2024 ■ Vol. 20 ■ no. 3

Teaching Quantitative Methods Vignettes

A Short Tutorial on Validation in Educational and Psychological Assessment

Angel Arias [a]

[a]School of Linguistics and Language Studies, Carleton University

Abstract ■ Validity theory has evolved considerably over time, transitioning from a simplistic notion of what tests purport to measure to a more nuanced understanding encompassing the interpretation and uses of test scores. This latter understanding of validity places equal importance on how test scores are interpreted and how tests are used in society. Despite the central role validity plays in assessment, there exists a divergence of perspectives regarding the inclusion of test consequences in validation processes. While some argue for a focus solely on the inference of score meaning, others advocate for a comprehensive consideration of test use implications. This tutorial aligns with the latter perspective, drawing heavily on the Standards for Educational and Psychological Testing and socioculturally sustaining assessment. A pedagogical activity is presented through concept maps to enhance understanding of validity and validation among graduate students in social sciences, humanities, and health sciences. The activity consists of three scaffolded steps: (1) a pre-teaching concept map, (2) a gap-filling concept map, and (3) the development of culturally sustaining perspectives for test validation and guiding research questions to address sources of validity evidence..

Keywords ■ validity, culturally sustaining validation, concept mapping, research methods, sources of validity evidence. **Tools** ■ CmapTools.

angel.arias@carleton.ca

10.20982/tqmp.20.3.v024

Acting Editor ■ Sébastien Béland (Université de Montréal)

Reviewers ■ One anonymous reviewer

Concept to be presented

Measurement validity theory (hereafter, validity) has been described as the most fundamental concept and important process in educational measurement (Randall et al., 2022; Sireci, 2009). It is a fundamental concept and an important process because it provides the foundation and guidelines to evaluate test quality and the role tests play in society. Validity has evolved considerably from the notion that a test is valid if it measures what it purports to measure (Buckingham, 1921) to the notion that it [validity] "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of test" (American Educational Research Association et al., 2014, p. 11, hereafter, the *Standards*). The former definition views validity as an inherent property of a test, whereas the latter defines validity as a characteristic related to the interpretation and uses of test scores. Although this distinction may appear trivial, it is crucial to underscore that for a test to be deemed useful for a specific purpose, supporting that it effectively measures its intended construct(s) is not sufficient to justify test use (Arias & Sireci, 2021; Sireci, 2016).

Consequently, evaluating such a claim constitutes a pivotal aspect of the validation process because test score use always entails some form of action or consequence (e.g., university admissions, accreditation, migration purposes, etc.). A closely related concept to validity is validation, which is "the process of constructing and evaluating arguments for and against the interpretation of test scores and their relevance to the proposed use (American Educational Research Association et al., 2014, p. 11). The *Standards* is an evolving document that has been subjected

to several iterations to reflect advances in validity theory and validation. This document also represents the consensus definition of validity within assessment-centred disciplines. The *Standards* offer important guidelines to consider when building validity arguments for test score interpretation and use and recommend gathering validity evidence based on (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing. This positionality of validity constitutes a long-standing collaboration that has prevailed since the 1950s (cf. American Psychological Association, 1952). Although the *Standards* have developed and evolved in the United States, their international relevance has been acknowledged and supported (Sireci, 2016; Zumbo, 2014).

Although the existence of a living document such as the *Standards* might imply a shared understanding of validity and validation processes across assessment-centred disciplines and among validity theorists, this assumption is, in reality, a chimera –a fire-breathing "concept" with a lion's head, a goat's body, and a serpent's tail. This does not necessarily represent a weakness or flaw in validity and validation. Instead, it suggests a rich and ongoing scholarly dialogue on such an important concept in evaluating the quality of tests and the consequences of their use in society. The different views on the role consequences and test use should play in validation endeavours have spurred a key disagreement on what validity entails. Several scholars have argued that validity should solely concern test score meaning and inferences, advocating for the exclusion of any repercussions caused by test use from validity considerations. They argue that including such repercussions introduces clutter that hinders clarity, making it challenging to harmonize evaluative judgments (Cizek, 2012; Cizek, 2020; Mehrens, 1997; Popham, 1997). Furthermore, Cizek (2020) introduced a framework titled "Comprehensive Framework for Defensible Testing," in which the validity of score meaning and justification of test use are addressed but fall outside the explicit scope of validity.

