Development and evaluation of assessments for counseling professionals

By: A. Stephen Lenz and Kelly L. Wester

Lenz, A. S., & Wester, K. L. (2017). Development and evaluation of assessments for counseling professionals. *Measurement and Evaluation in Counseling and Development, 50,* 201-209.

Abstract:

It is imperative that counselors understand how to critically evaluate assessments before using them to make clinical decisions. This evaluation can be conducted through integrating the 5 sources of validity. Each source of validity is discussed, along with methods to appraise psychometric quality, throughout this special issue.

Keywords: Assessment | evaluation | validation | validity

Article:

Validity, reliability, and fairness in assessment practices might be the most integral, yet misinterpreted concepts within the fields of educational and psychological assessment. The *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) provided guidelines for the development and evaluation of assessment validity that have substantial implications for professional counselors in clinical and research settings alike. Such a resource is invaluable when considering that prudently selected, administered, scored, and interpreted assessments can shape the course and trajectory of an individual's development and wellness across the life span. Similarly, evidence-supported treatments and evidence-based practices that are nested within sensible, pragmatic, assessment-based data frameworks are instrumental for informing best practices for educational and public mental health policy. The confluence of these variables within the current sociopolitical climate has triggered a moment within our profession's history wherein a functional understanding of measurement concepts is imperative. Alas, educational experiences and continuing education activities for many counseling professionals have been underrepresented by plainly spoken illustrations of assessment development and evaluation

activities. Therefore, the aim of this special issue is to provide such support that is a useful guide to counseling practitioners, researchers, and students.

Such an endeavor seems judicious given the updates to the *Standards* since the previous explication in *Measurement and Evaluation in Counseling and Development* (*MECD*) by Goodwin and Leech (2003), as well as our current zeitgeist of assessment access, use, and interpretation in which we find ourselves as counseling professionals. On one hand, the *Standards* appear to have shifted to speak not just to the assessment pedant, but to a broader range of consumers of assessment scores including graduate students and policymakers. This is particularly evident in the field of counselor education wherein the proliferation of assessments accompanying various professional competencies has abounded. On the other hand, unsupervised assessment use appears to be at an all-time high. One need look no further than the Internet or a smartphone's application store to access ready-made assessments intended to quantify complex human experiences such as social-emotional competence, language proficiency, postpartum depression, and posttraumatic stress disorder. Although this unprecedented access is undeniably convenient and might prompt individuals toward better understanding their lived experiences or help-seeking behaviors, there is also a risk that misinterpretations of results could have deleterious implications. Hattie (2014) shared this sentiment, in part, noting that this level of access without scrutiny of implications for those completing assessments represents "a critical shift from seeing tests only as a sample of items, a scoring system, and a score to seeing tests as reports about performance" (p. 34), without adequate attention to psychometric properties or context of scores. Consequently, our hope is that the content of this special issue will not only guide the development and evaluation of assessment content for use among counseling professionals, but also encourage a degree of critical thinking that supports situating the assessment experience across multiple layers of theory and evidence that promotes fairness for clients and stakeholders.

Considering the Characteristics Defining Validity Within the *Standards*

We preface this introductory contribution with some general commentary related to the concept of validity as a foundation of prudent assessment practices that are intended as preliminary to the articles in this special issue. It is our belief that the discussion of processes and features representing sources of validity evidence is best situated within the context of some very important characteristics. Although scholars have offered both objections and affirming overtures about validity depicted within the *Standards*, some important considerations about validity have persisted as generally agreed on (Markus, 2016; Newton & Shaw, 2016). Namely, these characteristics include (a) the distinct relationship between validity and validation activities; (b) the integrated, synergistic nature of validity evidence;, (c) the satisfaction of all standards does not preclude utility of assessment scores; (d) validity is based on inferences about scores, not an assessment or test intrinsically; (e) defining the purpose of assessment scores cannot be underscored enough; and (f) without careful attention to the consequences of testing, wide access to assessments poses varied degrees of threat to test takers. To each of these points, we submit modest discourse based on our experiences using informal and formal assessments in educational, clinical, and research settings.

