยเฉพาะแนวคิดที่มองความตรงตามโครงสร้างเป็นแบบองค์รวม(Unified view of construct validity) ซึ่งเป็นการนำเอาองค์ประกอบที่แตกต่างกันโดยสิ้นเชิงมารวมกันเป็นกรอบการทำงานแบบรวมเป็นหน่วยเดียว

สาระสำคัญที่นำเสนอมีสองประการ ประการแรกเพื่อนำเสนอทฤษฎีความตรง (Theory of Validity) ในแบบตรงไปตรงมา รวมถึงการอภิปรายมุมมองที่หลากหลายของแนวคิดเรื่องความตรง (Validity) และแนวคิดเกี่ยวกับความตรงตามโครงสร้าง (Construct validity)

ประการที่สอง เนื่องจากทฤษฎีความตรงตามโครงสร้างนั้นมีความสำคัญยิ่งยวด มีส่วนเกี่ยวข้องครอบคลุมความตรงประเภทอื่นๆซึ่งอยู่ภายใต้ความตรงตามโครงสร้าง จึงนำแนวคิดดังกล่าวมาขยายให้เห็นภาพอภิปรายโดยมองความตรงตามโครงสร้างเป็นแนวคิดแบบหน่วยเดียว(The expanding of the unitary concept of construct validity) จากนั้นนำเอาแนวคิดที่อภิปรายข้างต้นมานำเสนอเป็นกรอบความคิดที่แสดงความคล้ายคลึงกันระหว่างกระบวนการตรวจสอบความตรงตามโครงสร้าง (Construct validation) ที่เกิดขึ้นในการนำไปปฏิบัติ (Operations) และในบริบทของการทดสอบทางภาษา (Language testing context)

คำสำคัญ: ความตรงตามโครงสร้าง ความตรง การทดสอบทางภาษา

Considerations of Construct Validity in Language Testing Context

Suttinee Chuanchaisit
Business English Department, School of Humanities,

University of the Thai Chamber of Commerce, Bangkok, Thailand

Introduction

The most important and fundamental characteristic of any measurement procedure is validity. The term validity refers to whether or not the test measures what it claims to measure. We find that on a test with high degree of validity, the items will be closely connected to the test's intended focus. There are several ways to estimate the validity of a test including content-, construct-, and criterion-related validity.

Bachman and Palmer (1996) emphasized that several people seem to look only at test outcomes and ignore the test process itself. Evidence of test usefulness should be both qualitative and quantitative. McNamara (1996, p.7) supports that "the validity of second language performance assessments involves more than content- and criterion- related aspects of validity; the larger issue of construct validity has been insufficiently considered".

Although the notion of validity has evolved over the years, the concept of construct validity has been particularly confusing to many, being claimed to be the most difficult form of validity to understand (Alderson et al., 2001). Importantly, if we find it difficult to define construct validity, we will find it even more difficult to measure.

Therefore, the major contribution of this article is to provide a brief overview of what validity is, as background for

readers who are not keen on this issue. The researcher then discusses the concept of construct validity focusing on what construct validity really accomplishes in both traditional and unified views. Then, to reflect construct validity as a whole validity, the researcher attempts to make some sense of the unified concept of validity to demonstrate a new model of construct validity: "One for All-All for One: Expanding the unitary concept of construct validity", and also to illustrate this with examples from language testing context. Finally, this leads to a proposed theoretical unified framework of construct validity, to provide the logical framework both for guiding test design or research conduct and for developing a substantive grounded set of procedures for the use of a particular assessment.

What is validity?

In order to provide readers who lack solid language testing background, it is important to briefly review what validity is.

Since the processes of measurement are diverse and tend to be complicated, it is not surprising that some believe that the concept of validity must also be complicated. Basically, however, there are just two main perspectives on validity: the traditional and the unified views. The details of each are outlined below.