However, examining test use and its consequences is an integral aspect of the testing process, and validity theorists have strongly advocated for the inclusion of these factors in validation research (cf. Kane, 2013; Messick, 1989; Sireci, 2016; Zumbo, 2023). As educators and psychologists, "it is rare that anyone measures for the sheer delight one experiences from the act itself. Instead, all measurement is, in essence, something you do so that you can use the outcomes" (Zumbo, 2009, p. 60). In keeping with this perspective, Sireci (2016, p. 231) asked, "Can we have test interpretation without test use?" He explains that it could be possible, but why would we develop tests that will never be used for a practical purpose? This positionality is consistent with the *Standards* and variations of argument-based approaches to validation (cf. Chapelle, 2021; Kane, 2006; Kane, 2013).

Because test use and consequences play an important role in assessments and complement score-based inferences, this tutorial on validity and validation adopts the stance in the *Standards for Educational and Psychological Testing* and argument-based approaches to validation. That is, validity refers to the interpretation and uses of test scores, and validation consists of gathering validity evidence for both the interpretation and uses of scores. For simplicity of operationalization, this vignette draws heavily on the *Standards* to illustrate how to teach this concept to graduate students in social sciences, humanities, and health sciences who may be tasked with constructing validity arguments for measurement instruments. As such, the application activity will be based on the five sources of validity evidence: (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing.

Validating the interpretation and uses of test scores in the 21st century requires validators to consider the context in which tests are ultimately used seriously. The definitions of validity and validation have coexisted and have developed in contexts of white supremacist hegemony that have overlooked shifts in educational landscapes, such as the growing presence of migrant students and the increasing diversity of individuals taking major, life-changing, high-stakes tests and responding to survey items. In turn, this promotes dysconcious racism and exacerbates racial inequalities and disparities in education (Randall et al., 2022). King (1991, p. 135) describes dysconscious racism as "a form of racism that tacitly accepts dominant white norms and privileges." When embarking on validation research, it is highly recommended that validation frameworks be operationalized consistently with the targeted population for which the test score interpretation and uses are made. This implies drawing on transdisciplinarity to enhance the quality of the required validity evidence. For instance, sociocultural-oriented assessment (Bennett, 2023), asset-based pedagogies (Flint & Jaggers, 2021), and critical race theory (Delgado & Stefancic, 2017) can potentially inform validation research. Thus, when embarking on test development, constructs need to be defined ecologically to encompass the inherent characteristics of the target examinees and the context in which the assessment instrument is ultimately used. In turn, validation frameworks should also address the culturally sustaining aspects of the assessment instrument to evaluate the usefulness and validity of test score-based interpretations (cf. Randall, 2021; Randall et al., 2022, 2024). The following section introduces the application activity and attempts to address validation from a culturally sustaining perspective (Paris, 2012).

Table 1 ■ Sources of validity evidence and respective research questions

Source of validity evidence	Culturally sustaining perspectives	Guiding research question
Test content	e.g., To what extent does the test include construct relevant features that are important and embedded within the cultures and values of the targeted examinees?	e.g., To what extent are test items congruent with the specified knowledge, skills, and abilities of the targeted construct?
Response processes		
Internal structure		
Relation to other variables		
Consequences of testing		

Pedagogical strategy

Prior to this activity, the chapter on validity in the *Standards*,[1] Randall (2021), and Randall et al. (2022) must be assigned as required readings. Because validity and validation can be abstract concepts, teaching how to gather validity evidence to interpret and use test scores for a particular purpose can be daunting. Readings on this concept can be very dense (e.g. Messick, 1989), and prospective validators can often feel overwhelmed. Validation research cannot be conceptualized in terms of a one-off study. Instead, it is the joint responsibility of the test developer and the test user, and it is a very intensive research program (Newton & Shaw, 2013). Nevertheless, validation research needs to be accessible to scholars who do not necessarily specialize in this area of measurement. The activity presented in this vignette includes three parts: (1) a pre-teaching concept map example on key concepts in validity theory, (2) a gap-filling concept mapping task, and (3) the development of culturally sustaining perspectives in test validation and guiding research questions. Parts 1 and 2 serve as the foundation for Part 3.

