Soc Indic Res (2011) 103:219–230
DOI 10.1007/s11205-011-9843-4

Validity and the Consequences of Test Interpretation and Use

Anita M. Hubley · Bruno D. Zumbo

Accepted: 23 February 2011 / Published online: 22 April 2011

Abstract The vast majority of measures have, at their core, a purpose of personal and social change. If test developers and users want measures to have personal and social consequences and impact, then it is critical to consider the consequences and side effects of measurement in the validation process itself. The consequential basis of test interpretation and use, as introduced in Messick's (Educational measurement, Macmillan, New York, pp. 13–103, 1989) progressive matrix model of unified validity theory, has been misunderstood by many measurement experts, test developers, researchers, and practitioners. The purposes of this paper were to (a) review Messick's unified view of validity and clarify his consequential basis of test interpretation and use, (b) discuss the kinds of questions evoked by value implications and social consequences and their role in construct validity and score meaning, (c) present a reframing of Messick's model and a new model of unified validity and validation, (d) bring the concept of multilevel measures under the same validation umbrella as individual differences measures, and (e) offer some thoughts and directions for more explicit consideration of value implications, intended social consequences, and unintended side effects of legitimate test interpretation and use. This paper has implications for the interpretation, use, and validation of both individual differences and multilevel measures in education, psychology, and health contexts.

Keywords Consequential validity · Early development instrument · Educational achievement · Psychological assessment · Testing · Multilevel measures · Social consequences · Test interpretation · Validity · Value implications · Values

> 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. 66)

A. M. Hubley (✉) · B. D. Zumbo
Department of ECPS, University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: anita.hubley@ubc.ca

As Zumbo's quotation above highlights, the vast majority of measures have, at their core, a purpose of personal and social change. Ultimately, measures used in research, testing, assessment, and evaluation have implications, or are used, for ranking, intervention, feedback, decision-making, or policy purposes. Explicit recognition of this fact brings the often-ignored and sometimes maligned concept of consequences out of the shadows. If you want measures to have personal and social consequences and impact, then you must evaluate the intended consequences and unintended side effects of measurement when validating the inferences and uses made from tests and measures.

According to Messick (1989), validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 13). Messick's unified model of validity has been endorsed by the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME 1999) and the consequences of testing were identified in the *Standards* as one of five sources of validity evidence. While many test developers and test users strive to provide and examine Messick's evidential basis for test interpretation and use, these same groups tend to fall short when it comes to explicitly thinking and writing about the consequential basis of test interpretation and use. For example, a review of the 2005 *Mental Measurements Yearbook* (a leading resource that provides evaluative summaries of validity evidence for hundreds of published instruments) found that validity evidence reported in the literature has been gathered using outmoded frameworks and that some sources of validity evidence (including consequences) were essentially ignored (Cizek et al. 2008). Whereas Cizek et al. concluded from this that test developers and users reject consequences as evidence in validity research, we argue instead that most do not know about, do not understand, or have misinterpreted Messick's characterization of consequences.

Our goals in this paper are to (a) review and clarify Messick's unified view of validity and consequential basis of test interpretation and use, (b) discuss the kinds of questions evoked by value implications and social consequences and their role in construct validity and score meaning, (c) present a reframing of Messick's model and a new unified model of validity and validation, (d) introduce the concept of multilevel measures (as seen with the Early Development Instrument (EDI), some educational achievement tests, or health and social indicators) with an eye towards bringing them under the same validation umbrella as individual differences measures as this will allow us to recognize the commonality of validation practice while identifying subtle areas of distinction, and (e) offer some thoughts and directions for more explicit consideration of value implications, social consequences, and side effects of legitimate test interpretation and use with both individual differences and multilevel measures in education, psychology, and health contexts.

1 Unified Validity Theory

There are several points about unified validity theory, as presented in the 1999 *Standards* and described in whole or part by others (e.g., Anastasi 1986; Cronbach 1971; Hubley and Zumbo 1996; Kane 2006; Loevinger 1957; Messick 1989; Zumbo 2007) that are worth highlighting. Validity is about the inferences, interpretations, actions, or decisions that are based on a test score and not the test itself. It refers to the degree to which all of the accumulated evidence supports the intended interpretation of test scores for the proposed purpose. Moreover, validity is about whether the inference one makes is appropriate, meaningful, and useful given the individual or sample with which one is dealing and the

context in which the test user and individual/sample are working. That is, one cannot separate validity from the sample from which, or the context in which, the information was obtained (Zumbo 2009).