The Distinct Relationship Between Validity and Validation

The *Standards* identified validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). Within this statement, there is a distinct emphasis on estimations of validity being a matter of degree of indication rather than a categorical status. By contrast, validation refers to the many processes and practices implemented by researchers to accumulate evidence supporting or dispelling suppositions about the usefulness of assessment scores. In some cases, validation activities yield clear evidence for or against a particular use and interpretation, but in others, validation confirms and dispels the usefulness of scores when explaining the experiences of individuals across various intersections of identity such as age, gender, cultural expression, diagnostic presentation, educational achievement, and vocational interest. Taken together, validation is an ongoing process of estimating the usefulness of assessment scores for explaining the experiences and characteristics of individuals for a well-defined construct. From the information garnered through this process, counselors can make inferences about the degree to which a particular assessment's scores will support accurate interpretations and inferences about their population of interest.

Validity Is Inferred From Scores and Uses, Not the Assessment Itself

Does validity refer to whether or not something measures what it is supposed to? Sure it does, in about the same degree that putting yourself in someone else's shoes explains the relational intricacies that underlie communicating empathy. A precondition to discussing validation procedures is the proposition that validity is inferred from assessment scores and the ways they are used, not the assessment itself. That is to say, discussions of validity are always going to be in reference to the degree that scores representing a construct successfully explain important experiences or characteristics of individuals and groups. Furthermore, although scores might lead to declarations of some degree of validity based on available evidence, the application of those validity inferences could not possibly be useful for all people, for all purposes, and therefore for all interpretations (Drummond, Sheperis, & Jones, 2016). Thus, some amount of incredulity is to be expected not only when assessments are described as valid, but especially when they are treated as categorically valid indiscriminately across counseling settings and client populations. More accurately, some scores are transferable to some settings and populations, but careful consideration of important characteristics is an ethical imperative given the high-stakes nature of assessments within clinical practice.

Validity Is a Unitary Concept

Within the *Standards*, validity is regarded as a unitary concept in which the degree of evidence is accumulated for or against use of scores for a particular purpose that is centered on a construct of interest. Although the degree of validity that can be inferred from an assessment procedure or set of scores is informed by multiple sources of evidence, sources of validity evidence are not synonymous with types of validity. Therefore, it is important to implement a systematic approach such as that proposed by Kelly, O'Malley, Kallen, and Ford (2005) when considering validation evidence that recognizes the synergistic interplay between findings that lead to a unitary representation. This prospect is complicated by trends indicating that common practices for reviewing validity evidence still present findings in separate categories without integration

into a unified depiction (Cizek, Koons, & Rosenberg, 2011; Hogan & Agnello, 2004). This point underscores the importance for counselors to not solely rely on published reviews when making decisions about whether an assessment might be useful for a particular purpose. Instead, each decision should be made in consideration of the totality of relevant evidence in juxtaposition to characteristics representing the intersection of setting and population.

Validity Evidence Encompassing All Standards Does Not Imply Utility

Although the *Standards* provide clear guidance for the validation activities that promote validity estimation, they also caution against viewing the five sources of validity evidence as a checklist. This is because not all types of validity evidence are required for every assessment to be determined as useful and such a checklist approach would undermine the value of professional expertise. Instead, the evidence required for developing an integrated estimation of validity should be based on the proposed uses of assessment results (AERA, APA, & NCME, 2014). With this in mind, counselors are compelled to consider implementing validation activities that depict a clear fit between the proposed uses of assessment scores and the type of evidence provided by each source of validity evidence.