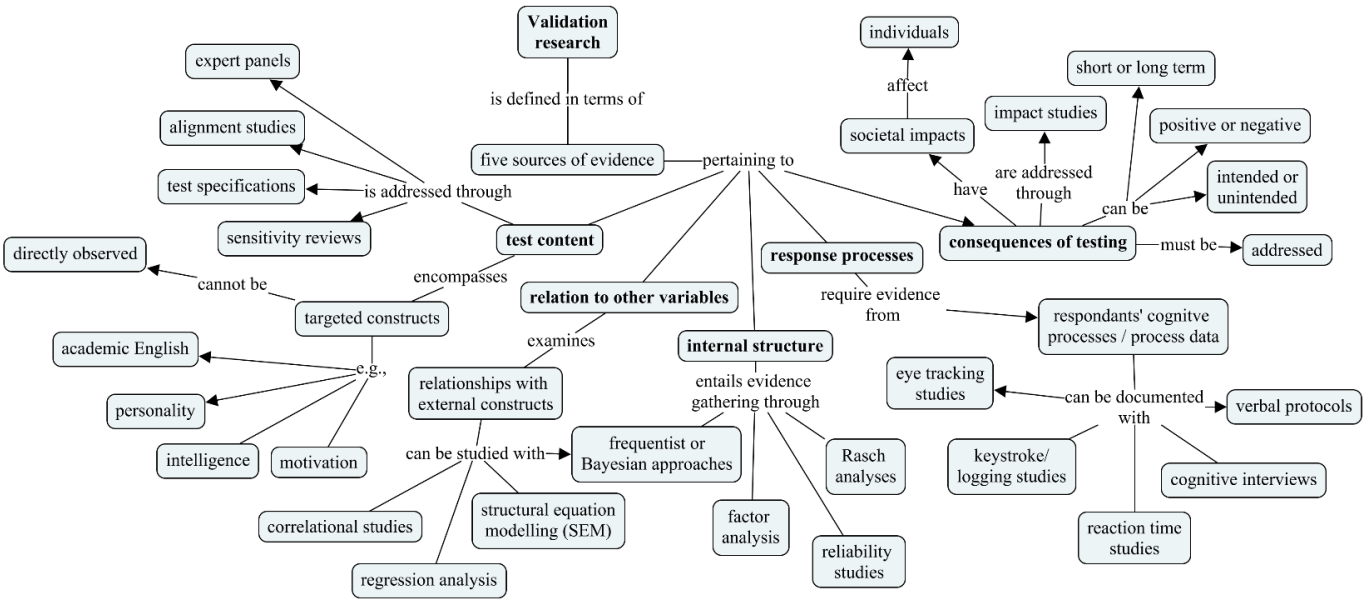
This tutorial draws on concept mapping to develop the application activity and to illustrate how this strategy can enhance learning abstract concepts such as validity and validation. Concept maps are diagrams with labelled nodes representing concepts and links denoting concept relationships (Nesbit & Adesope, 2006). Concept mapping "involves defining the topic, identifying, and listing the most important or "general" concepts that are associated with that topic, ordering the concepts from top to bottom in the mapping field, and adding and labelling linking phrases" (Cañas et al., 2003, p. 8). Concept mapping requires students to identify the connections between concepts and utilizes dual coding, where students learn from the map's text labels and visual structure. For some students, this technique can reduce cognitive load by allowing them to focus on essential relationships rather than decoding written text.

Moreover, concept mapping enhances long-term retention in the repeated study of texts and serves as an effective task when used as a retrieval-based learning activity (Blunt & Karpicke, 2014). In this regard, scenario-based assessment situations will be provided to contextualize the activity for students. Figure 1 provides Part 1 of the application activity. This concept map example illustrates the connections between key validity and validation concepts. Ideally, include this figure as a slide on a validity lecture before having students complete Parts 2 and 3 of the activity.