Under the unified view, validity is all about the construct and hence the meaning of scores. The process of validation involves presenting evidence and a compelling argument to support the intended inference and to show that alternative or competing inferences are not more viable. One refers to types of validity *evidence* rather than distinct types of validity. Furthermore, evidence is intended to inform an overall judgement; therefore validation is not meant to be just a piecemeal activity. Messick (2000) identified six distinguishable aspects of construct validity evidence: content, substantive processes, score structure, generalizability, external relationships, and consequences of testing. He and others (e.g., Hubley and Zumbo 1996; Zumbo 2007, 2009) have argued strenuously that validity cannot rely solely on any one of these complementary forms of evidence in isolation from the others. Another key element of construct validation has to do with construct underrepresentation and construct-irrelevant variance. Construct underrepresentation occurs when the measure fails to include important dimensions of the construct whereas construct-irrelevant variance means that variance due to other distinct constructs, variance due to the method used, and unreliable or error variance are also present. In fact, construct underrepresentation and construct-irrelevant variance are always present to some extent; the goal is to minimize their presence.

Finally, validation is an ongoing process. The unified model provides us with a regulative ideal that gives us something to strive for and governs our validation practice (Zumbo 2009). But, as Messick (1989, p. 13) points out, "Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what test scores mean." Thus, we can think of this process as being similar to the idea of repairing a ship while at sea (Zumbo 2009). It is also important to recognize that the validity of our inferences may change over time. Unified validity theory recognizes that values and cultural or societal norms change. This can affect theory and/or how we view or operationalize constructs. Language also changes over time and affects the meaning, comprehension, or relevance of test items. In fact, measurement, theory, and research all affect one another. The end result is that we must continually validate the inferences we make.

Messick (1989) developed a model, referred to as the progressive matrix of validity, that emphasizes what one needs to consider when validating inferences from our measures (both in terms of interpretation of scores and then use of scores; see Table 1). He organized

Table 1 Messick's progressive matrix of validity facets

		Function	
		Inferences from & interpretation of a test	Use of or decisions made using a test
Basis for justification	Evidential basis	Construct validity	Construct validity + relevance and utility
	Consequential basis	Construct validity + value implications	Construct validity + relevance and utility + value implications + social consequences

his matrix in terms of function (i.e., interpretation vs. use) and the basis for justifying validity (i.e., evidential basis vs. consequential basis).

2 Evidential Basis for Test Inferences and Use

It is important to remember that Messick (1989, 2000) argued that construct validity is the whole of validity because validity is about the *meaning* of the scores. What *is* the construct represented by the scores obtained on a measure? Construct validity evidence (e.g., content-related evidence, score structure, external relationships) is placed within a nomological network and provides the evidential basis for the interpretation of scores. In line with Zumbo (2009), we use the expression 'nomological network' to suggest a contextualized and pragmatic framework. Like Messick (1998), we envision this network as a fundamental aspect of the science of measurement that provides a "framework for deriving empirically testable consequences of construct theory and a foil for framing plausible rival hypotheses to challenge construct meaning" (p. 40). Importantly, Messick (1998, 2000) also considered the intended and unintended consequences of testing to be part of the evidential basis for the interpretation of scores when that information contributes to our understanding of score meaning.

To provide evidence to support the use of test scores and to make decisions based on those scores, one also needs evidence of the 'relevance and utility' of the test score inferences and actions. That is, you need to take into account the particular sample with whom, and the context in which, you are working when deciding if there is adequate construct validity evidence to support the inferences made from a measure. What you are asking is whether the inferences made will be relevant to this group and useful for the intended purpose or use of the test.

3 Consequential Basis of Test Inferences and Use

The aspect of the matrix that perhaps most reflects Messick's thinking and yet is very often misunderstood is the consequential basis for interpretation and use. The consequential basis is not about poor test practice. Rather, the consequences of testing refer to the unanticipated or unintended consequences of *legitimate* test interpretation and use (Messick 1998). There are two aspects to the consequential basis of testing: value implications and social consequences.

