A Turn toward Specifying Validity Criteria in the Measurement of Technological Pedagogical Content Knowledge (TPACK)

Robert F. Cavanagh
Curtin University

Matthew J. Koehler
Michigan State University

Abstract

The impetus for this paper stems from a concern about directions and progress in the measurement of the Technological Pedagogical Content Knowledge (TPACK) framework for effective technology integration. In this paper, we develop the rationale for using a seven-criterion lens, based upon contemporary validity theory, for critiquing empirical investigations and measurements using the TPACK framework. This proposed seven-criterion lens may help researchers map out measurement principles and techniques that ensure reliable and valid measurement in TPACK research. Our critique of existing TPACK research using these criteria as a frame suggests several areas of theorizing and practice that are likely impeding the press for measurement. First are contradictions and confusion about the epistemology of TPACK. Second is the lack of clarity about the purpose of TPACK measurement. Third is the choice and use of measurement models and techniques. This article illustrates these limitations with examples from current TPACK and measurement-based research and discusses directions and guidelines for further research.
(Keywords: Technological Pedagogical Content Knowledge framework, TPACK, reliability, validity, measurement, assessment)

Since initial publication in 2006 by Mishra and Koehler, the Technological Pedagogical Content Knowledge (TPACK) framework for effective technology integration (see Figure 1, p. 130), has had a significant impact on research and practice around educational technology (Koehler, Shin, & Mishra, 2011). Application of the framework by researchers and practitioners to inform design of interventions such as professional development has led to the development of measures to quantify effects and potential gains (Graham, Cox, & Velasquez, 2009; Guzey & Roehrig, 2009). Although this empirical imperative is a powerful rationale for developing measures, measurement is also often viewed as the optimal means of establishing the validity of theoretical frameworks and models. The validation of the frame-

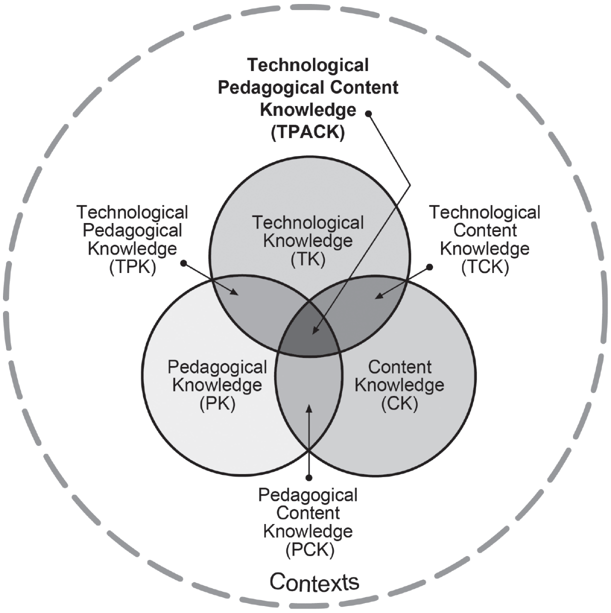


Figure 1. The TPACK framework (reproduced with permission from http://tpack.org)

work as a model of technology integration is a second driver of the proliferation of TPACK measures.

The growth in both the number and variety of the TPACK measures being explored warrants a critical look at the quality and validity of the measures being used (Koehler, Shin, & Mishra, 2011). In the sections that follow, we examine these issues through the lens of contemporary validity theory and then propose a multistep approach for examining validity in empirical investigations of TPACK.

This work is grounded in the construct of validity advanced by Messick (1995). According to Messick (1995, p. 741), validity "is an overall judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores and other modes of assessment." Messick (1998) was emphatic about this approach being unified in contrast to the multiple-type conception that previously prevailed. He also reframed these types of validity as forms of evidence and stated:

> What is singular in the unified theory is the kind of validity: All validity is of one kind, namely, construct validity. Other so-called separate types of validity—whether labeled content validity, criterion-related validity, consequential validity, or whatever—cannot stand alone in validity arguments. Rather, these so-called validity types refer to complementary forms of evidence to be integrated into an overall judgment of construct validity. (p. 37)

Table 1. Validity Evidence Criteria

Types of Evidence	Description	Examples of Application
1. Content evidence	The relationship between the instrument's content and what the instrument seeks to measure	Specification of research questions, development of a construct model, writing of items, selection of a scaling model
2. Substantive evidence	Explanation of observed consistencies in the data by reference to a priori theory or hypotheses	Comparing TPACK scores of teachers who have completed TPACK training with those who have not
3. Structural evidence	Confirmation of subconstructs or components in the construct model	Conducting Confirmatory Factor Analysis
4. Generalizability evidence	Individual items are not biased toward particular groups or situations	Testing that each item in a test of TPACK elicits similar responses from males and females with the same overall TPACK level
5. External evidence	Similar results are obtained when different tests are applied to measure the same construct	Comparing findings from observational schedules and document analysis
6. Consequential evidence	Consideration of how results could impact on persons and organizations	Discussing findings with stakeholders
7. Interpretability evidence	Communication of the qualitative meaning of scores	Providing a construct map that explains key points on the scale

The current version of the Standards for Educational and Psychological Testing published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) embody this unified conception: "Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (AERA, APA, & NCME, 1999, p. 11). The evidence requires documentation of all aspects of instrument development and administration, from initial theorizing to assessing the consequences of interpreting the results.

Messick (1995) provided a six-criterion framework for the organization of evidence. The criteria were the content, substantive, structural, generalizability, external, and consequential aspects. Wolfe and Smith (2007a) added an additional aspect from the Medical Outcomes Trust Scientific Advisory Committee (1995), evidence of the interpretability aspect. Application of the seven-criterion framework has not been restricted to psychometric test development. Significantly, it has been used in the assaying of phenomenological research that used rating scales, surveys, and observational instruments (Cavanagh, 2011a; Young & Cavanagh, 2011).

The seven criteria are introduced in Table 1, along with examples of how each can be applied. In this paper, we employ these criteria to audit TPACK-based empirical research and the measures employed in this research. The following sections explain the seven aspects of validity evidence in more detail and how these could be manifest in TPACK measurement.