Once students have examined the sample concept map, have them work in groups to complete Part 2. They will complete the concept map found in Figure 2 using the list of concepts in the box. This will assess their ability to link

[1]The 2014 edition of the *Standards* is open access and can be retrieved from www.testingstandards.net/open-access-files.html. The *Standards* are periodically subjected to revisions and the latest edition should be used for up-to-date pedagogical purposes.

Figure 1 ■ Concept Map on Validity Research in Educational and Psychological Assessment



important concepts in validation with their corresponding definitions. Encourage students to use the validity chapter of the *Standards* to complete this part of the activity.

Part 3 of the activity requires students to develop culturally relevant perspectives of the testing situation and guiding research questions to address each of the sources of validity evidence in the *Standards*. Instruct students to use the assessment situations in Appendix A to develop the culturally sustaining perspectives and the guiding research questions of a validation endeavour for each of the sources of validity evidence proposed in the *Standards*. Note that students or instructors can develop other scenarios to fit the class context or interest. Use Parts 1 and 2 as scaffolding steps to complete Part 3. An example of a culturally relevant perspective and a guiding question are provided in Table 1.

It is recommended that Part 3 of the activity be completed in groups. The instructor should assist as needed. Gathering validity evidence can be a daunting task for the neophyte student. This tutorial is addressed to graduate students in assessment-centred programs who have enrolled in research methods classes (both qualitative and quantitative).

Conclusion

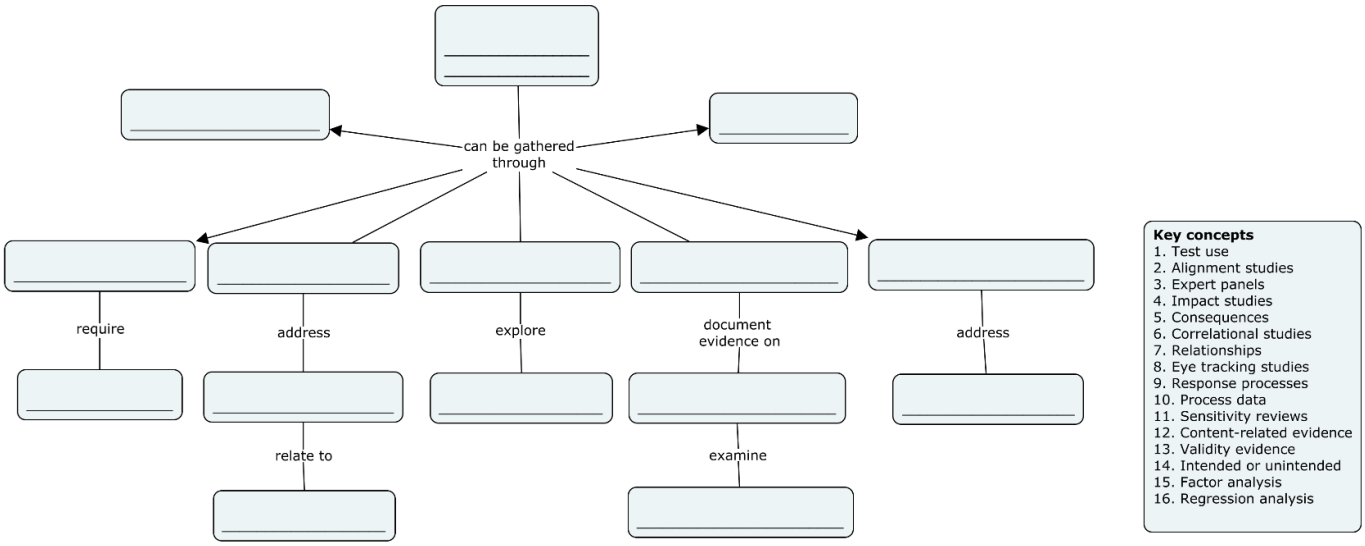
This concise tutorial on validity briefly summarized what the concept means for assessment-centred disciplines and proposed a three-step activity to teach and evaluate critical concepts and processes in validation research. Validation efforts require test developers and users to center consequences in evidence-gathering endeavours. The *Standards* (AERA et al., 2014) provide a theoretical framework to address validation research in high-stakes contexts. Five sources of validity evidence should be addressed when building a validity argument for the interpretations and uses of test scores. These sources of evidence are based on (1) test content, (2) response processes, (3) internal structure, (4) relation to other variables and (5) consequences. In order to decolonize and democratize validation research, the field needs to draw on relevant perspectives from other sources of knowledge such as critical race theory (Delgado & Stefancic, 2017), ecological systems theory (Bronfenbrenner, 1979; Zumbo, 2023), cultural validity (Basterra et al., 2011), and critical positionalities on validation frameworks (Randall, 2021; Randall et al., 2022, 2024).