In the **traditional view**, it is evident that early validity theory held multiple lines of thought. Validity standards which were first codified in 1954 indicated to the test users the degree to which the test was capable of achieving certain aims (APA, 1954). In other words, tests of validity then aimed to answer the question: "Does this test measure what it is supposed to measure?" (Kaplan & Saccuzzo, 2005, p. 134). Four types of validity were identified corresponding to different aims of testing. Cronbach (1984) argued that these should be four

different methods of inquiry rather than types of validity. Thus, it might be emphasized that they are not distinct categories and not tactual types of validity, but instead different approaches to the establishment of evidence for validity. They approaches are presented as follows:

Content validity indicates how well the test represents the subject matter content and behaviors selected (Hatch & Farhardy, 1982). Its evidence can be established by evaluating whether test items are a good sample of the conceptual domain that the test is designed to cover (content representation) and whether the test items relate to the content domain (content relevance). *Predictive validity* is called for when a test is used to predict future performance and necessitates collecting criterion data later than the test. *Construct validity* is needed when making inferences about unseen traits such as intelligence or anxiety. Validation of construct tends to provide an answer to the question "What does this test really measure?" (Bachman, 1990, p. 256). In order to establish construct validity evidence, researchers have to simultaneously define some constructs and develop the instruments to measure them. Shohamy (1994, p. 120) suggested that evidence should be collected from multiple perspectives. *Concurrent validity*, a separate type of validity involving an external criterion, is more appropriate when a new test is proposed as a substitute for a less convenient measure that is already accepted (e.g., a multiple-choice history test in place of a difficult-to-score essay examination). Concurrent validity data might also serve as a shortcut approximation of longitudinal predictive data.

The unified view of validity has been generally accepted, reflecting most closely the notion of construct validity (Cronbach, 1984). Messick (1995) argued that the traditional view of validity is fragmented and incomplete. He proposed the

notion of unified validity, the unifying force of which tends to be the meaningfulness and trustworthy interpretability of test scores and their implications, i.e. construct validity. It can be seen as a multi-faceted construct that sought out multiple evidence sources.

As a unitary concept, Messick (1989) 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 on other modes of assessment" (p. 13). His view convinces us to accept that validity is not the judgment of a test itself, but the properties of the test being judged.

This unified view has been endorsed by the measurement profession as a whole in Standards for Educational and Psychological Testing (Bachman, 1990, p. 236). The unitary view implies the following:

(1) Test validation requires empirical support.

(2) Theories that define what the test measures (its construct) are needed to make a valid test.

(3) Appropriate interpretation of the scores depends on the test construct.

Content and criterion-related validity are significant, but they are the elements of the construct validity that "embrace the entire evidence basis for test (score) interpretation; including content and criterion-related lines of validity evidence" (McNamara, 1996). (For more details, refer to the section on expanding the unitary concept of construct validity).

In discussing language test validity at this point in time, it may be said that its conceptualization is generally accepted in either the traditional or the unified view. As Brown (2000) claimed, either the traditional view or unified view is held by

virtually all psychometricians inside or outside of language testing. However, validity in the view presented here is conceptualized as the overarching quality of construct validity with all other measurement validity labels falling beneath it. The details are provided in the next section.

What exactly is construct validity?

This section covers several areas of thought regarding construct validity. Both traditional and unified views of construct validity are raised to be the underlying issue. A foundation in the richness of this idea may provide the reader with an overview of what construct validity really constructs. However, it would be easier to understand the concept of construct validity by first understanding what a construct is. A construct is an attribute, proficiency, ability or skill that is processed in our brain and is defined by established theories. It is something that exists in theory and has been observed to exist in practice such as students' overall English language proficiency.