Sources of Validity Evidence

Reliability and validity ultimately are the twin pillars of psychometric quality of assessments; however, validity is the foundation providing evidence of assessment quality. Validity is "the most fundamental consideration in developing tests and evaluating tests" (AERA, APA, & NCME, 2014, p. 11). The foundations of sound assessment development, use, and interpretation rest on the integration of several sources of validity evidence. Although validity is conceptualized as a unitary concept, there are nevertheless some sources of validity evidence that provide a reasonable impetus for integration by consumers of assessment scores. It is the accumulated evidence across these sources of validity that supports the interpretation of the assessment scores for their intended purpose. These sources of validity include evidence based on (a) assessment content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) evidence of consequential validity.

Before diving into the specific sources of validity, it should be noted that although validity is the foundation of psychometric quality, it is infrequently reported among assessments that have been developed (Cizek, Bowen, & Church, 2010; Cizek, Rosenberg, & Koons, 2008; Hogan & Agnello, 2004). Within the flagship journal of counseling, *Journal of Counseling & Development*, over half (60%) of assessments used lacked validity information (Wester, Borders, Boul, & Horton, 2013), bringing into question the quality and appropriateness of the assessments and the interpretation of the scores. To use underdeveloped, and potentially invalid, assessments can bring about unintended and grave consequences. A better understanding of the five sources of validity is required to be able to both influence adequate instrument development and help reviewers and clinicians recognize sources of validity evidence to better integrate for an overall understanding of the validity of a test and its scores. Specific ways to test for these sources of validity are provided throughout this special issue.

Evidence Based on Assessment Content

Although evidence based on content might seem like one of the simplest forms of validity to address, it might not be so easy. Evidence of content validity has evolved over time and has more recently been debated as a form of validity evidence. If validity refers to the interpretation of scores, the belief is that the content itself does not equate an assessment score; yet, without content validity, other sources of validity might suffer. In its simplest form, content validity refers to the degree to which the content of an assessment is consistent with and represents the intended construct. This includes a clear operational definition of the domain measured by the assessment, individual assessment items that are relevant to the domain, and all aspects of the domain intended to be assessed are accurately and adequately measured (Sireci & Faulkner-Bond, 2014). To be lacking in content validity would mean that an assessment does not measure what it claims to measure. This could equate to construct underrepresentation, irrelevance to a particular subgroup of individuals, or missing the mark on the construct altogether. The implications of lacking content validity are perilous, as it could mean inadequate treatment, misdiagnoses, or providing certifications and licensure to individuals who are not adequately prepared.

Evidence of validity based on assessment content is typically assessed through the use of content or subject matter experts. Asking content experts to engage in ranking or rating of items to determine the degree to which they represent the domain specified or asking them to match items to the section or domain they believe the item to represent are ways to provide evidence of validity based on assessment content. Complementary quantitative strategies, such as Lawshe's content validity ratio (Ayre & Scally, 2014; Lawshe, 1975), have been developed to estimate the consensus among content experts and facilitate developing a corpus of items that has a high probability of representing the intended constructs. Although content validity might seem simple, it is an important component of the quality of an assessment, as ultimately all other forms of validity hinge on the quality and strength of the content of an assessment. Understanding content of an instrument can also be connected to understanding individuals' response process to each item and the assessment as a whole.

Evidence Based on Response Processes

Compared to the long-standing professional discussions of what we consider the more traditional forms of validity (e.g., content, internal structure), evidence of validity based on response processes is a newer form of validity evidence, included in the *Standards* for the first time in 1999. Prior versions of the standards referred to this source of validity as related to construct validity (AERA, APA, & NCME, 1985). Cizek, Rosenberg, and Koons (2008) found that this source of validation is explored and stated in less than 2% of the validation studies of assessments. However, this might be due to it being a more newly discussed source of validity, or the confusion around how to explore and analyze response processes, than a matter of professional negligence. Evidence based on response process refers to “the fit between the construct and the detailed nature of the performance or response actually engaged in by the test takers” (AERA, APA, & NCME, 2014, p. 15). This equates to focus and understanding of the attention, perception, language, knowledge, and higher reasoning of individuals as they respond to items on an assessment. This includes whether test takers might be guessing on items on the assessment, misunderstanding a word or concept, or whether something else (e.g., anxiety,

boredom, social desirability) could be influencing individuals while responding to items on the assessment.