References

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

American Psychological Association. (1952). Committee on test standards. technical recommendations for psychological tests and diagnostic techniques: Preliminary proposal. *American Psychologist*, 7(8), 461–475.

Figure 2 ■ Gap-filling Concept Map for Key Terms in Validation Research



Arias, A., & Sireci, S. G. (2021). Validez y validación para pruebas educativas y psicológicas: Teoría y recomendaciones. *Revista Iberoamericana de Psicología*, *14*(1), 11–22. doi: 10.33881/2027-1786.rip.14102.

Basterra, M., Trumbull, R., & Solano Flores, G. (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity* (Routledge, Ed.).

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessement*, *28*(2), 83–104. doi: 10.1080/10627197.2023.2202312.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. doi: 10.1037/a0035934.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Harvard University Press.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium–XIV. *Journal of Educational Psychology*, *12*(5), 271–275. doi: 10.1037/h0066019.

Cañas, A. J., Coffey, J. W., Carnot, M. J., Feltovich, P., Hoffman, R. R., Feltovich, J., & Novak, J. D. (2003). *A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support*. Institute for Human; Machine Cognition.

Chapelle, C. A. (2021). Argument-based validation in testing and assessment.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge, Taylor & Francis Group.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43. doi: 10.1037/a0026975.

Delgado, R., & Stefancic, J. (2017). *Critical race theory: An introduction*. New York University Press.

Flint, A. S., & Jaggers, W. (2021). You matter here: The impact of asset-based pedagogies on learning. *Theory into Practice*, *60*(3), 254–264. doi: 10.1080/00405841.2021.1911483.

Kane, M. T. (2006). Validation. In *Educational measurement* (4 ed., pp. 17–64). American Council of Education; Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi: 10.1111/jedm.12000.

King, J. E. (1991). Dysconscious racism: Ideology, identity, and the miseducation of teachers. *Journal of Negro Education*, *60*(2), 133. doi: 10.2307/2295605.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurements: Issues and Practice*, *16*(2), 16–18. doi: 10.1111/j.1745-3992.1997.tb00588.x.

Messick, S. (1989). *Validity* (R. L. Linn, Ed.; 3rd ed.). American Council on Education; Macmillan.

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76*(3), 413–448. doi: 10.3102/00346543076003413.

Newton, P. E., & Shaw, S. D. (2013). *Validity in educational and psychological assessment*. SAGE Publications.

Paris, D. (2012). Culturally sustaining pedagogy. *Educational Research*, *41*(3), 93–97. doi: 10.3102/0013189X12441244.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9–13. doi: 10.1111/j.1745-3992.1997.tb00586.x.

Randall, J. (2021). "color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practices*, *40*(4), 82–90. doi: 10.1111/emip.12429.

Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2024). Our validity looks like justice. does yours? *Language Testing*, *41*(1), 203–219. doi: 10.1177/02655322231202947.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170–178. doi: 10.1080/10627197.2022.2042682.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Information Age Publishing.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education*, *23*(2), 226–235. doi: 10.1080/0969594X.2015.1072084.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). Information Age Publishing.

Zumbo, B. D. (2014). What role does, and should, the test *standards* play outside of the united states of america? *Educational Measurement: Issues and Practice*, *33*(4), 31–33. doi: 10.1111/emip.12052.

Zumbo, B. D. (2023). A dialectic on validity: Explanation-focused and the many ways of being human. *International Journal of Assessment Tools in Education*, *10*(Special Issue), 1–96. doi: 10.21449/ijate.1406304.