The corpora of reports on TPACK measurement used in the study were identified by a literature search in conjunction with the second author's extensive familiarity with TPACK literature. The theoretical model applied in the study was the Wolfe and Smith (2007a; 2007b) seven-criterion framework. This framework was adopted *a priori* as a vehicle for identifying validity evidence rather than simply classifying typical features of TPACK research or counting the occurrence of these features in the TPACK literature. We selected studies because they exemplified one or more aspects of validity evidence of relevance to TPACK measurement. However, locating examples of all of the types of evidence was difficult and, for some types of evidence, not successful. For example, specification of scaling models and testing the technical quality of items are very rare and only found in one study (i.e., Jamieson-Proctor, Finger, Albion, Cavanagh, Fitzgerald, Bond, & Grimbeek, 2012), which applied many, but not all, of the AERA, APA and NCME standards for instrument construction. This potential over-reliance on one study is a limitation of this paper and will hopefully be overcome as advances are made in attention to validity in future TPACK research.

We commence by examining the content aspect of validity that begins with the reason for measurement. Then we examine a sequence of activities that lead from clarification of the construct of interest to the design of the instrument.

Evidence of the Content Aspect

Purpose

The evidence of content aspect of validity includes clear statements of the *purpose* of a study or instrument development process that are made before other activities are attempted. Asking research questions is one widely used method of expressing the intent of an investigation. For example, in a study of TPACK measures, Koehler, Shin, and Mishra (2011, p. 18) made their purpose clear by posing two research questions: "What kinds of measures are used in the TPACK literature?" and "Are those measures reliable and valid?"

Also related to articulating a clear purpose for a study or measure is specifying the domain of inference, the types of inferences, and potential constraints and limitations.

Domain of inference. Specifying the *domain(s) of inference* situates the anticipated outcomes of an investigation within an established body of theory or knowledge and provides additional evidence of the content. The domains could be curricular (relating to instruction), cognitive (relating to cognitive theory), or criterion-based (knowledge, skills, and behaviors required for success in a particular setting). For example, the domain of inference of

TPACK is curricular due to the pedagogical component (Mishra & Koehler, 2006; Koehler & Mishra 2008) and also criterion-based due to its contextual specificity and situational dependence (Doering, Scharber, Miller, & Veletsianos, 2009).

Type of Inferences. The *types of inferences* delimit the intended conclusions or judgments to be made from a study or instrument. Presumably, TPACK studies or measures could be designed to make inferences about mastery, individual teachers, systems, or groups of teachers. To date, TPACK measurements have primarily sought to measure individual teachers' TPACK (Roblyer & Doering, 2010; Schmidt, Baran, Thompson, Mishra, Koehler, & Shin, 2009), although there have been notable attempts to study groups of teachers as well (e.g., Finger, et al., 2012).

There is also an element of mastery underpinning TPACK through the implication that high technology integration results from high levels of, and interaction between, technological, pedagogical, and content knowledge. Schmidt et al. (2009, p. 125) explained, "At the intersection of these three knowledge types is an intuitive understanding of teaching content with appropriate pedagogical methods and technologies."

Potential constraints and limitations. *Potential constraints and limitations* can also be identified that comment on the logistics, resource issues, or methodological exigencies. For example, Harris, Grandgenett, and Hofer (2010) identified a methodological limitation when they criticized self-report methods in TPACK research. The authors explained that "the challenges inherent in accurately estimating teachers' knowledge via self-reports—in particular, that of inexperienced teachers—are well-documented" (Harris, et al., 2010, p. 1).

Instrument Specification

Following the definition of the purpose, a set of *instrument specifications* are developed. This task involves describing *constructs*, a *construct model*, and then a *construct map*.

Constructs. Wilson (2010) described a *construct* as "the theoretical object of our interest" (p. 6) and saw it resulting from knowledge about the purpose of designing an instrument and the context in which it is to be used. He also considered a *construct* to be part of a theoretical model that explains phenomena. Importantly, the construct should sit within a well-established body of knowledge, and one of the *purposes* of a study is to contribute to extant theory in this *domain of inference*. The *construct model* and this theory are *a priori* considerations that require specification prior to other measure construction activities.

The TPACK framework could be viewed as a representation of one construct, a trait or ability of teachers that is not directly observable but is latent and indicated by observable behaviors. For example, Koehler et al. (2011, p. 6) explained that the "TPACK framework connects technology to curriculum content and specific pedagogical approaches and describes how teachers' understandings of these three knowledge bases can interact with one another to produce effective discipline-based teaching with educational technologies" (p. 6).

Alternatively, TPACK could be viewed as a composite of the seven constructs of Figure 1 (p. 130), each of which is sufficiently different from the others to warrant separate specification (Schmidt et al., 2009). The seven constructs comprise three types of knowledge—technological knowledge (TK), pedagogical knowledge (PK), and content knowledge (CK); and three types of knowledge about the interactions between technology, pedagogy, and content—pedagogical content knowledge (PCK), technological pedagogical knowledge (TPK), technological content knowledge (TCK); and then the interaction between PCK, TPK, and TCK—technological pedagogical content knowledge (TPACK). Additional complexities are contextual dependency on situational variables (e.g., subject discipline), which needs to be accommodated in both the unified and the multi-component representations, and the possibility of perhaps as few as three components (Archambult & Barnett, 2010) or more than seven components.

Empirical studies that use TPACK to guide research have tended to focus on one specific aspect of TPACK. Angeli and Valanides (2009) researched a strand within an alternative TPACK framework they termed ICT-TPCK; Harris et al. (2010) studied the quality of technology integration; and Jamieson-Proctor et al. (2012) evaluated TPACK confidence and usefulness. In these cases, models supplemented the more general TPACK model utilizing Venn diagrams that altered the focus on the phenomenon of interest.