Validation studies exploring the response process have found that having difficulty understanding a concept in the assessment can affect assessment reliability and internal structure. Having a response process, such as social desirability, anxiety, or inattentiveness, can alter what is being measured—or the construct of the assessment; it can therefore be understood how earlier versions of evidence based on response process were subsumed under the construct validity in the standards. Assessing for evidence related to response process is most frequently explored through cognitive and in-depth interviews, but has sometimes been done through observations and eye tracking, as well as response time (Padilla & Benitez, 2014). Knowing the evidence related to response process of a measure might alter how a researcher or clinician uses the measure or interprets the scores. Given that response process can affect scores on an assessment, from mild interference with responses to specific items to altering the entire construct measured, exploring and understanding the response process among each subgroup of individuals is essential. Response process can ultimately affect the internal structure of an assessment.

Evidence Based on Internal Structure

The extent to which items on the assessment interrelate and correspond to the framework of the presumed construct provide evidence on internal structure. There are three facets of internal structure, including dimensionality, measurement invariance, and reliability. Dimensionality refers to whether the scores should be unidimensional, or a multifactor or bifactor model. This can affect whether the assessment is compiled into one overall score or multiple scores, or can be used in both ways (having multiple subscale scores while also having one overall composite score). Consider the example of the Toronto Alexithymia Scale (TAS–20; Bagby, Parker, & Taylor, 1994), with which clinicians can use the overall total score to determine the degree to which a client exhibits alexithymia, but can also break the TAS–20 down into three subscales: difficulty describing feelings, difficulty identifying emotions, and externally oriented thinking. Dimensionality of the assessment can affect the intended use, scoring, and interpretation of the assessment.

Measurement invariance can provide information about individual items and how they might function similarly or differently across subgroups of clients. This differential functioning could be unintentional concerns with content or cognitive and response processes, or could be intentional and appropriate, but needs to be understood and examined. Finally, the aspect of reliability refers to the consistency of assessment scores across repeated administration, indicating that responses from participants on the items remain similar across time. Each of these aspects is connected to an assessment's internal structure, and affects overall validity because it affects how we assess and interpret scores on an assessment, which in turn impact our students and clients. Evidence of internal structure is typically provided through factor analysis and goodness-of-fit models, exploring assessment scores and item responses across subgroups of individuals, as well as correlations between multiple iterations of the assessment. Both measurement invariance and misuse of dimensions on an assessment can result in inaccurate assessment scores, which then create a situation that could alter treatment decisions. Thus, the internal structure of an assessment is influential in the final scores received and subsequent

interpretation. This in turn would affect the previous forms of validity evidence mentioned, but also relations to other variables.

Evidence Based on Relations to Other Variables

Evidence based on relations to other variables is typically the most mentioned and tested aspect of validity provided on assessments. This source of validity helps to validate the construct being measured and can affect diagnoses and types of treatments provided, or adjustments to educational curriculum. It is usually represented by discussions of convergent, discriminant, criterion, and predictive validity, all of which provide evidence substantiating that the construct purported to be measured by the assessment is in fact being measured. This is typically tested using inferential statistics such as correlations and regressions with scores from other assessments or behavioral observations.

To know that an assessment measures a particular construct, or can predict future behavior, is a foundation that is depended on within clinical work. As an example, although depression is known to be related to suicidal behaviors, whether a depression inventory can actually predict suicidal behavior is another question. Through validation studies, it was found that assessments, such as the Beck Depression Inventory, can in fact be used as valid assessments to predict future suicide attempts (Desseilles et al., 2012). It is through exploring the relationships between current and future behaviors and scores on assessments that we can provide evidence of the construct, as well as the appropriate clinical use of an assessment in practice. Without this source of validity evidence, we would not be able to clearly affirm if the assessment measured the stated construct, which has serious ramifications for clinical diagnoses and treatment protocols. As you could assume, not measuring the stated construct would have ramifications for the next source of validity discussed, consequential validity.