Traditional view of construct validity

Traditionally, construct validity has been defined as the experimental demonstration that determines "whether a test measures what it supposes to measure". Such an experiment could take the form of a differential-group study, wherein the performances on the test are compared between two groups: the one that has the construct and the other that does not have the construct. If the result is that the group with the construct performs better than the group without the construct, it provides evidence of construct validity of the test. However, under the auspices of this traditional validity, constructs tend to be validated only through the analyses of external measurements. It has been argued that if the question of "whether a test measures what it supposes to measure" is a question of the meaning of test

scores, the internal process underlying test scores should be taken into consideration. This idea has led to the unified view of construct validity proposed by Messick (1989).

Historically, during early 1950s, the first formal articulation of the concept of construct validity came from the idea of the nomological network. Cronbach and Meehl (1955) elaborated the model of theory testing by developing the concept of "nomological net." The construct to be measured was located in a conceptual space showing its hypothesized connections to other constructs and observed behaviors. These theoretical relationships were then tested empirically through correlational and experimental studies. However, there was a weak point. It did not provide practicing researcher with a way to actually establish whether or not their measures had construct validity.

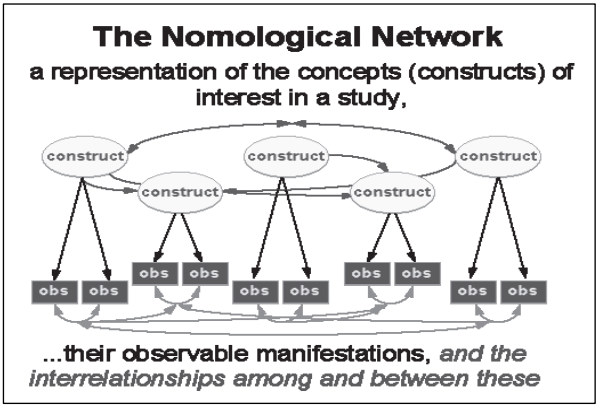

Figure 1. Nomological network (Cronbach&Meehl, 1955, p.290).

Figure 1 above represents the nomological network proposed by Cronbach and Meehl (1955). What Cronbach and Meehl (1955) were trying to do was to link the conceptual/theoretical realm with the observable one, as this is the central concern of construct validity. However, while the

nomological network idea may work as a philosophical foundation for construct validity, it does not provide a practical and usable methodology for actually assessing construct validity.

In 1959, the development of the Multi-trait Multi-method Matrix (MTMM) took on this task by emphasizing a methodological approach to construct validity (Campbell & Fiske, 1959). In order to claim that the measure had construct validity under the MTMM approach, the researcher had to demonstrate that both *convergent* and *discriminant* validity existed in the measure. Convergent validity was demonstrated when the researcher showed that the measure, which theoretically is supposed to be highly interrelated, was indeed highly interrelated (in practice). And the discriminant validity was presented when the researcher demonstrated that the measures that should not be related to each other were in fact not related.

There is one thing that the nomological network and the MTMM ideas have in common and is an underlining theme in both, the idea of "pattern". As a theoretical pattern, the researcher does not necessarily have a theory on how the programs and measures related to each other. As an observed pattern, however, the researcher provides evidence through observation that the programs or measures actually behave that way in reality. Therefore, if we claim construct validity, we are claiming that our theoretical pattern corresponds with our observed pattern (how we think the world works, matches with how things operate in reality).

However, another broader view of language testing has emerged. In educational measurement circles, all three types of validity (content, criterion-related, and construct validity) are taken to be different facets of a single unified form of construct

validity. This unified view of construct validity is considered a new development by many of the language testers around the world. This view is clarified in next part.

Unified view of construct validity

The unified concept of construct validity first started with a study by Loevinger (1957) which claimed that construct validity reflected the whole of validity from a scientific point of view. Later, the idea was expanded in Messick's studies (1975, 1980, 1989) which presented a unified and expanded theory of validity supporting the idea that construct validity embraces and subsumes all other forms of validity, namely content validity and concurrent validity. They are considered to be sub-parts of construct validity. Simply speaking, the view was that if the identified observations could define the construct, content representativeness and content relatedness would then be a prerequisite to construct validity.