Construct models. There are many sources of information that can assist in depicting a *construct model*. Wolfe and Smith (2007a) listed real-world observations, literature reviews of theory, literature reviews of empirical research, reviews of existing instruments, expert and lay viewpoints, and content and task analyses. Constructs can have internal and external models. An *internal model* typically comprises components, facets, elements or factors, and the hypothesized relations between these components. The TPACK models above are examples of internal models. Another example of an internal model is represented in Table 2 (Jamieson-Proctor et al., 2012, p. 5). The construct model for the *T*eaching Teachers for the Future (TTFF) TPACK Survey has seven components: TPACK, TPK, TCK, confidence to support student learning, confidence to support teaching, usefulness to support student learning, and usefulness to support teaching.

External models represent relations between the target construct and other constructs. Constructs associated with context (e.g., racial identity, learning environment, professional development) and how these relate to TPACK could constitute external models. An early version of the TTF instrument (Jamieson-Proctor, Finger, Albion, Cavanagh, Fitzgerald, Bond, & Grimbeek, 2012) contained a set of items on teacher efficacy. These items

Table 2. The Conceptual Structure of the TTF TPACK Survey

TPACK Framework Dimension	Scale: Confidence to Use ICT to:	Scale: Usefulness of ICT to:
TPACK	Support student learning	Support student learning
TPK, TCK	Support teaching	Support teaching

were intended to measure what was at the time considered a construct related to TPACK.

Construct maps. The *construct map* requires qualification of the *construct model* by providing a coherent and substantive definition of the content of the construct and a proposal of some form of ordering of persons or of the tasks administered to persons (Wilson, 2010). From a content perspective, the extension of Shulman's (1986; 1987) conception of pedagogical content knowledge (PCK) by the addition of technological knowledge (TK) has produced the integrative TPACK model (Graham, 2011). However, the PCK model and associated definitions have been criticized for imprecision and thus being "a barrier to the measurement of PCK" (Graham, 2011, p. 1955). This in turn has led to problems when defining the TPACK construct and the need for ongoing work in this area to resolve these issues (Koehler, Shin, & Mishra, 2011).

The issue of definitional precision is not peculiar to TPACK measurement. Wilson (2010, p. 28) referred to it as the "more complex reality of usage" and suggested some constructs should be conceptualized as multidimensional and represented by several discreet construct maps. He also recommended initial focus on one dimension at a time and development of a simple model on the assumption that complications can be dealt with later. This approach is compatible with the transformative view of TPACK that focuses on change and growth of teachers' knowledge over time rather than on discriminating between different types of TPACK knowledge (Graham, 2011). It is also consistent with the general objectives of measurement—interpersonal comparison of capabilities or dispositions, comparison of an individual's capabilities or dispositions at different times, or comparison of the difficulty the tasks comprising a measure present to persons.

The notion of ordering of persons or of instrument tasks has been successfully applied in construct mapping of TPACK. Harris, Grandgenett, and Hofer (2012) developed a rubric to rate experienced teachers on four forms of technology use when planning instruction. Twelve scorers assessed curriculum goals and technologies, instructional strategies and technologies, technology selection, and fit using a scoring rubric that described four levels of each form of technology use. They rated curriculum goals and technologies "strongly aligned" (scored 4), "aligned" (scored 3), "partially aligned," (scored 2), and "not aligned" (scored 1). The goal of this exercise was evaluating teachers' TPACK by ordering of persons.

The ordering of tasks assumes that different tasks present varying degrees of difficulty to the persons attempting the tasks. An example of a task-ordered rubric is the six facets of learning for understanding developed by Wiggins and McTighe (1998; 2005). The facet of *explanation* was postulated to vary in degree from naïve to sophisticated. Five levels were defined—naïve, intuitive, developed in-depth, and sophisticated. A naïve understanding was described as "a superficial account; more descriptive than analytic or creative; a fragmentary or sketchy account of facts/ideas or glib generalizations" (Wiggins & McTighe, 1998, p. 76). In contrast, sophisticated understanding could be demonstrated by "an unusually thorough, elegant, and inventive account (model, theory, or explanation)" (Wiggins & McTighe, 1998, p. 76). The facets of a learning rubric describe student behaviors at each level to differentiate between levels as well as to order the levels. Such a system of ordering is important when the construct of interest is hypothesized to be cognitively developmental with the attainment of lower-level tasks prerequisite to mastering those at higher levels. In the Wiggins and McTighe (1998; 2005) construct map, naïve explanations are easier to provide than intuitive explanations, which in turn are easier to provide than developed explanations (Cavanagh, 2011). This ordering informs theorizing about students learning for understanding. A developmental view of TPACK learning in which teacher cognition progresses through developmental stages would also require the identification of similar sequences of levels for the construct map and then the development of instrument items.

Item development. *Item development* concerns making choices about *item formats* such as multiple choice, rating scales, and performance assessments. This can be informed by following the recommendations of *item writing guidelines* about content/semantics, formatting, style, stem statements, response scales, and response choices. *Regular reviews* such as expert reviews, content reviews, and sensitivity (targeting) reviews can be conducted throughout all stages of instrument development. For example, seven TPACK experts reviewed the validity and face value of the rubric developed by Harris et al. (2012) to assess observed evidence of TPACK during classroom instruction.

Scoring model. A detailed construct map with an internal structure that orders persons and tasks informs selection of a scoring model. Significantly, it is the ordering that provides a foundation for the instrument being a measure. A *scoring model* shows how observations or responses to items are numerically coded. Right or wrong answers provide dichotomous data that could be scored 0, 1. Rating scales produce polytomous data that can be scored using the successive integers 0, 1, 2, and 3. Rating scales can show the degree of agreement of respondents to a stem statement, and while this is related to the overall strength of the trait of interest in persons, it is the ordering within the construct map that constitutes the measure.

The number and labeling of response categories is crucial to the performance of a rating scale instrument (Hawthorne, Mouthaan, Forbes, & Novaco, 2006; Preston & Colman, 2000). Another related issue is use of a "neither disagree or agree" category and the reasons for the selection of this category (Kulas & Stachowski, 2001). The scoring model for the TTF TPACK Survey instrument (Jamieson-Proctor et al., 2012) comprised seven response categories scored 0 (not confident/useful); 1, 2, 3 (moderately confident/useful); 4, 5, 6 (extremely confident/useful); plus an additional "unable to judge" category scored 8 and coded as missing data. We collected data using Qualtrics online survey software.