3.1 Value Implications

Value implications challenge us to reflect upon (a) the personal or social values suggested by our interest in the construct and the name/label selected to represent that construct, (b) the personal or social values reflected by the theory underlying the construct and its measurement, and (c) the values reflected by the broader social ideologies that impacted the development of the identified theory (Messick 1980, 1989). There are evaluative overtones to any construct label or name of a measure. Messick (1980) points to phrasing such as 'flexibility versus rigidity' as opposed to 'confusion versus control' as an example. Consider the difference between using a test name such as the 'Early Development Instrument' versus the 'School Readiness Inventory' versus 'Developmental Immaturity Scales'. The point is not whether one uses a positively worded or negatively worded label

or name, but that the label or name one chooses might cause researchers to investigate the construct differently or practitioners to view the measure and its use differently. In naming a construct or measure, it is important to strive for consistency among its theoretical importance, empirical evidence of its meaning, and relevant value implications and connotations (Messick 1980).

Hubley and Zumbo (1996) used a depression measure for older adults as an example to stimulate discussion of value implications. The kinds of value implication questions that one might ask with such a measure include: what values are reflected by (a) an interest in identifying and measuring negative symptomatology, (b) terming this constellation of symptoms 'depression', (c) focusing specifically on older adults, (d) using a deficit model or values of normality/abnormality to frame the construct, (e) using a cognitive-behavioural theory to guide the content and item development, (f) using theories of aging to guide scale content and format, and (g) use of a self-report format to elicit this information? It is important to reflect upon and understand the values that underlie our constructs, measures, and measurement because they impact the meaning of the test scores, the relevance and utility of inferences made with different samples, contexts, and time periods, and the consequences of test use. Recall that value implications are relevant to both test interpretation and use. Values often are the impetus for test score use and decisions and, thus, value implications related to a test may emerge or become clearer when it is used in a particular social context (cf. Linn 1997; Messick 1995).

There is an important feedback loop between theory and research that was identified by Hubley and Zumbo (1996). For example, it is common to hear that more women are depressed than men. This statement is based on research results using our existing measures of depression. In turn, when evaluating measures of depression (e.g., using a known-groups validation method), one would expect to see a greater proportion of women to be depressed than men or the measure will not be seen to be sensitive to a known gender difference. But note that there is a circular process here in that findings based on our measures help shape theory while theory and research help shape our evaluation of measures. If there is bias among the items on existing depression measures, then that bias also contaminates our theories (and vice versa). In addition, theory and research findings contribute to shaping our values. We need to be more cognizant of and question how our values may shape our theories, the development and evaluation of our measures, and our interpretations of research findings. Likewise, our research findings, measures, and approaches to measurement shape our theories, discourse, and ultimately our value systems.

3.2 Social Consequences

Social consequences refer to consequences for society stemming from the use of a measure. Knowing the care with which Messick selected terminology and with which he wrote, we have often wondered why he consistently chose the adjective "social" as opposed to simply the term "consequences" without any modifier. A common dictionary definition of the word 'social' references human society and hence social consequences implies societal consequences and the welfare of human beings as members of society. We can only speculate that he was drawing a distinction between society and the individual, and shining a light upon the former. Although some writers have argued that social consequences have no place in validity, their argument tends to be based on a misconception that social consequences are about test use and, in particular, test *misuse* (cf. Brennan 2006; Mehrens 1997; Popham 1997). First, the focus is on consequences, not use. Second, Messick (1998) did not view test misuse or illegitimate test use to be part of his consequences of testing.

Indeed, although they might be important concerns, he saw the consequences of test *misuse* as irrelevant to the nomological network and score meaning and thus outside of construct validity and the validation process.

Social consequences of legitimate test use can be positive or negative and both are important in terms of validity. While the test developer and test user are often more concerned about unanticipated *negative or adverse* effects resulting from test use, we argue that one also needs to consider positive effects when considering validity and score meaning. Again, our focus, from a validity standpoint, is about effects that are traceable to sources of invalidity such as construct underrepresentation and construct-irrelevant variance. Because these consequences contribute to the soundness of score meaning, they are an integral part of construct validity and the validation process (Messick 1989, 2000).

Let's go back to the example of a depression measure for older adults. If the measure is being used legitimately to screen for or describe depression levels in the general community, one needs to consider the intended and unintended social consequences of finding very small or very large numbers of depressed elderly. We need to consider how such findings might impact, for example, theories about depression and mental health, theories about aging, the presence or funding of community mental health programs, physicians' diagnoses of older patients' reported symptoms, and/or group health plan coverage and rates. Furthermore, we must consider how these impacts affect score meaning and use.