Regarding *aspects of construct validity*, to speak of validity as a unified concept does not imply that validity cannot be usefully differentiated into distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked (Messick, 1989). In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are the contextual, substantive, structural, generalizable, external, and consequential aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989b). Following a description of these six aspects, some of the validity issues and sources of evidence bearing on each are highlighted below:

(1) The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989b);

(2) The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks;

(3) The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957; Messick 1989b);

(4) The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test criterion relationships (Hunter, Schmidt, & Jackson, 1982);

(5) The external aspect includes convergent and discriminating evidence from multi-trait multi-method (note: write as *multi-trait multi-method*) comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & GIeser, 1965);

(6) The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989b).

There researcher hereby claims that construct validity can describe all the internal processes underlying test scores thus better serving the broader measurement community, as

constructs exist across all assessment contexts. Thus, this concept might be viewed as the construct validity-centered theory.

It might be concluded then that either the traditional view or the unified view of construct validity is held by virtually all psychometricians inside and outside of language testing. Therefore, construct validity can be said to be well-accepted, one way or the other. Keep in mind that construct validity does not refer to the question of whether or not the test really measures intelligence, but only to the question of how well certain score interpretations are supported by the evidence, focusing more on test use.

One for All – All for One: Expanding the unitary concept of construct validity

Validity becomes a unified concept, and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely, construct validity.

The unified view of validity seems to be the integrated evaluative judgments of the degree to which empirical evidence and theoretical rationales support inferences and actions based on test use and score interpretation (Messick, 1989). This inspired the researcher to view the unitary concept of construct validity as "One for All – All for One" in order to reflect construct validity as a whole validity theory.

What is proposed in this study is a model, Figure 2, asserting that construct validity includes the integration of two major validities: *evidence-based validity* (the relevance of the test to the particular applied purpose) and *consequence-based validity* (the utility of the test scores in the applied setting), as major components of construct validity. This originates from the

thought that empirical evidence itself is not strong enough to code whether the measurement sensitive to variation is a targeted attribute. The judgment of the consequences of score interpretation and test use is also needed for consideration along with the empirical evidence. Details of the components of construct validity beginning with the evidence-based validity followed by consequence-based validity including their subordinate terms of validity are shown and described as follows:

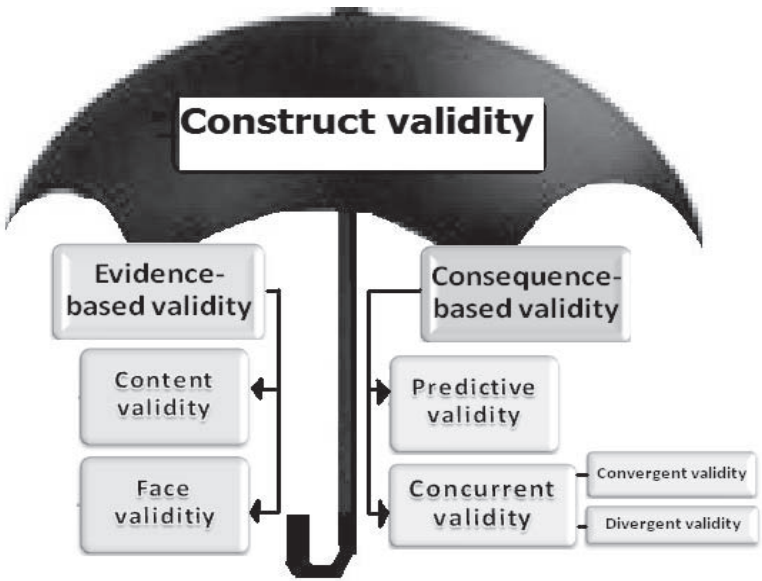

Figure 2. "One for All, (note: use a comma "One for All, All for One") All for One": Model of construct validity in an expanded and unified view [modified from Cronbach and Meehl (1955) and Messick (1989)].