Scaling model. The data obtained directly from instrument administration are termed *raw data* because they require processing by scaling into a meaningful form. Without scaling, the use of raw scores is limited to the presentation of frequencies, and even mathematical operations as basic as estimating a mean score should be undertaken with caution (Doig & Groves, 2006). A *scaling model* such as the Rasch Model (Rasch 1980) can be applied to raw scores to calibrate these on a linear scale. The intervals on a linear scale are equal in the same way as the markings on a yardstick. This enables comparison of person scores according to their magnitude and not just their order.

We analyzed the TTF TPACK Survey student scores using the Rasch Rating Scale Model (Andrich, 1978a; Andrich, 1978b; Andrich, 1978c; Bond & Fox, 2007; Jamieson-Proctor, Finger, Albion, Cavanagh, Fitzgerald, Bond, & Grimbeek, 2012). Data from four groups of like-named items (i.e., TPK/TCK Confidence, TPK/TCK Usefulness, TPACK Confidence, TPACK Usefulness) were subject to separate scaling, and then we equated scaled scores on an interval scale (Jamieson-Proctor, Finger, Albion, Cavanagh, Fitzgerald, Bond, & Grimbeek, 2012). The generation of interval data enabled accurate comparison of student responses on four scales between the two occurrences of instrument administration at the national level and also within the 39 universities/higher education providers that participated in the project.

Item technical quality. Evidence of *item technical quality* can be garnered by testing how well data from individual items meet the requirements of an item-response measurement model. For example, in its simplest form, the Rasch Model requires the probability of a person completing a task to be a function of that person's ability and the difficulty of the task. Persons with high ability are more likely to complete difficult tasks than those with lower ability. Conjointly, easy tasks are likely to be completed by both low- and high-ability persons. Rasch Model computer programs such as RUMM2030 (RUMMLab, 2007) or Winsteps (Linacre, 2009) test how well the responses to an item display this property by estimating fit statistics. Common reasons for items having poor fit to the model include the item not discriminating between persons of different ability and the responses being confounded by an attribute of the persons different to the trait being measured.

Rasch Model analysis of the *TTF TPACK Survey* data using the WINSTEPS computer program (Linacre, 2009) identified six items with misfitting data. These were stepwise removed from subsequent analyses until all the remaining items showed adequate fit to the model's requirements for measurement. The items removed and their respective scales were:

- **Scale TPK/TCK Confidence Combined:** Teach strategies to support students from Aboriginal and Torres Strait Islander backgrounds; access, record, manage, and analyze student assessment data
- **Scale TPK/TCK Usefulness Combined:** Teach strategies to support students from Aboriginal and Torres Strait Islander backgrounds; manage challenging student behavior by encouraging the responsible use of ICT
- **Scale TPACK Confidence Combined:** Communicate with others locally and globally
- **Scale TPACK Usefulness Combined:** Communicate with others locally and globally (Jamieson-Proctor, Finger, Albion, Cavanagh, Fitzgerald, Bond, & Grimbeek, 2012. p. 8)

Another consideration in rating scale instruments is the functioning of the rating scale categories. There is a diversity of views on the optimum number of response categories (Hawthorne, Mouthaan, Forbes, & Novaco, 2006; Preston & Colman, 2000). There are also many reasons, which are often unclear, for selecting a "neither disagree or agree," "undecided," or "not sure" category as a middle category (Kulas & Stachowski, 2001). Optimizing the response scale is possible by analysis of pilot and trial data using the Rasch Rating Scale Model (Andrich, 1978a; Andrich, 1978b; Andrich, 1978c). For an item, a Category Probability Curve is produced from plotting the responses to each category in the response scale against the ability of the persons. An ideal pattern of responses would show the more capable respondents choosing the most difficult to affirm categories and the less capable respondents choosing the easier to affirm categories. For the seven-category response scales used in the TTFF study, some of the provided response options were not used as intended. Consequently, "adjacent response categories were combined as required to achieve satisfactory Category performance" (Jamieson-Proctor, et al., 2012, p. 8).

The preceding section on the content aspect of validity described the key activities in the construction of a measure, and methods for ensuring these are implemented as intended. The content activities are sequential and iterative but require implementation in conjunction with the other six aspects of validity evidence. With this in mind, the following six sections examine substantive, structural, generalizability, consequential, and interpretability evidence of validity.

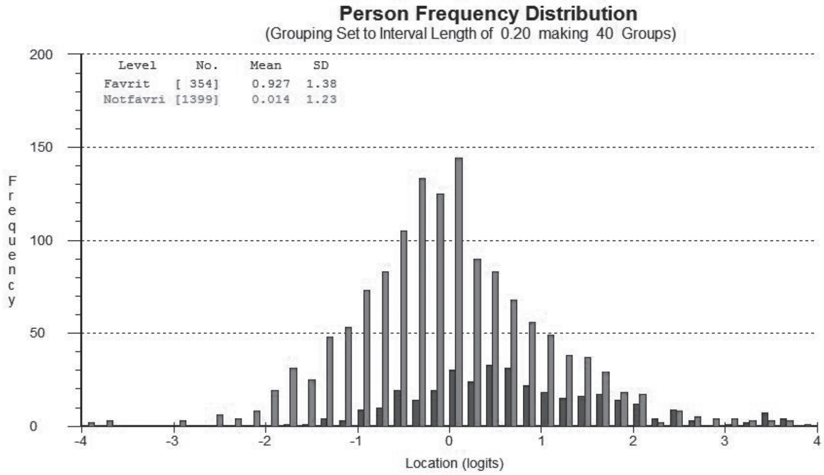


Figure 2. Frequency distributions of student engagement scores for favourite and nonfavorite subjects (*N*=1743).

Evidence of the Substantive Aspect

The substantive aspect of validity can be evidenced by the extent to which the theoretical framework, an a priori theory, or the hypothesis informing an investigation can explain any observed consistencies among item responses. This section examines each approach.