Remember too that construct validity, relevance and utility, value implications, and social consequences all work together and impact one another in test interpretation and use. If the measure is to be used, for example, with a different cultural group (e.g., Aboriginal peoples, immigrant groups) than the original test development sample, we must ask (a) if the meaning of the scores is the same with this group, (b) if the scores are relevant and useful for this group given our purpose and context, (c) about the role and impact of values in this case, and (d) about the social consequences stemming from this particular test use (cf. Messick 2000; Willingham 2002; Willingham and Cole 1997). It is essential to consider whether a newly studied cultural group conceives of or values the construct in the same way as the original group upon which the construct or measure was developed. Again, this is a question about the degree to which the obtained scores reflect construct underrepresentation and/or construct-irrelevant variance. Messick (1989) identified the test user as being in the best position to evaluate the meaning of scores obtained in a given context and to determine the extent to which the intended meaning of those scores may have been eroded by contaminating influences within that context.

Not all social consequences are due to sources of invalidity. For example, if a negative social consequence (such as financially penalizing schools for children's poor test performance) is the result of test misuse due to external political beliefs or policies (such as the 2001 'No Child Left Behind Act' in the US), then such a consequence is outside the realm of validity. It is important for test developers and users to be cognizant of the relationship between score meaning and social consequences. If social consequences occur that are traceable to construct underrepresentation and/or construct-irrelevant variance, then the construct and/or measure need to be modified to incorporate these findings; if they are not, then they are not part of validity (Messick 1998).

4 Reframing Messick's Progressive Matrix: A New Model of Validity and Validation

The value of Messick's (1989) progressive matrix was that it built upon the state of affairs in validity theory throughout the 1960s–1990s and highlighted the consequential basis as

an understudied and underappreciated element of validation work. There have been numerous discussions about the role of value implications and, most notably, consequences in validity as a result of the model presented in the progressive matrix. This was, we suspect, partly Messick's intention in presenting the matrix. However, a negative outcome of the matrix is that the consequential basis appears to be given equal weighting to the evidential basis. This has, in part, contributed to misunderstandings about the meaning and intention of the consequential basis as well as protests about the burden involved in dealing with it. In fact, much of Messick's later work (e.g., Messick 1995, 2000) was spent clarifying his points and correcting misunderstandings about value implications and social consequences. We argue that Messick's matrix could be greatly simplified and consequences placed more appropriately and simply as part of the evidential basis for test interpretation and use. Table 2 depicts our reframing of Messick's matrix.

In this revised framework, the consequential basis is not set out as a seemingly separate entity from the evidential basis and its placement also better reflects Cronbach's (1988, p. 4) view that "the argument must link concepts, evidence, social and personal consequences, and values". The risk in this reframing of Messick's (1989) matrix is that test developers and users will feel less pressure to consider the value implications and social consequences inherent in their work. We hope, however, that this reframing makes its clearer that value implications and social consequences are inherent to score meaning and are not part of a new or separate 'consequential validity' (cf. Messick 1995). Our reframing of the matrix is purposefully depicted using a similar format to Messick's progressive matrix. This allows readers to easily compare and contrast the two frameworks and makes it more apparent that only utility separates the facets of validity relevant to test score interpretation versus use.

Even revised, we find Messick's (1989) matrix to be unsatisfying. Figure 1 shows our re-envisioning of a contemporary unified validity and validation framework, paying greater attention to the role of theory and values at each step, types of evidence included in construct validation, and the role of intended consequences and unintended side effects. Our new framework is also consistent with Zumbo's (2009) depiction of validity and validation as an integrative cognitive judgment involving a form of contextualized and pragmatic view of explanation.