Evidence-based validity

Evidence-based validity emphasizes whether the measurement is a good reflection of the construct. Construct validity comprises the evidence and rationale supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables. Therefore, evidence-based validity focuses on using tests involving the empirical investigation of content validity and relevance/utility, since content validity is seen as one of the domains of constructs in a unified concept of validity. There are two subordinate terms of validity of the evidence based validity.

Content validity

As a part of evidence-based validity, content validity refers to specific evidence of relevance and utility of the test that supports general construct validity evidence. Lennon (1956) and Messick (1989) include evidence of content relevance and representativeness as well as of technical quality. This may include any documented evidence that targets content where processes and competencies are actually present in test items.

Content validity plays an important role in language testing context. Both the content relevance and representativeness of assessment tasks are traditionally appraised by expert professional judgment, documentation of which serves to address the content aspect of construct validity. In other words, it involves gathering the judgment of reliable experts who simply analyze the test by comparing the content with a coinciding statement of the content domain--what the content ought to be. In essence, the content domain refers to test objectives/ specifications, while the sample refers to test items. The experts

will be able to review the items and comment on whether or not the items cover a representative sample of the behavior domain.

Face validity

Face validity is very closely related to content validity. It can provide evidence of representativeness for construct validity in a less formal way. While content validity depends on a theoretical basis for assuming if a test is assessing all domains of a certain criterion, face validity relates to whether a test appears to be a good measure or not. It refers to tests' "surface credibility or public acceptability" (Ingram, 1977, p. 18).

In the field of language testing, face validity essentially involves an intuitive judgment by people who are not necessarily experts, focusing on the degree to which items in a measurement instrument appear, for their face value, to measure the single construct that they intend to measure. Anyone, who might be administrators, or test-takers themselves, could look over the test, and might be able to develop an informal opinion on whether or not the test is measuring what it is supposed to measure. The process would be conducted at the same period as the evaluation of content validity.

With regard to the relationship between content validity and face validity, there are fundamentally two groups of thoughts. The first group of scholars sees face validity as different and separate from content validity (e.g. DeVellis, 1991; Kerlinger, 1973). The other (e.g. Carmines & Zeller, 1979; Nunnally, 1967) considers face validity and content validity to be two sides of the same coin, essentially viewing the measurement of a face validity assessment instrument as an indirect approach to the assessment of content validity.

Consequence-based validity

Besides evidence-based validity, consequence-based validity plays an important role to construct validity. Consequence-based validity tends to cover all of the consequences of a test relating to test score interpretation, including such considerations as accuracy in measuring intended criteria, the impact on test scores, and the social consequences of a test's interpretation and use (Messick, 1989; Gronlund, 1998; McNamara, 2000; Brindley, 2001; Brown, 2004). It includes evidence and rationales for evaluating the intended and unintended outcomes that result from using an assessment in a particular way to support specific interpretation (Messick, 1995). Unintended consequences only affect judgments about validity to the extent that they can be traced back to a source of invalidity in the test. Consequence-based validity involves making judgments of *value implications* (defined as the contexts of implied relationships to good/bad score interpretations) and of the *social consequences* (defined as the value contexts of implied consequences of test use and the tangible effects of actually applying the test).

As Messick (1989) emphasizes, for a fully unified view of validity, it should be recognized that the appropriateness, meaningfulness, and usefulness of test score-based inferences depends also on the social consequences of the testing. Hence, social values and social consequences should not be ignored in considerations of validity. These lead to the substantive aspects of consequence-based validity, which are predictive validity and concurrent validity.

Predictive validity

Predictive validation is most common with proficiency testing, tests which are intended to predict how well a person

will perform in the future (Anderson et al, 1995). Predictive validity shares similarities with concurrent validity in that both are generally measured as correlations between a test and some criterion measure. The results of predictive validity can be seen as a factor of consequence-based validity since it evaluates outcomes that are derived from using an assessment in a particular way.