For example, the literature on student engagement suggests that it is characterized by enjoyable experiences in the classroom and a favorable disposition toward the material being learned and toward the classroom environment (Shernoff, 2010). Students describing their favorite class would be expected to have higher engagement scores than those describing a nonfavorite class. We used RUMM2030 to calculate engagement scores for data from the Survey of Student Engagement in Classroom Learning (Cavanagh, 2012). Figure 2 presents the frequency of scores (person locations measured in logits) for students reporting their favorite subjects and those reporting a nonfavorite subject. The scores for favorite subject were statistically significantly higher than those for the nonfavorite subjects (i.e., mean score favorite .93 logits and mean score nonfavorite .01 logits, F=147.7, p< .000).

A similar approach for gathering substantive evidence could be used with TPACK construct models and data. There are likely particular groups of teachers with attributes anticipated to be associated with high TPACK scores. These could be teachers who have completed postgraduate courses in technology integration, teachers who have received substantial professional development in technology integration, teachers who have been recognized for outstanding technology use in their classroom, teachers who have received awards for innovative technology use in the classroom, and/or teachers selected to mentor or train colleagues in technology integration.

Evidence of the Structural Aspect

The *structural* aspect of validity concerns the construct model and map, for example, by ascertaining if the requirements of a unidimensional measurement model are met when a unidimensional trait is measured. There are both traditional and contemporary methods for collecting evidence about construct structure. The traditional approach is to conduct a Principal Components Factor Analysis of raw scores (dichotomous or polytomous) to examine correlations and covariance between items by identifying factorial structure in the data. Provided there is sufficient data in relation to the numbers of items in the scale under scrutiny, this method is well accepted in TPACK research. Notwithstanding, smaller data sets and large instruments (many items) have required a multiscale approach. Schmidt et al. (2009) developed a 75-item instrument measuring preservice teachers' self-assessments of the seven TPACK dimensions: 8 TK items, 17 CK items, 10 PK items, 8 PCK items, 8 TCK items, 15 TPK items, and 9 TPACK items. However, the sample included only 124 preservice teachers, which precluded a full exploratory factor analysis of data from all 75 items but did allow separate analyses of the seven dimensions. In this study (Schmidt et al. 2009), factor loadings were estimated, "problematic" items were "eliminated," and Cronbach's alpha reliability coefficient was computed for data from the retained items in each scale. This process provided evidence of the internal structure of the seven dimensions but did not confirm a seven-dimension construct model of TPACK. Similarly, the TTF TPACK Survey data were subject to two exploratory factor analyses: one for the 24 TPK and TCK items and one for the 24 TPACK items. We found two-factor solutions in both cases, with the confidence data loaded on one factor and the usefulness data loaded on the second factor (Jamieson-Proctor, Finger, Albion, Cavanagh, Fitzgerald, Bond, & Grimbeek, 2012). The results provide confirmatory evidence of the construct model in Table 2 (p. 135).

Another approach to garnering evidence of dimensionality uses the Rasch Model. The linear Rasch measure is extracted from the data set after the initial Rasch scaling, and then a Principal Components Factor Analysis of the residuals is conducted. The assumption underlying this process is that variance within the data should be mainly attributable to the Rasch measure and that there will be minimal structure and noise in the residual data. Application of this approach to phenomena that are clearly multivariate requires separate Rasch Model analyses for each variable. This was the case with the TTF TPACK Survey data. We used four Rasch Model analyses and took the sound data to model fit in the four scales as evidence of the structure within the four-component construct model presented in Table 2 (Jamieson-Proctor, et al., 2012).

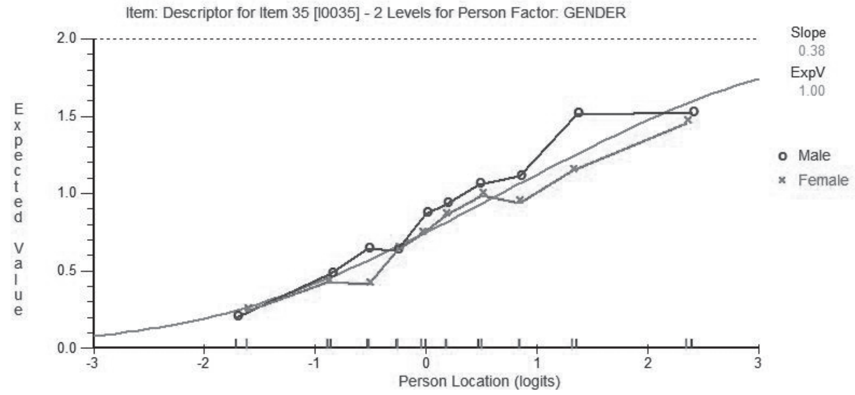


Figure 3. Item characteristic curve for Item 35 (*N*=1745).

Evidence of the Generalizability Aspect

Wolfe and Smith (2007b) explained "the generalizability aspect of validity addresses the degree to which measures maintain their meaning across measurement contexts" (p. 215). For example, consider an item for which the success rate does not differ between males and females. A lack of this property of an item is referred to as differential item functioning (DIF). Testing for DIF typically proceeds by generating an Item Characteristic Curve and plotting observed scores for class intervals of groups of persons of interest. Figure 3 displays this information for Item 35 ("My test scores are high") from the Survey of Student Engagement in Classroom Learning (Cavanagh, 2012). When the observed responses of boys and girls with the same engagement level are compared, the more highly engaged boys responded more affirmatively than the more highly engaged girls (F=15.05, p< .001). The item has functioned differently for males and females.

A similar approach for gathering generalizability evidence could be used with TPACK models and data. Ideally, there should be no difference in scores for a TPACK item between groups of teachers with the same overall score, such as between groups of male and female teachers, city and rural teachers, or experienced and inexperienced teachers. This does not negate the overall instrument discriminating between different groups; it merely avoids bias at the item level.

Evidence of the External Aspect

The relation between a measure and an external measure of a similar construct can show the external aspect of validity. For example, the developers of the TTF TPACK Survey acknowledged the importance of using external measures: "As with all self-report instruments, data collected with this instrument should be complemented with other data collection methodologies to overcome the limitations associated with self-report instruments"

(Jamieson-Proctor, Finger, Albion, Cavanagh, Fitzgerald, Bond, & Grimbeek, 2012, p. 9). For similar reasons, Harris et al. (2010; 2012) assessed the quality TPACK through examination of detailed written lesson plans and also semi-structured interviews of teachers. However, the extent to which a second measure can be independent of the first is difficult to establish, particularly when both measures share a common construct model or measure a similar construct.

