Editorial

On Consequential Validity

Dragos Iliescu¹ and Samuel Greiff²

¹ Faculty of Psychology and Educational Sciences, University of Bucharest, Romania

² Institute of Cognitive Science and Assessment (COSA), University of Luxembourg, Luxembourg

Problem Statement and Intention of This Editorial

Our understanding of validity has evolved from what seemed at that time innovative, and which is now interesting merely from a historical perspective (Camara & Brown, 1995). Most of the standard steps in the evolution of assessments are well-known, taught to students, applied by professionals, and are everyday knowledge in the repertoire of psychometricians: from “validity is the correlation of test scores with some criterion” to “validity is the degree to which a test measures what it purports to measure,” then later to “all validity is construct validity” and finally to “validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA, APA, & NCME, 1999, p. 9). Sireci (2012) discussed these evolutions and broke them down into three eras of validity theory: the empirical era (starting around 1900 and focused primarily on criterion-related validity and factor analysis), the theoretical era (starting around 1920 and focused primarily on factor analysis and construct validity), and the practical era (starting around the 1970s and advancing the argument-based approach).

But some of the interesting and potentially disrupting proposals have not made the impact they could have. One of these is the concept of *consequential validity*. In this editorial, we draw attention to consequential validity by outlining what it proposes, and how this could have an impact on better assessments – encouraging the development of stronger research in this important direction, which *EJPA* would be thrilled to host.

Samuel Messick and His View on Validity

Consequential validity has been proposed by Messick (1989) as an integral part of the argumentation on the validity of a test, pointing to the need to investigate the

social consequences (both actual and potential) of testing. Although critics of the construct noted years later that “there is no agreement at present about the importance or meaning of the term consequential validity” (Mehrens, 1997, p. 17), we note here the exact definition proposed by Messick: consequential validity is the aspect of (construct) validity that “appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially regarding sources of invalidity related to issues of bias, fairness, and distributive justice” (Messick, 1995, p. 745).

Messick proposed two dimensions for what he called “the consequential basis of validity”. First, he proposed that test score labels need to have accurate descriptions, in order to guide interpretation by stakeholders. At a first glance, this is an integral part of a more classical approach to validity – after all, if a test should measure what it intends to measure, the test developer should first of all lucidly label what the test intends to measure. Messick was more concerned with oversimplification and professional slang in test score labeling and how this could produce the wrong understanding in stakeholders or other audiences. For instance, labeling a score “resilience quotient” may lead stakeholders to assume that resilience is unidimensional, despite the literature showing that it is a multidimensional construct. Second – and more important to our discussion here – he proposed that test developers and users should appraise the potential and actual consequences of using the respective test. This component was later extended by Shepard (1993, 1997) to include both intended and unintended consequences of test use.

Consequential validity has attracted different levels of attention depending on the specific fields. It has sparked little to no enthusiasm in clinical and occupational testing but has probably been more discussed in the educational domain than in any other domain of test usage (Koretz, 2008). In educational settings, it was connected to the consequences that testing in schools can have on policies regarding school funding, curriculum and instructional content, and others. To this day, consequential validity is, however, one of the concepts related to validity that is least

utilized by test developers and users (Cizek et al., 2008) and possibly still the most debated (Lees-Haley, 1996). When offering validity information on a test, most efforts of test developers will be focused on convergent information, that is, evaluating whether the proposed assessment samples the intended content domain (of knowledge, behaviors, processes) consistently, fairly, and authentically (Messick, 1989; Wiliam, 2000), but oftentimes completely ignoring consequential information. We argue that this may well be the effect of a definitional bias of psychometricians: it is the direct result of how we define “the test,” where we draw boundaries between the test and test-related processes, and for which of these components we, therefore, take responsibility. We believe that tests as a force for good in the world will improve if assessment researchers develop more awareness toward and dedicate more of their attention to aspects of consequential validity.

Examples of How Consequential Validity Can Be Conceived

There are many ways in which to conceive consequential validity. Let us look at two examples. As a first example, we consider achievement tests, which are used widely in educational settings (classrooms and schools). They generate important data for students, teachers, parents, and other stakeholders and are considered an integral part of educational management. A good test will measure student achievement correctly, covering the target content and generating reliable scores on the competence of the assessed student. These scores will be highly predictive for the intended individual outcomes, such as student grades in high-stakes exams, admission to higher education, or student dropout. This looks like the perfect picture of a valid test. But sometimes the administration of such a test will also produce other outcomes: negative feedback on student performance on a course may lead to demotivation, may lead that student to embrace other courses and areas of learning, or may lead to actual drop out from school; and the mere existence of that test may motivate teachers to teach to the specific test and not to the broader capacity that should be developed. This, in its entity, begs the question of whether the test should only be judged on its capacity to produce good data, or also on its consequences? If the test is also judged based on such consequences, then test developers will need to assume responsibility for these. If so, the very definition of “test” may shift to include variables and processes that lead to such consequences; in this case, individual feedback on test scores may become an integral part of how this achievement test is conceived.

As a second example, we consider cognitive ability and personality tests, which are widely used in occupational settings. They generate important data for decision-makers, such as hiring managers, and are one of the most important tools used in personnel selection. Imagine a consultant who sets up a selection system using such tests, based on the predictive validities for these tests computed on a sample of employees. Test scores are then used by hiring managers to make hiring decisions. One year later, the consultant engages in a follow-up study and investigates the validity of the selection system, only to find out that the validity of the decisions based on the test is far lower than the validity predicted based on predictor-criterion correlations. This could be due to the fact that hiring managers do not understand the test scores or reports well enough in order to build them into their decision making, and therefore either ignore some or all test scores, expose themselves selectively to test results, or misconstrue their meaning, thus leading to mis-decisions that could have been avoided. Should these tests only be judged on their capacity to show good correlations with the intended occupational criteria, or also on their actual consequences for the organization? If the tests will also be judged based on such consequences in terms of hiring decisions, then the consultant will need to assume responsibility for these. If so, the definition for “tests” in this case may shift to include lucid feedback reports to managers, and specialized training.