When it comes to assessing predictive validity, the process involves establishing that the scores from a measurement procedure (e.g., a test or survey) make accurate predictions about the construct they represent. Examples of such constructs may include intelligence, achievement and depression. Such predictions must be made in accordance with theory; that is, theories should tell us how scores from a measurement procedure predict the construct in question. In order to be able to test for predictive validity, the new measurement procedure must be taken after the well-established measurement procedure. By after, we are referring to a period of time typically ranging from a few weeks to a few months or even years between the two measurements.

With regards to language testing context, there were a number of predictive validity studies attempting to analyze the relationship between various English proficiency test results and academic outcomes (e.g. Criper & Davies, 1988; Wall et al, 1994). However, the findings were mixed. Graham (1987) provided several reasons as to why the relationship between these two variables is problematic. First, the issue of the exact nature of language proficiency is still continually debated. Second, it relates to the difficulties of testing language proficiency. Third, there are a number of moderating variables affecting test-takers' academic performance. The nature of

relationship between all the variables is complex and thus not easy to determine.

Concurrent validity

Concurrent validity is similar to predictive validity in the sense that we assess both the concurrent validity and predictive validity of a measurement procedure when two different measurement procedures are carried out. However it is different in that the two procedures are assessed at the same time for concurrent validation. Concurrent validity is established when the scores from a new measurement procedure are directly related to the scores from a well-established measurement procedure for "the same construct"; that is, there is consistent relationship between the scores from the two measurement procedures. This gives us confidence that the two measurement procedures are measuring the same thing.

In order to demonstrate the validity of this type, it is important to show that it correlates highly with indices of the TLU that one might theoretically expect it to correlate with, and also that it does not correlate significantly with variables that one would not expect it to correlate with (Bachman, 1990, p.250, Campbell & Fiske, 1959). This statement implies that there might be two modes of comparison: their similarities and their differences. Therefore, the following two sub-types of concurrent validity could be involved: *Convergent validity* which refers to the degree to which the measurement is similar to (converges on) other measurements that it theoretically should be similar to and, in contrast, *discriminant validity*, which refers to the degree to which the measurement is not similar to (diverges from) other measurements that it theoretically should not be similar to. For example, to demonstrate the discriminant validity of a test of language ability, one might correlate the scores on the

language test with scores on tests of arithmetic skills, where low correlations would be evidence of this validity.

Both convergent and discriminant evidence are basic to construct validation. Of special importance among these external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the assessment scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose. For example, in the context of language testing, the scores must differentiate individuals in the same way on both measurement procedures; that is, a test-taker that gets a high score on one test (i.e., the well-established measurement procedure) should also get a high score on the new measurement procedure. This should be mirrored for test-takers who get a medium and low score, meaning the relationship between the scores should be consistent. If the relationship is inconsistent or weak, the new measurement procedure does not demonstrate concurrent validity.

A proposed unified framework of construct validity

Messick's (1989) point of view is that all types of validity are unified and can be seen as construct validity. This concept along with ideas gathered from other mentioned theorists, namely, Messick (1989); Cronbach & Meehl (1955); Trochim (2006); Brown (2004); and Bachman & Palmer (2000) have greatly influenced the researcher's perspective of construct validity and have led to the conceptualization of a model demonstrating the similarity between construct validation

procedures that take place in operations and in a language testing context (see Figure 3).