Evidence of the Consequential Aspect

The consequential aspect of validity centers on judgments about how the score interpretations might be of consequence. When measures are used in high-stakes testing, the consequences for students, teachers, and schools can be significant and sometimes the source of serious concern. Measuring TPACK is unlikely to have such consequences, but applications that compare teachers against one another or against benchmarks for performance management purposes could be seen as less benign. TPACK researchers should consider potential consequences, and such consideration is further evidence for establishing consequential validity.

Evidence of the Interpretability Aspect

The interpretability aspect of validity concerns the qualitative interpretation of a measure in terms of how well its meaning was communicated. Figures and graphical displays can assist the reader in understanding the meaning of an instrument and the properties of its data. The TTF TPACK Survey was developed to test for change in TPACK in Australian preservice teachers who were provided with six months of specialized instruction in technology integration. The results of this testing were presented as graphics such as Figure 4 (Finger et al. 2012, p. 12). This is an item-by-item display of scores from the first survey administration and of scores from the second survey administration for the confidence items. Rasch Model equating procedures have enabled all the scores to be plotted on the same scale. The improvement in scores for all the items is obvious.

Another useful display is an item map that plots the difficulty of items and the ability of persons on the same scale. Figure 5 is the item map for a scale measuring student engagement and classroom learning environment (Cavanagh, 2012, p. 9). The scale is marked in logits from 3.0 to -3.0. The student scores are located on the scale, and × indicates 10 students. The students with the most affirmative views are located toward the top of the distribution. The location of an item shows the difficulty students experienced in affirming the item. The items located towards the top of the distribution were more difficult to affirm than those below. The items are numbered according to their labeling in the instrument. Item 41 (“I start work as soon as I enter the room”) and Item 48 (“Students do not stop others from working”) were the most difficult to affirm, whereas Item 7 (“I make an effort”)

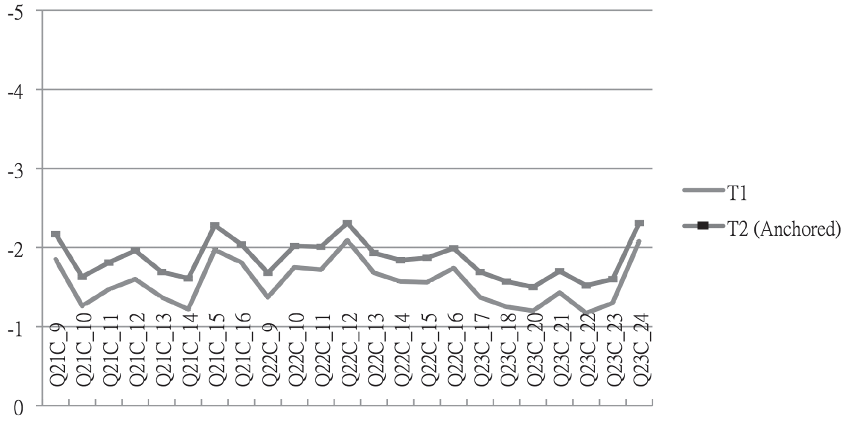
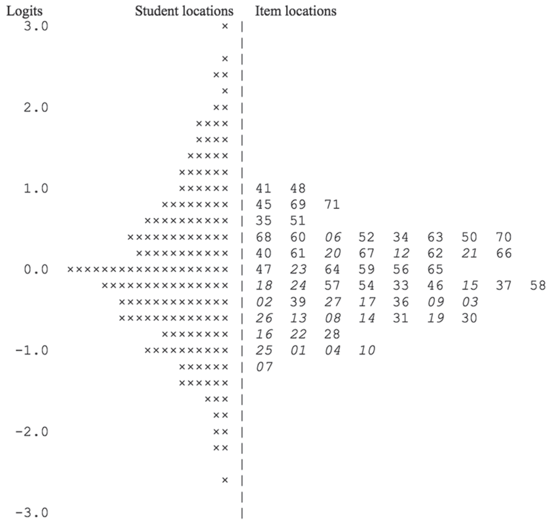


Figure 4. Confidence to facilitate student use.



```
Logits              Student locations  Item locations
 3.0                              × |
                                    |
                                  × |
                                 ×× |
                                  × |
 2.0                             ×× |
                               ×××× |
                               ×××× |
                              ××××× |
                             ×××××× |
 1.0                         ×××××× | 41  48
                           ×××××××× | 45  69  71
                         ×××××××××× | 35  51
                       ×××××××××××× | 68  60  06  52  34  63  50  70
                        ××××××××××× | 40  61  20  67  12  62  21  66
 0.0        ××××××××××××××××××××××× | 47  23  64  59  56  65
                    ××××××××××××××× | 18  24  57  54  33  46  15  37  58
                      ××××××××××××× | 02  39  27  17  36  09  03
                      ××××××××××××× | 26  13  08  14  31  19  30
                           ×××××××× | 16  22  28
-1.0                     ×××××××××× | 25  01  04  10
                             ×××××× | 07
                             ×××××× |
                                ××× |
                                 ×× |
-2.0                             ×× |
                                 ×× |
                                    |
                                  × |
                                    |
-3.0                                |
```

Figure 5. Item map for engagement and learning environment items.

was easy to affirm. The relation between student scores and item difficulty enables predictions to be made about student responses. Students with locations below 1.0 logits are unlikely to affirm Items 41 and 48. Conversely, those with locations above 1.0 logits are likely to affirm these items.

For TPACK measurement, the calibration of items as illustrated in the item map would enable profiling of TPACK for many teachers at different times and in different situations. It would also accurately show changes in TPACK over time for individual teachers. The scaling of person scores and item difficulty scores is essential for constructing an item map; raw scores are not suitable for this purpose.