Evidence for Validity and Consequences of Testing

When considering validity, most thoughts and attempts are focused on the development of the assessment, such as item creation, evidence of content, internal structure, and the other sources of validity evidence already mentioned. However, one source of validity evidence less focused on, but exceedingly important, is the evidence of the consequences of testing, or consequential validity. Cronbach (1988) argued for the importance of consequential testing when he stated that negative consequences from the interpretation of scores on an assessment should invalidate the assessment, even if those consequences were not from any flaws in the assessment itself. Evidence of validity and consequences of testing refers to the soundness of the proposed interpretation and use of assessment scores. A focus on the consequences of an assessment is important, as assessments are used not only to gather information, but to inform clinical treatment, pharmaceutical treatment, and diagnosis, to mention only a few.

Three criteria for evaluation of consequential validity were noted by Kane (2006), including clarity, coherence, and plausibility. Thus, validation of the consequences of testing entails a clear statement of the proposed interpretation and uses of the assessment scores, an evaluation of the interpretation and uses provided with supporting evidence, and a statement and exploration of potential alternative consequences from the interpretation and use of the assessment

(Lane, 2014). Consequences of an assessment could be intended or unintended, but all must be considered and explored. For example, when creating a scale to test suicidal behaviors, the intended use of this assessment might be to identify risk level of engaging in or attempting suicide. The assessment developer might hope to be able to categorize individuals into no risk, slight risk, moderate risk, and high risk categories to assist counselors in making informed clinical decisions regarding immediate and future treatment. One intended consequence or use of this assessment might be to identify risk level and determine whether the client needs to be admitted to an inpatient treatment facility to stabilize or adjust medication, versus being sent home under the care and support of family until the next scheduled counseling session. However, an unintended consequence might be that the individual is stigmatized due to risk level of suicide, treated differently by family members and friends, or taken less seriously due to low level of risk of suicidal behavior potentially resulting in increases in self-harm behaviors.

Lane (2014) provided three questions to consider in validation of consequences of assessments: (a) What does the assessment proclaim to do? (b) What are the arguments for and against the intended claims? (c) What are the unintended outcomes or consequences (both positive and negative) of the assessment scores? The answers to these questions can provide evidence of consequential validity by providing clarity of the intended use and interpretation of the assessment, logical links and evidence of how these interpretations can lead to decisions and actions, and empirical evidence that the interpretations and actions are credible and fall within the scope of the intended use of the assessment.

Although each of these sources of validity can be explored discretely, it is the integration of these sources that provides a sound validity argument. Each source of validity usually cannot be explored or answered within one study, but this is more of a long-term process that ultimately never ends (AERA, APA, & NCME, 2014). Information can be gleaned from validation studies in the development of the assessment, but also inferences can be drawn from each subsequent study that uses the assessment. These sources of evidence assimilate into the unitary concept of validity. Having knowledge of these sources allows researchers and counselors to use assessments with confidence, while denouncing the use of invalid, haphazardly created assessments.

Rationale and Organization of Special Issue Content

This special issue of *MECD* has been prepared as a support to educators, clinicians, and scholars whose professional activities could be supported through clear, plainly spoken depictions of strategies that support inferences about the degree of validity associated with a particular assessment. Such an endeavor is warranted when considering that many master's-level counselors might have received only one course in advanced psychometrics and doctoral-level counselor educators often rely on exemplars from related professions to guide best practice. Therefore, this special issue is intended to provide a practical support for master's-level counselors and doctoral-level counselor educators who are interested in the theories and methods that will not only support contributions to the knowledge base available to within *MECD*, but also support increased precision in the measurement of counseling outcomes by practitioners and scholars alike.