Appendix A: Assessment Situations and Handout

Assessment situation for health sciences programs

Construct: Nutritional self-efficacy can be defined as a measurable belief system regarding one's perceived ability to engage in various aspects of healthy eating behaviours. Assessment of nutritional self-efficacy typically involves evaluating an individual's confidence in their capability to perform specific tasks related to diet and nutrition.

1. Test Content

For assessing nutritional self-efficacy, validity evidence based on test content would ensure that the assessment instrument includes items that capture various aspects of individuals' confidence in making healthy food choices. This might include questions about food selection, meal planning, portion control, and overcoming dietary challenges.

2. Response Processes

Conducting cognitive interviews with individuals who represent the target population for the assessment can provide insights into how they interpret and respond to the items. For example, participants might provide feedback on whether the language used in the items is clear and whether the scenarios presented in the questions resonate with their real-life experiences.

3. Internal Structure

Statistical analyses, such as factor analysis, can help determine whether the items on the assessment instrument cluster together as expected to form coherent dimensions of nutritional self-efficacy. For instance, factor analysis might reveal underlying factors such as confidence in meal planning versus confidence in resisting unhealthy food temptations.

4. Relations to Other Variables

Correlational analyses can be conducted to explore the relationships between scores on the nutritional self-efficacy assessment and relevant variables, such as dietary behaviour or health outcomes. For instance, individuals with higher scores on the assessment might demonstrate healthier eating habits or have better outcomes related to chronic disease management.

5. Consequences of Testing

Studies can be designed to examine the potential impact of using the nutritional self-efficacy assessment on individuals and society. For example, researchers might investigate whether participation in a nutrition education program tailored to individuals' self-efficacy levels leads to improvements in dietary habits and overall health outcomes.

Assessment situation for psychology programs

Construct: Personality encompasses the unique combination of enduring traits, behaviours, and cognitive patterns that define an individual's distinct identity. It encompasses how individuals perceive, think, feel, and behave across various contexts, reflecting their consistent tendencies and reactions over time. These traits, including extraversion, conscientiousness, openness, agreeableness, and neuroticism, shape individuals' interactions with others, their choices, and their responses to life's challenges. Personality is influenced by a complex interplay of environmental factors and life experiences, contributing to the diversity of human behaviour.

1. Test Content

Content validity for a personality assessment involves ensuring that the items on the instrument adequately represent the breadth and depth of personality traits. This might include items related to the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) or other relevant dimensions of personality.

2. Response Processes

Cognitive interviews can be conducted with individuals to understand how they interpret and respond to the items on the survey questionnaire. Participants might provide feedback on whether the items accurately capture their personality traits and whether the response options are appropriate for expressing their thoughts and feelings.

3. Internal Structure

Factor models can be used to examine the internal structure of the personality assessment and determine whether the items cluster together to form coherent dimensions of personality. For example, factor analysis might reveal distinct factors representing different aspects of extraversion, such as sociability, assertiveness, and excitement-seeking.

4. Relations to Other Variables

Correlational analyses can explore the relationships between scores on the personality assessment and other relevant variables, such as job performance, interpersonal relationships, or mental health outcomes. For instance, individuals with higher scores on measures of extraversion might be more likely to excel in roles that require social interaction and communication skills.

5. Consequences of Testing

Studies can be conducted to examine the impact of personality assessment on individuals and organizations. For example, researchers might investigate whether personality assessment results are used to inform hiring decisions and whether employees' personality traits predict job satisfaction and turnover rates within an organization.

Citation

Arias, A. (2024). A short tutorial on validation in educational and psychological assessment. *The Quantitative Methods for Psychology*, *20*(3), v24–v31. doi: 10.20982/tqmp.20.3.v024.

Received: 12/02/2024 ~ Accepted: 20/05/2024

Extended activity metadata

Concept illustrated	Validity and validation	*Type of activity*	In-class with instructor
Prerequisite	Research methods	*Types of data*	Conceptual chapters
Co-requisite	Test development	*Computation by*	CmapTools
Suitable class size	10–30 students	*Duration*	45-60 minutes