A Checklist for Researchers

The preceding sections have examined seven aspects of validity evidence, and, where possible, examples of TPACK measurement were used to illustrate these aspects and situate them within the epistemology and methodology of TPACK research. Table 3 lists the definitions of the seven aspects to provide a tabular recount of the key considerations in mounting an argument for validity. The table could be used as a checklist for TPACK researchers to assess the validity of their research, either *a priori* when designing TPACK measures or *post hoc* to evaluate existing TPACK measures.

The use of the checklist requires comment. First, it is more than a simple seven-item list; the content exemplifies contemporary understandings of validity and validity theory. The underlying approach and its major features have been explained in this paper, but this explication has been limited. Users of the table would likely benefit by consulting some of the original sources referenced in the text. Second, statistics such as correlation coefficients or the results of an exploratory factor analysis are often put forward as proof of validity. Statistical evidence is just one aspect of an argument for validity, and an argument relying on only this form of evidence would be weak. Third, the application of the checklist should be on the availability of evidence rather than simply whether attention has been given to each particular aspect, although this would be a useful starting point. The notion of validity being an argument requires the provision of evidence to convince others, and the checklist is simply a vehicle for stimulating and organizing evidence collection. Fourth, the availability of extensive evidence of all seven aspects is an optimal situation and, in reality, not attainable in many educational studies. This limitation is methodological and mainly centered on the instrument specification process within the content aspect. The use of measurement models that examine the properties of data at the level of individual items and persons can ensure instrument specification complies with the content evidence requirements. Detailed and persuasive evidence is available when Item Response Theory and Rasch Model methods are used.

While the iterative nature of instrument construction might suggest that the sequencing of the seven aspects could be varied, there are some strong reasons for commencing with the content aspect. The rationale for this view derives from a scientific approach to educational research, including TPACK research, that is very consistent with Messick's (1995) view of validity. In both, primacy is given to substantive theory informing decisions about instrumentation. The research is driven by theory rather than theory being generated from existing data; in terms of validity, specification of the construct model, particularly the construct map, precedes selection of data collection methods and analyses. When the checklist is used *post hoc*, this matter is more important for principled rather than pragmatic reasons. However, when using the checklist *a priori* at the commencement of a study, substantive theory and the findings of previous research require clarification

Table 3. A Checklist of Validity Evidence

Aspect of evidence	Definition	
1. Content	The relevance and representativeness of the content upon which the items are based and the technical quality of those items	
	Purpose	Domain of inference Types of inferences Potential constraints and limitations
	Instrument specification	Construct selection Construct model Construct map Item development Scoring model Scaling model Item technical quality
2. Substantive	The degree to which theoretical rationales relating to both item content and processing models adequately explain the observed consistencies among item responses	
3. Structural	The fidelity of the scoring structure to the structure of the construct domain	
4. Generalizability	The degree to which score properties and interpretations generalize to and across population groups, settings, and tasks, as well as the generalization of criterion relationships	
5. External	What has traditionally been termed convergent and discriminant validity and also concerns criterion relevance and the applied utility of the measures	
6. Consequential	The value implications of score interpretation as a basis for action	
7. Interpretability	The degree to which qualitative meaning can be assigned to quantitative measures	

(Wolfe & Smith, 2007a, p. 99)

before progressing to methodological decisions. In this situation, the order of the seven aspects is important.

The final consideration in the use of the checklist is that it is neither exhaustive nor the only way to conceptualize an argument for validity. For example, in the hard sciences, where causal relations exist between variables, the dominant form of validity is predictive validity. Notwithstanding, we believe that an argument is required, and this needs to reflect all aspects of an instrument development process or of an empirical investigation.

Conclusion

One purpose of this paper was to stimulate discussion about the validity of TPACK measures and measurement. A second purpose was to use contemporary validity theory as a framework to examine the principles and practices applied when dealing with validity issues in TPACK measurement. The analysis suggests several types of validity evidence that are not characteristic of current TPACK measurement activities, and that identification of these factors could provide the impetus for improvement of TPACK measurement. In particular, the content and substantive aspects of validity evidence are especially challenging.

TPACK theory is still in its infancy, as is the measurement of TPACK. It is timely to consider concerns such as validity from the perspective of mainstream epistemologies and methodologies. Maturation of TPACK research and measurement requires nurture and sustenance from well-established fields of research and methodologies.

Acknowledgment

The authors gratefully acknowledge the assistance of Joshua Rosenberg with the preparation of this manuscript.

Author Note

Robert F. Cavanagh is a professor in the School of Education at Curtin University, Perth, Australia. His research interests focus on the measurement of student, teacher, and classroom attributes conducive to improved learning and instruction. Please address correspondence regarding this article to Rob Cavanagh, School of Education, Curtin University, Kent St., Bentley 6102, Australia. Email: r.cavanagh@curtin.edu.au.

Matthew J. Koehler is a professor in the College of Education at Michigan State University, East Lansing. His research interests focus on the design and assessment of innovative learning environments and the knowledge that teachers need to teach with technology.

References

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

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581–594. doi:10.1177/014662167800200413

Andrich, D. (1978b). Rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. doi:10.1007/BF02293814

Andrich, D. (1978c). Scaling attitude items constructed and scores in the Likert tradition. *Educational and Psychological Measurement, 38*(3), 665–680. doi:10.1177/001316447803800308

Angeli, C., & Valanides, N. (2009). Epistemological and methodological issues for the conceptualization, development, and assessment of ICT-TPCK: Advances in technological pedagogical content knowledge (TPCK). *Computers and Education, 52*(1), 154–168. doi:10.1016/j.compedu.2008.07.006

Archambault, L. M., & Barnett, J. H. (2010). Revisiting technological pedagogical content knowledge: Exploring the TPACK framework. *Computers and Education, 55*(4), 1656–1662. doi:10.1016/j.compedu.2010.07.009

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Cavanagh, R. F. (2011a). Establishing the validity of rating scale instrumentation in learning environment investigations. In R. F. Cavanagh, & R. F. Waugh (Eds.), *Applications of Rasch measurement in learning environments research* (pp. 77–100). Rotterdam: Sense Publishers.

Cavanagh, R. F. (2011b). Confirming the conceptual content and structure of a curriculum framework: A Rasch Rating Scale Model approach. *Curriculum Perspectives, 31*(1), 42–51.