Readers will find three articles supporting the development of psychological and educational assessments through developing content-oriented evidence (Lambie, Blount, & Mullen, 2017/this issue) and use of cognitive interviewing (Peterson, Peterson, & Powell, 2017/this issue), as well as issues associated with translations and cross-cultural validation (Lenz, Gómez Soler, Dell'Aquilla, & Uribe, 2017/this issue). Six articles depict strategies for the evaluation of assessment validity and precision through exploratory factor analysis (Watson, 2017/this issue), confirmatory factor analysis (Lewis, 2017/this issue), Rasch methodology (Willse, 2017/this issue), estimating reliability and precision of scores (Bardhoshi & Erford, 2017/this issue), and strategies for establishing evidence with conceptually related variables (Balkin, 2017/this issue; Swank & Mullen, 2017/this issue). Two articles support the evaluation of fairness of assessment practices and interpretations through inspecting construct irrelevance and construct underrepresentation (Spurgeon, 2017/this issue) and use of assessment scores with individuals not represented in the normative sample (Hays & Wood, 2017/this issue).

Conclusions

Although we, and the authors of each article, made efforts to provide language and processes that are easily accessible and implemented by counselors in all settings, the actual implementation of the information from this special issue is up to each individual counselor. Of course, assessments can continue to be downloaded from smartphone apps and the Internet without exploring the validity. However, as noted, this results in decisions being made that can have perilous ramifications due to inappropriate interpretations. Yet, taking the time to explore sources of validity evidence—whether developing a measure or looking to use an existing measure—can result in the appropriate use of assessments and interpretation of scores. We recommend that researchers use this special issue to better understand not only how to develop assessments, but how to critique assessments they would like to use in their studies so that the knowledge they are putting forth into the field is accurate due to having assessments that yield valid scores; clinicians use this special issue to evaluate the contexts in which assessments can be used appropriately to make diagnostic and treatment decisions given the demographics and circumstances of their clients; and students use this special issue to enhance their understanding of the importance of validity so they can employ this knowledge throughout their future careers.

Acknowledgments

The guest editors would like to emphasize that this special issue would not have been possible without the dedicated knowledge and expertise provided by each author who contributed to this issue, as well as the reviewers who devoted time and effort to providing vital feedback that resulted in what we believe to be noteworthy contributions to the field of understanding validity and reliability.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: APA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*, 79–86. https://doi.org/10.1177/0748175613513808

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale-I: Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*, 23–32. https://doi.org/10.1016/0022-3999(94)90005-1

Balkin, R. S. (2017). Evaluating evidence regarding relationships with criteria. *Measurement and Evaluation in Counseling and Development*, *50*, 264–269. https://doi.org/10.1080/07481756.2017.1336928

Bardhoshi, G., & Erford, B. T. (2017). Processes and procedures for estimating score reliability and precision. *Measurement and Evaluation in Counseling and Development*, *50*, 256–263. https://doi.org/10.1080/07481756.2017.1388680

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*, 732–743. https://doi.org/10.1177/0013164410379323

Cizek, G. J., Koons, H. H., & Rosenberg, S. L. (2011). Finding validity evidence: An analysis using *The Mental Measurements Sourcebook*. In J. A. Bovaird, K. F. Geisinger & C. W. Buckendahl (Eds.), *High stakes testing in education: Science and practice in K-12 settings* (pp. 264–278). Washington, DC: American Psychological Association.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397–412. https://doi.org/10.1177/0013164407310130

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

Desseilles, M., Perroud, N., Guillaume, S., Jaussent, I., Genty, C., Malafosse, A., & Courtet, P. (2012). Is it valid to measure suicidal ideation by depression rating scales? *Journal of Affective Disorders*, *136*, 398–404. https://doi.org/10.1016/j.jad.2011.11.013

Drummond, R. J., Sheperis, C. J., & Jones, K. D. (2016). *Assessment procedures for counseling and helping professionals* (8th ed.). Columbus, OH: Pearson.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new *Standards for Educational and Psychological Testing*: Implications for measurement courses. *Measurement*

and Evaluation in Counseling and Development, *36*, 181–191. https://doi.org/10.1080/07481756.2003.11909741