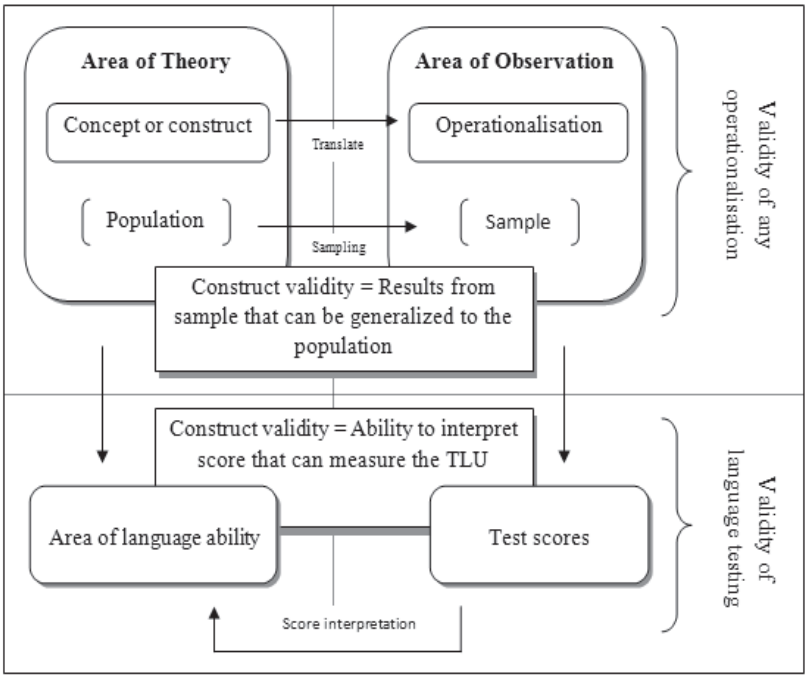

Figure 3. A proposed unified framework of construct validity.

Figure 3 presents the concept of construct validity proposed. This concept is applied in both general operations, such as in the research areas, and specific contexts (such as language testing). Both contexts share the core concept of construct validity. It can be seen that there are two paths in the model above; the left side shows theoretical relations (area of theory) and the right side illustrates empirical relations (area of observation). If the two paths match, then it can be said that construct validity is obtained.

Construct validity is then perceived to be a process of sample validation in the research methodology. Concept or construct in the area of theory acts as population, while the operationalization in the area of observation plays the role of sample. Therefore, construct validation might be equivalent to the process of sampling validation which is done in order to test whether samples can be generalized to the population. Construct validity is the agreement between the theoretical concept/ constructs and the operationalization in the measurement.

More specifically, in a language testing context, a test can be considered valid for a construct if the empirical relations between test scores match the theoretical relations between constructs. By the same logic, the measured TLU is the population, and the ability of test scores to reflect the TLU is the sample. So, construct validity refers to the ability to interpret scores that can measure the TLU.

Conclusion

This paper concentrates on the concept of construct validity in the context of language testing. The concept of validity in general was outlined and Messick's (1989) unified concept of validity presented. Then, notions of construct validity and its traditional and unified views were demonstrated and discussed. Messick (1989) considers construct validity as the one unifying and overarching framework for conceptualizing validity evaluations. Logical analysis of test content and empirical confirmation of hypothesized relationships are both essential to defending the validity of test interpretations, however, neither is sufficiently alone.

Therefore, in order to reflect construct validity as a whole validity, next section, some senses of the unified concept of validity were demonstrated as a proposed model of construct

validity: "One for All, All for One: Expanding the unitary concept of construct validity, see Figure 2.

The unified validity framework meeting the requirements distinguished two interconnected facets of validity as a unitary concept (Messick, 1989ab). One facet is evidence-based validity which refers to justification of the testing based on appraisal of evidence supportive of score meaning, while the other is consequence-based validity which refers to the outcome of testing in either interpreted or applied use. These two facets are equivalent in effecting the degree of validity.

Finally, a proposed unified framework of construct validity was presented wherein different points of view were gathered to construct a conceptualized model demonstrating the similarity between construct validation procedures taking place in operational and language testing contexts. In operationalization, construct validation might be equivalent to the process of sampling validation which is done in order to test whether or not samples can be generalized to the population. In contrast, in the language testing context, construct validity refers to the ability to interpret scores that can measure the TLU.