Cavanagh, R. F. (2012). *Associations between the classroom learning environment and student engagement in learning: A Rasch model approach*. Paper presented at the meeting of the Australian Association for Research in Education: Sydney, Australia.

Doering, A., Scharber, C., Miller, C., & Veletsianos, G. (2009). GeoThentic: Designing and assessing with technology, pedagogy, and content knowledge. *Contemporary Issues in Technology and Teacher Education, 9*(3), 316–336. Retrieved from http://www.citejournal.org/vol9/iss3/socialstudies/article1.cfm

Doig, B., & Groves, S. (2006). Easier analysis and better reporting: Modeling ordinal data in mathematics education research. *Mathematics Education Review Journal, 18*(2), 56–76. doi:10.1007/BF03217436

Finger, G., Jamieson-Proctor, R., Cavanagh, R., Albion, P., Grimbeek, P., Bond, T., Fitzgerald, R., Romeo, G., & Lloyd, M. (2012). *Teaching teachers for the future (TTF) project TPACK survey: Summary of the key findings.* Paper presented at ACEC2012: ITs Time Conference, Perth, Australia. Available at: http://bit.ly/ACEC2012_Proceedings

Graham, C. R. (2011). Theoretical considerations for understanding technological pedagogical content knowledge (TPACK). *Computers & Education, 57*(3), 1953–1960. Retrieved from http://www.sciencedirect.com/science/article/pii/S0360131511000911

Graham, C., Cox, S., & Velasquez, A. (2009). Teaching and measuring TPACK development in two preservice teacher preparation programs. In I. Gibson et al. (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2009* (pp. 4081–4086). Chesapeake, VA: AACE. Retrieved August 19, 2013, from http://www.editlib.org/p/31297

Guzey, S. S., & Roehrig, G. H. (2009). Teaching science with technology: Case studies of science teachers' development of Technological Pedagogical Content Knowledge (TPCK). *Contemporary Issues in Technology and Teacher Education, 9*(1), 25–45. AACE. Retrieved August 18, 2013 from http://www.editlib.org/p/29293

Harris, J., Grandgenett, N., & Hofer, M. (2010). Testing a TPACK-Based Technology Integration Assessment Rubric. In D. Gibson & B. Dodge (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2010* (pp. 3833–3840). Chesapeake, VA: AACE. Retrieved August 18, 2013, from http://www.editlib.org/p/33978

Harris, J., Grandgenett, N., & Hofer, M. (2012). Using structured interviews to assess experienced teachers' TPACK. In P. Resta (Ed.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2012* (pp. 4696–4703). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/40351

Hawthorne, G., Mouthaan, J., Forbes, D., & Novaco, R. W. (2006). Response categories and anger measurement: Do fewer categories result in poorer measurement? Development of the DAR5. *Social Psychiatry Psychiatric Epidemiology, 41*(2), 164–172. doi:10.1007/s00127-005-0986-y

Jamieson-Proctor, R., Finger, G., Albion, P., Cavanagh, R., Fitzgerald, R., Bond, T., & Grimbeek, P. (2012). *Teaching Teachers for the Future (TTF) project: Development of the TTF TPACK survey instrument.* Paper presented at ACEC2012: ITs Time Conference, Perth, Australia. Available at: http://bit.ly/ACEC2012_Proceedings

Koehler, M. J., & Mishra, P. (2008). Introducing TPCK. In AACTE Committee on Technology and Innovation (Ed.), *Handbook of technological pedagogical content knowledge (TPCK) for educators* (pp. 3–29). London: Routledge.

Koehler, M. J., Shin, T. S., & Mishra, P. (2011). How do we measure TPACK? Let me count the ways. In R. N. Ronau, C. R. Rakes, & M. L. Niess (Eds.), *Educational technology, teacher knowledge, and classroom impact: A research handbook on frameworks and approaches* (pp. 16–31). Hershey, PA: Information Science Reference.

Kulas, J. T., & Stachowski, A. A. (2001). Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators. *Journal of Research in Personality, 47*, 254–262. doi: 10.1016/j.jrp.2013.01.014

Linacre, J. M. (2009). *Winsteps* (Version 3.68) [Computer Software]. Beaverton, OR: Winsteps.com.

Medical Outcomes Trust Scientific Advisory Committee. (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin, 3*, 1–4.

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. doi:10.1037/0003-066X.50.9.741

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*(4), 35–44. doi:10.1023/A:1006964925094

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. doi:10.1111/j.1467-9620.2006.00684.x

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15. doi:10.1016/S0001-6918(99)00050-5

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Roblyer, M. D., & Doering, A. H. (2010). *Integrating educational technology into teaching* (5th ed.). Boston, MA: Allyn & Bacon.

RUMMLab. (2007). *RUMM2020 Rasch Unidimensional Measurement Models.* RUMM Laboratory Pty Ltd.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. doi:10.3102/0013189X015002004

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22. Retrieved from http://hepg.org/her/abstract/461

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological pedagogical content knowledge (TPACK): The development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education, 42*(2), 123–149.

Shernoff, D. J. (2010). *The experience of student engagement in high school classrooms.* Saarbrucken, Germany: Lambert Academic Publishing.

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandra, VA: Association for Supervision and Curriculum Development.

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Alexandra, Virginia: Association for Supervision and Curriculum Development.

Wilson, M. (2010). *Constructing measures: An item response approach*. New York: Routledge.

Wolfe, E.W., & Smith, E.V. (2007a). Instrument development tools and activities for measure validation using rasch models: Part I–instrument development tools. *Journal of Applied Measurement, 8*(1), 97–123. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17215568

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II–validation activities. *Journal of Applied Measurement, 8*(2), 294–234. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17440262

Young, A., & Cavanagh, R. F. (2011). An investigation of differential need for psychological services across learning environments. In R. F. Cavanagh & R. F. Waugh. (Eds.), *Applications of Rasch measurement in learning environments research* (pp. 227–244). Rotterdam: Sense Publishers. ISBN 978-94-6091-491-1.

Manuscript received July 12, 2013 | Initial decision July 30, 2013 | Revised manuscript accepted August 29, 2013