Consequential validity may be more easily integrated into discussions about validity if testing professionals adopt the understanding that validity as a concept does not apply to instruments but to inferences (Cizek et al., 2008; Messick, 1989). This is the very basis of the current understanding of validity (AERA, APA, & NCME, 2014), as an integrated collection of knowledge stemming from empirical evidence and theoretical reasoning, more specifically, “evidence and theory [that] support the interpretations of test scores for proposed uses of a test” (p. 11). In such an integrated corpus of knowledge, the decisions that appear as a result of, and are connected to the interpretation of test results need to be factored in. However, this embracing of the domain of inferences is, we should note, the crux for the most ardent critiques of consequential validity: that measurement quality and the inferences based on the data can be fundamentally different. As Mehrens (1997) put it: “This confounding of inferences about measurement quality with treatment efficacy (or decision-making wisdom) seems unwise to me.” (p. 17); he then goes on to describe how such confounding would be seen in the medical profession: if a medic takes the temperature of a patient, the decisions he/she will make based on this measurement are fundamentally in a different domain than the actual measurement.

Here, we wish to avoid this debate (albeit acknowledging its importance); instead, our angle is different and urges test developers and researchers to look at consequential validity because of the ever-changing nature of our definition of what a test is and because more and more elements that have an impact on consequences are adopted in this definition. Indeed, the definition of what we consider “the test” has evolved through time from narrow to increasingly encompassing. We adhere to a broader definition and encourage test developers to increase their awareness of components that can legitimately be considered an intrinsic part of the test. We believe that adherence to either a narrower or broader definition will shape psychometric practice by motivating test developers to assume responsibility for more aspects of test usage.

The traditional definition of what a test is has, unfortunately, been narrow, and has focused on the items: a procedure or method that comprises a set of standardized items (e.g., stimuli, questions, or tasks) that are scored in a standardized manner and are used to examine and possibly evaluate individual differences (e.g., emotions, cognitions, attitudes, knowledge, skills, abilities, competencies) (Anastasi & Urbina, 1997; Cronbach, 1990). Arguments have also been made in favor of a broader definition (Iliescu, 2017), encompassing among others: the technical manual, training materials for the training of test users (including the delivery of these materials or training sessions), certification for test users (including the actual certification scheme and process), test reports (both the design of the reports and development of the information technology system that generates the reports), protection of the test (both in terms of access of qualified professionals and protection of intellectual property), or how the test will be published and made available to test users. This list can certainly be extended to considerations regarding how feedback is given and received on a specific test, or how the results of a specific test are built into decisions for individuals, groups, communities, and society. Embracing such a broader definition of what “a test” is, will implicitly include aspects that relate to the consequences of testing.

What We Propose

We, therefore, propose that test developers and other researchers embrace their responsibility for these supplementary facets of a test and generate substantive research in these understudied areas. To this end, through this editorial, we raise awareness of the understudied topic of consequential validity. In *EJPA*, we explicitly welcome empirical studies on aspects of consequential validity, addressing for example:

1. Report development: how reports (automated or not) for psychological tests should be constructed for maximum effect on different stakeholders;
2. Feedback on test scores: what form of test feedback is more effective for specific outcomes, in specific stakeholders;
3. Individual consequences of testing: what the intended and unintended effects of testing an individual with a specific test are and how they can be amplified, respectively mitigated;
4. Social consequences of testing: how a specific test or testing program contributes to change in groups, communities, and society.

All these questions relate to the field of consequential validity: they discuss the effects of testing or elements that contribute to the effects of testing. Extant research on such aspects is unexpectedly scarce, and we strongly believe that more robust research on these and connected issues will increase our understanding of how testing generates positive change, and our capacity to develop the position of testing and assessment as a force for good in society.

References

- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. AERA.
- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. AERA.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th. ed.). Prentice Hall.
- Camara, W. J., & Brown, D. C. (1995). Educational and employment testing: Changing concepts in measurement and policy. *Educational Measurement: Issues and Practice*, 14(1), 5–11. <https://doi.org/10.1111/j.1745-3992.1995.tb00845.x>
- Cizek, G. J., Rosenbert, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68(3), 397–412. <https://doi.org/10.1177/0013164407310130>
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Harper Collins.
- Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations*. Cambridge University Press.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.
- Lees-Haley, P. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist*, 51(9), 981–983. <https://doi.org/10.1037/0003-066X.51.9.981>
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16–18.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). ACE & NCME.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances

- as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405–450). AERA.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–24. <https://doi.org/10.1111/j.1745-3992.1997.tb00585.x>
- Sireci, S. G. (2012, July 5). *What have we learned from 100 years of validity theory and test validation?* State-of-the-Art address delivered at the 8th Annual Conference of the International Test Commission, Amsterdam, The Netherlands.
- Wiliam, D. (2000). The meanings and consequences of educational assessments. *Critical Quarterly*, 42(1), 105–127.

Published online June 3, 2021

Dragos Iliescu

Faculty of Psychology and Educational Sciences
University of Bucharest
Sos. Panduri 90
050657 Bucharest
Romania
dragos.iliescu@fpse.unibuc.ro