Hattie, J. (2014). The last of the 20th-century test standards. *Educational Measurement: Issues and Practice*, *33*(4), 34–35. https://doi.org/10.1111/emip.12053

Hays, D. G., & Wood, C. (2017). Stepping outside the normed sample: Implications for validity. *Measurement and Evaluation in Counseling and Development*, *50*, 282–288. https://doi.org/10.1080/07481756.2017.1339565

Hogan, T. P, & Agnello, J. (2004). An empirical study of reporting practices regarding measurement validity. *Educational and Psychological Measurement*, *64*, 802–812. https://doi.org/10.1177/0013164404264120

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kelly, A. P., O'Malley, K. J., Kallen, M. A., & Ford, M. E. (2005). Integrating validity theory with use of measurement instruments in clinical settings. *Health Research and Educational Trust*, *40*, 1605–1619. https://doi.org/10.1111/j.1475-6773.2005.00445.x

Lambie, G. W., Blount, A. J., & Mullen, P. R. (2017). Establishing content-oriented evidence for psychological assessments. *Measurement and Evaluation in Counseling and Development*, *50*, 210–216. https://doi.org/10.1080/07481756.2017.1336930

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*, 127–135. https://doi.org/10.7334/psicothema2013.258

Laswshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lenz, A. S., Gómez Soler, I., Dell'Aquilla, J., & Uribe, P. M. (2017). Translation and cross-cultural adaptation of assessments for use in counseling research. *Measurement and Evaluation in Counseling and Development*, *50*, 224–231. https://doi.org/10.1080/07481756.2017.1320947

Lewis, T. F. (2017). Evidence regarding the internal structure: Confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, *50*, 239–247. https://doi.org/10.1080/07481756.2017.1336929

Markus, K. A. (2016). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy & Practice*, *23*, 252–267. https://doi.org/10.1080/0969594X.2015.1060191

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word “validity” and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, *23*, 178–197. https://doi.org/10.1080/0969594X.2015.1037241

Padilla, J. L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*, 136–144. https://doi.org/10.7334/psicothema2013.259

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development*, *50*, 217–223. https://doi.org/10.1080/07481756.2017.1339564

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*, 100–107. https://doi.org/10.7334/psicothema2013.256

Spurgeon, S. L. (2017). Evaluating the unintended consequences of assessment practices: Construct irrelevance and construct underrepresentation. *Measurement and Evaluation in Counseling and Development*, *50*, 275–281. https://doi.org/10.1080/07481756.2017.1339563

Swank, J. M., & Mullen, P. R. (2017). Evaluating evidence for conceptually related constructs using bivariate correlations. *Measurement and Evaluation in Counseling and Development*, *50*, 270–274. https://doi.org/10.1080/07481756.2017.1339562

Watson, J. C. (2017). Establishing evidence for internal structure using exploratory factor analysis. *Measurement and Evaluation in Counseling and Development*, *50*, 232–238. https://doi.org/10.1080/07481756.2017.1336931

Wester, K. L., Borders, L. D., Boul, S., & Horton, E. (2013). Research quality: Critique of quantitative articles in the *Journal of Counseling & Development*. *Journal of Counseling & Development*, *91*, 280–290. https://doi.org/10.1002/j.1556-6676.2013.00096.x

Willse, J. T. (2017). Polytomous Rasch models in counseling assessment. *Measurement and Evaluation in Counseling and Development*, *50*, 248–255. https://doi.org/10.1080/07481756.2017.1362656

Author information

A. Stephen Lenz, PhD, LPC, is an Associate Professor of Counselor Education at Texas A&M University--Corpus Christi. He teaches and researches in the areas of holistic approaches to student development, assessment development, and counseling outcome research and program evaluation.

Kelly L. Wester, PhD, LPC, NCC, is a professor at the University of North Carolina at Greensboro. She teaches and researches in the areas of research training and ethics in counseling, along with nonsuicidal self-injury.