Our new model of validity and validation highlights several key features. First, one can envision that, based on a construct, one develops a test/measure to which one ascribes test score meaning and inference. From test score meaning and inference emerge (a) intended social and personal consequences, but also (b) unintended social and personal side effects of legitimate test use. Unlike Messick, we argue there may be personal as well as social impacts. In addition, we think it is helpful to use different terms to distinguish between intended consequences and unintended side effects. Importantly, consequences and side effects of legitimate test use may also influence test score meaning and inference, which

Table 2 Hubley and Zumbo's reframing of Messick's matrix

	Function	
	Inferences from and interpretation of test scores	Use of, or decisions made based on, test scores
Evidential basis	Construct validity + relevance + value implications + social consequences	Construct validity + relevance and utility + value implications + social consequences

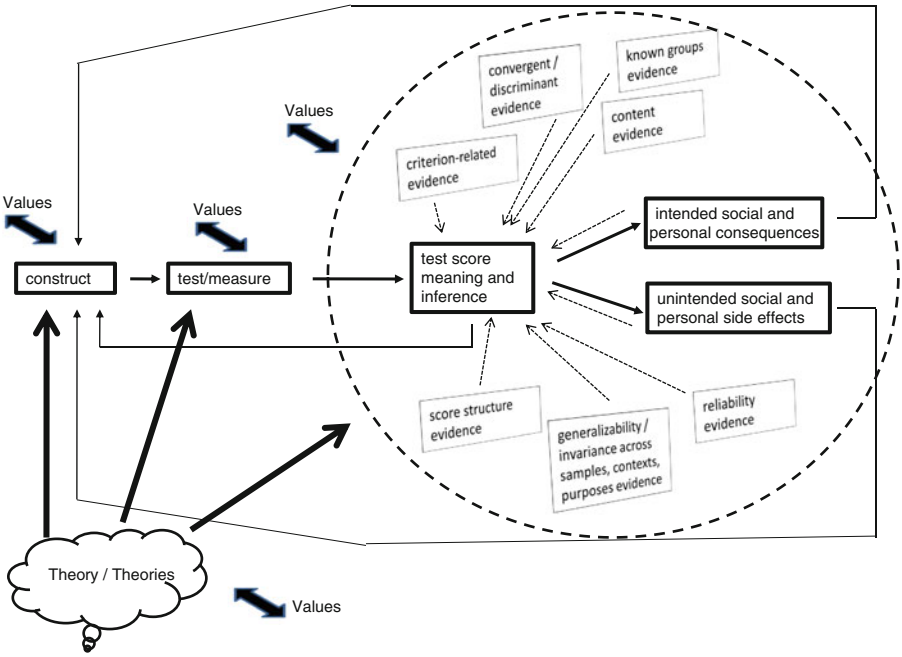

Fig. 1 Hubley and Zumbo's revised unified view of validity and validation

makes them relevant to the validation process. Test score meaning and inference are effected and shaped by several forms of validity evidence, which are enclosed within the large dashed circle and include but are not necessarily limited to: criterion-related, convergent/discriminant, known groups, content, score structure, reliability, and generalizability/invariance evidence as well as intended social and personal consequences and unintended social and personal side effects. The centrality of the large dashed circle is meant to signify that construct validity is at the core of this unified view of validity and validation. It should be apparent that theory or theories influence the construct, the test/measure, and construct validity evidence. The 'theory/theories' we are referring to include the theory related to the construct, theories related to the sample and context, and psychometric theory and models. Finally, we can see that the effect of values is pervasive throughout the framework and related to theory/theories (broadly defined), the construct, test/measure, and construct validity as well as validation choices and decisions.

5 Validation of Individual Differences Versus Multilevel Constructs and Measures

Often, validity and validation are discussed in the context of individual differences constructs in which the scores from tests/measures are interpreted and/or used at the individual person level. However, there are many cases in education, psychology, and health in which multilevel constructs and measures are employed. Indeed their use is growing in popularity in everyday assessment and evaluation practice. Unfortunately, multilevel measures tend to be discussed outside of the purview of construct validity. Our new framework of validity and validation, however, applies equally well to test/measures based on either individual differences or multilevel constructs.

Zumbo and Forer (2011, p. 1) define a multilevel construct as "a phenomenon that is potentially differentially meaningful both in use and interpretation at the level of individuals and at one or more levels of aggregation." It is critical to note that this definition "allows for measures that are used and scores [that] are reported only at the aggregate level (e.g., NAEP and some international assessments such as TIMSS or PISA) as well as for measures that are used and scores [that] are reported at both the individual and aggregate levels (e.g., statewide educational assessments)."