Biodata

Suttinee Chuanchaisit is a lecturer from the School of Humanities at the University of the Thai Chamber of Commerce in Bangkok, Thailand, also currently takes a position of the Director of UTCC Language Center. She earned a doctorate in English as an International Language (International Program), with a concentration in Language Assessment and Evaluation, from Chulalongkorn University. Her areas of interest include Language Testing, Test-taking Strategies, Learning Strategies, and Factors affected Language Learning Performance.

She can be reached via e-mail at drsuttinee@gmail.com or +66(0)2697-6416. The address is The University of the Thai Chamber of Commerce, 126/1 Vibhavadi Rangsit Road, Din Daeng, Bangkok 10400, Thailand.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for education and psychological testing*. Washington, DC: American Educational Research Association.

Alderson, J. C. & Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (2001). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Allen, M. J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press.

Anastasi, A. (1982). *Psychological Testing*. New York: Macmillan Publishing Company.

Anastasi, A. (1988). *Psychological Testing*. New York, NY: MacMillan Publishing Company.

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

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

Bachman, L. F., & Palmer, A. (2000). *Language testing in practice*. Oxford: Oxford University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Brindley, G. (2001). Language assessment and professional development. In

C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin. (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies*. (pp. 126-136). Cambridge: Cambridge University Press.

Brown, J. D. (2000). What is construct validity? *JALT Testing and Evaluation SIG Newsletter, 4*(2), 7-70.

Brown, J. D. (2004). Performance assessment: Existing literature and directions for research. *Second Language Studies*, *22*(2), 91-139.

Campbell, D.T., & Fiske, D.W. (1959).Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carmines, E. G. & Richard A. Z. (1979). *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: how can we go straight? In W.B. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement over a decade*. (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.), New York, NY: Harper Row.

Cronbach, L. J., & Gieser, G. C. (1965). *Psychological tests and personnel decisions*. (2nd ed.). Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

DeVellis, R.F. (1991). Scale Development: Theory and applications. *Applied Social Research Methods Series, 26*. Newbury Park: Sage.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Gronlund, N. E. (1998). *Assessment of student achievement*. (6th ed.). Boston, MA: Allyn & Bacon.

Hatch, E. & Farhady, H. (1982). *Research Designs and Statistics for Applied Linguistics*. California: Newbury House Publisher.

Hunter, J. E., Schmidt, F. L., &Jackson, G.B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage Publications.

Ingram, E. (1977). Basic Concepts in Testing. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods* (pp. 195-216). Oxford: Oxford University Press.

Johnson, E. J. (2001). Digitizing consumer research. *Journal of Consumer Research, 28*(2), 331-336.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praege.

Kaplan, R.M., & Saccuzzo, D.P. (2005). *Psychological Testing: Principles, applications, and issues*. Belmont, CA: Thomson Wadsworth.

Kerlinger, F. (1973). *Foundations of behavioral research*. New York, NY: Holt, Reinhart & Winston.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3* (Monograph Supplement 9), 635–694.

Lyman, H.B. (1963). *Test Scores and What They Mean*. NJ: Prentice-Hall.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T.F. (1997). Interaction in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466.

McNamara, T.F. (2000). *Language Testing*. Oxford: Oxford University Press.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-46). Hillsdale, NJ: Lawrence Erlbaum.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective*. Englewood Cliffs, NJ: Prentice Hall.

Shepard, L.A. (1993). Evaluating Test Validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 19*. Washington, DC: AERA.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 1*(2), 99-123.

Society for Industrial and Organizational Psychology, Inc. (SIOP). (1987). *Principles for the validation and use of personnel selection procedures*. (3rd ed.). College Park, MD: Author.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

Trochim, W. M. (2006). *Levels of Measurement*. Retrieved from http://www.socialresearchmethods.net/kb/measlevl.php.

Weir, C. J. (2005). *Language testing and validation: An evidenced-based approach*. Basingstoke: Palgrave Macmillan.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84*, 608-618.