Aggregate level data are often used for policy, planning, intervention, and even funding purposes. For example, in the educational field, aggregate data such as 'pass/fail' proportions or proportions of students 'meeting expectations' from state-wide assessments (e.g., state-wide Gr. 8 math exams) are used primarily for curriculum evaluation and planning. In the mental health field, the Center for Epidemiological Studies—Depression (CES-D) scale (Radloff 1977) was originally designed to measure and report community level rates of depressive symptomatology. Finally, the Early Development Instrument (EDI), the focus of this special issue, is a population-based measure of school readiness that is completed for individual children by their teacher, but the scores are aggregated to, and reported at, the school or neighbourhood/community level (Janus 2006).

It is important to remember the *level at which the test scores are interpreted*. When test scores are interpreted at the individual level, construct validity evidence is obtained at the individual level. When data from multilevel measures are reported, interpreted, and used at a group level (e.g., at the level of school, neighbourhood, country), construct validity evidence also needs to reflect this same group level of data (Zumbo and Forer 2011; Forer and Zumbo 2011). As with all measures, it is important for an array of empirical evidence to be compiled and a compelling argument put forward to support the intended inference(s) and to show that alternative or competing inferences are not more viable. Such evidence might include (multilevel) reliability, content-related evidence, (multilevel) factor structure, substantive processes, (multilevel) external relationships, articulation of values (e.g., in theory, the construct, score meaning, and use), and both intended social and personal consequences and unintended social and personal side effects of legitimate test use on the meaning of test scores. Forer and Zumbo (2011) describe multilevel construct validity and provide some conceptual and technical psychometric validity evidence for the EDI. Linn (2008, 2009), Forer and Zumbo (2011), and Zumbo and Forer (2011) challenged us to think about the potential errors in inference that can be made across levels of data—that is, what Zumbo and Forer refer to as ecological or atomistic fallacies of measurement data inferences. In fact, potential errors in inference across levels of data highlight that we need to think about cross-level consequences and the eventual trickle-down of the high stakes resulting from the use and reporting of multilevel assessment data. An example of these high stakes results are discussed by Linn (2006, 2008) and Kane (2006) when considering the consequences of test use in policy-making and program evaluation. Just because data are gathered and reported at the aggregate level does not mean there are not consequences or side effects at a personal level. When aggregate data are used to make policy decisions or funding/resource allocations, there can still be direct and immediate impacts on individuals.

Once an intended consequence or unintended side effect of legitimate test use has been identified, one then needs to explore what that consequence or side effect might mean for score meaning. For example, in the case of a state-wide math literacy (i.e., numeracy) test, consider (a) a potential intended consequence of increased high school graduation rates in the state and (b) a potential unintended side effect of increased rates of teachers teaching to the test. How do each of the above social consequences and side effects of use affect the

meaning of the state-level math literacy test scores, how 'math literacy' is conceptualized, and theories about math literacy, math achievement, math pedagogy, and cognitive development in children? Explicit consideration of social and personal consequences and side effects might enlighten us with respect to whether personal (e.g., gender, ethnicity, culture, language) and contextual factors (e.g., poverty, community support) are part of the math literacy construct or external to it. As summed up by Shepard (1997), "consequences are evaluated in terms of the intended construct meaning" (p. 8). When the social consequences and side effects of using a math literacy measure are not congruent with our societal values and goals regarding math literacy, and more broadly numeracy, such insights in the validation process may be used to adjust constructs, theories, and aspects of the measurement process until the desired congruence between values, purposes, and consequences is accomplished. In addition to state-wide educational assessments, these considerations apply to all national and international assessments in, for example, education, health, social policy, and psychology/mental health.

6 Conclusion

In unified validity theory, validity is all about the construct and, hence, the meaning of scores. What we validate are the inferences, interpretations, actions, or decisions that we make based on a test score. Thus, validity is about the degree to which our inferences are appropriate, meaningful, and useful given the individual or sample we are dealing with and the context in which we are working. Validation is about presenting empirical evidence and a compelling argument to support the intended inference and to show that alternative or competing inferences are not more viable. In particular, we aim to identify the degree to which construct underrepresentation and construct-irrelevant variance are problems. But, as noted by Cronbach and Meehl (1955), "both the test and the theory are under scrutiny" (p. 296) and if research findings obtained from test scores and the theory disagree, then this discrepancy needs to be resolved with either a new test, a new theory, or both because validation work applies to both the test and the explanatory theory (Shepard 1997; Zumbo 2009). Because changes (e.g., in our empirical knowledge, theoretical understandings, values, society) occur over time, the process of validation is an ongoing one.

Messick (1989) summarized his view of unified validity theory in a progressive matrix of validity facets in which he identified four facets, distinguishing between the functions of test interpretation and test use and between evidential and consequential bases for justifying these functions. Messick's goal may have been to highlight the importance of the consequential basis of test interpretation (i.e., value implications) and use (i.e., social consequences) in validity but, instead, the consequential basis generally has been sorely misunderstood and criticized for being an untenable addition to unified validity that unnecessarily burdens test developers and users.

Value implications were not a new addition to conceptions of validity; rather, Messick (1995) strenuously argued that values are already inherent to score meaning and his intention was to expose them to examination and debate. Everything we do—from our development of a construct and measure to our use of tests to our interpretation of the obtained scores to our validation approaches—reflects our values. Recall that validation is an ongoing process in which one presents the most reasonable case about the meaning of test scores based on what we know from research and test use at a given time and that the validity of our inferences may change over time. We need to be more cognizant of how, over time, (a) our values result in changes in language, interpretation, theory, use, and

consequences, and (b) these changes may reinforce or challenge our values, and that (c) both of these processes impact construct validity and the meaning of scores.

The concept that 'consequences' related to testing may contribute to validity is not a new idea, but certainly Messick was the key player in highlighting the potential role of social consequences in elucidating values and score meaning. The presentation of the consequential basis of validity as separate from the evidential basis has led to two distinct viewpoints among validity specialists. While everyone agrees that values and social consequences are important, detractors would likely argue that including these facets in a unified validity model is a case of construct irrelevant variance whereas supporters would argue that removing these facets from a unified validity model is itself an example of construct underrepresentation. Perhaps the greatest frustration for those who think that social consequences play an important role in validity has been the continued mistaken belief by many writers that social consequences are about test misuse rather than *legitimate* test interpretation and use.

Another issue raised by detractors is that examining social consequences is too great a burden. We do not think so. Examining how various consequences (and side effects) of legitimate test interpretation and use impact construct validity and score meaning is certainly a challenge. But this is a task that is necessary if we are to fully understand the constructs we are measuring. Moreover, we need to be more reflective, more thoughtful, and more aware of how values, theory, practice, and consequences are linked. We offered a reframing of Messick's (1989) progressive matrix of validity that we think is less confusing and places the consequential basis of validity in its proper place within a unified validity model. We also presented our new framework of a unified view of validity and validation that more clearly articulates social consequences as intended social and personal consequences and unintended social and personal side effects as well as shows the placement and role of theory/theories, values, consequences, and side effects in validity and validation. To be able to examine the values implicit in, and social and personal consequences and side effects stemming from, legitimate test interpretation and use requires that test developers and test users must be willing to think deeply about validity and not just be automatons in the validation process. To truly move our understanding of the meaning of test scores forward, we must be willing to do the work to examine what *all* of the available validation evidence tells us.

References

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: American Council on Education/Praeger.

Cizek, G. J., Rosenberg, S., & Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–412.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

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

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

Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research*. doi:10.1007/s11205-011-9844-3.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123*, 207–215.

Janus, M. (2006). *Early Development Instrument: An indicator of developmental health at school entry*. Monograph from the proceedings of the International Conference on Measuring Early Child Development, Vaudreuil Quebec.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*, 14–16.

Linn, R. L. (2006). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education, 19*, 5–15.

Linn, R. L. (2008). *Validation of uses and interpretations of state assessments*. Washington, DC: Council of Chief State School Officers.

Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 195–212). Charlotte, NC: IAP—Information Age Publishing, Inc.

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

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*, 16–18.

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

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

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

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

Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 3–20). Boston: Kluwer Academic Publishers.

Popham, W. J. (1997). Consequential validity: Right concern–wrong concept. *Educational Measurement: Issues and Practice, 16*, 9–13.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16,* 5–8,13, 24.

Willingham, W. W. (2002). Seeking fair alternatives in construct design. In H. I. Braun, D. N. Jackson, D. E. Wiley, & S. Messick (Eds.), *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Lawrence Erlbaum.

Willingham, W. W., & Cole, N. J. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 45–79). The Netherlands: Elsevier Science B.V.

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). Charlotte, NC: IAP—Information Age Publishing, Inc.

Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bovaird, K. Geisinger, & C. Buckendahl (Eds). *High stakes testing in education—science and practice in K-12 settings [Festschrift to Barbara Plake]*. Washington, DC: American Psychological Association Press (in press).