Samuel Messick's Consequential Validity

Robert E. Orton

1954-1998

University of Minnesota

Research in construct validity, begun with L.J. Cronbach and Paul Meehl's 1955 influential essay, continues to attract the attention of measurement specialists and philosophers of education alike.[1] Samuel Messick's unified but four-faceted explanation of test validity, developed over the last thirty years, is the current "received view."[2] One of the noteworthy features of Messick's view is his distinction between the "evidential" and the "consequential" bases of test validity. This distinction, which can be glossed initially as a difference between factual and value-laden aspects of testing, is interesting for the following reason. Though test users have always (to various degrees) acknowledged that their work is value-laden, Messick builds this normative element into his concept of validity — a notion traditionally associated with "truth."

This essay offers an interpretation of Messick's consequential validity, particularly as it helps to illustrate the practical aspects of testing. The main point to be argued here is that, though inferences based on test scores are value-laden, it is important to distinguish between technical and axiological dimensions of practicality, with Messick's evidential basis of test validity more closely associated with the former and his consequential basis with the latter. A related goal is to push against semantic holism, the philosophical "foil" here being W.V.O. Quine's arguments against the analytic-synthetic distinction.[3] Though it may be often difficult or impossible to distinguish between facts and stipulations, or between facts and values, a test interpreter ought to distinguish between stipulations and values, or so it will be argued. Failure to do so mixes technical and ethical issues, confuses "things" and "persons," and risks violating Immanuel Kant's Categorical Imperative.[4]

After a brief overview of the analytic-synthetic distinction and its relationship to construct validity, the discussion turns to two ways in which practical aspects of testing can be problematic: as technical and as ethical muddles. Parallels between this and Messick's evidential-consequential distinction are offered. The tone is for a shift away from semantic holism and toward a reductionism based on a moral distinction between persons and things.

Semantic Holism

Postpositivist philosophers of science are generally suspicious of a fact-value distinction.[5] This suspicion is related to a mistrust of attempts to distinguish between empirical facts and any sort of convention or "nongiven," be it a logical stipulation or a human value. The distinction between stipulations and facts is systematically ambiguous, the result of what Quine explains as a "radical untranslatibility."[6] Our theories face the tribunal of experience as a whole, and piecemeal decisions that this

or that term can be translated in a given way will always mix facts and stipulations. Using one of Quine's examples, a person pointing toward a rabbit and uttering the unknown phrase "gavagai" may be referring to "rabbit" or "rabbit part."[7] No one can know for sure. A decision as to how to translate "gavagai" can only be made by observing the behavior of a language user over time, and even then there is always a level of uncertainty regarding the denotation of a given term. A pragmatic acceptance that the use of a given term is "precise enough" for the purpose at hand goes deeper than any purported link between word and object.

The early positivists' attempt to individuate meaning is thus shown to suffer from a misplaced quest for certainty. Bertrand Russell's Supreme Maxim of Scientific Philosophizing: "Whenever possible, substitute constructions from known entities for inferences to unknown entities," places too great an emphasis on isolating the logical atoms of knowledge.[8] In place of a reductionism associated with the isolation of error, the cautious attitude is reflected in two very different maxims: 1) It is always possible that a given statement is linked to any other statement, and 2) Language "goes all the way down" in the sense that distinguishing between "the scheme" and "the content" falls away with Ockham Razor.[9] A holistic nominalism replaces the caution of the earlier logical atomism.

This rise of semantic holism coincides with the birth of construct validity. In their 1948 essay, Kenneth MacCorquodale and Meehl distinguish between "hypothetical constructs" and "intervening variables," the former of which play a major role in construct validity.[10] With intervening variables it is possible to distinguish the logical or stipulative meaning of a term, which is derived from its operational definition, from the empirical significance of a term, which follows from its observed relationships with other terms. With hypothetical constructs, by contrast, this version of the analytic-synthetic distinction cannot be drawn. The resulting holism did not sit well with early critics of construct validity, one referring to it as "the logical blur which inevitably surrounds the hypostatization of a research preference or a hunch about 'fruitfulness' into a systematic position."[11] Ultra-operationist critics notwithstanding, construct validity survived and continues to survive. In an oft-quoted phrase, Quine summarizes many of the issues regarding logical atomism and the analytic-synthetic distinction as follows: "Conventionality is a passing trait. Significant for classifying terms on the moving front of science, but useless for classifying terms behind the lines,"[12] Many constructs in educational and psychological measurement are on the "moving front" of science in that they are little more than abstractions from variables clustered using statistical techniques. In these cases, causal laws are, at best, constant conjunctions or empirical regularities — or they appear to be empirical regularities. Maybe there is not enough evidence to make a causal claim, meaning that even internal validity is lacking. In these cases, it may be possible to distinguish between formal or stipulative relationships, on the one hand, and experimentally determined or factual relationships, on the other. That is, it may be possible to draw an analytic-synthetic distinction.

In the natural sciences, by contrast, drawing an analytic-synthetic distinction is often pointless or arbitrary. Constructs such as "electron," which Hilary Putnam

calls "law cluster concepts," combine stipulative and empirically determined relationships in their very meaning.[13] Even the referents of these terms are so determined by practical knowledge closely linked to a "nomological network" that one accepts the whole theory associated with electrons, not just individual instances of individual laws. Separating the conventional or stipulative and the nonconventional or factual aspects of these terms is difficult or impossible.

The difficult cases — and these may be the lion's share of construct validity cases in the late twentieth century — are between these two extremes. Here it may not be clear whether or not the analytic-synthetic distinction is practical or impractical to make. Many interpretations of educational assessments are somewhere between knowledge-poor formal manipulation (Quine's "moving front of science") and knowledge-rich nomological networks that link Putnam's "law-cluster concepts." It is this middle ground that is the focus of the remainder of the discussion.

Technical and Moral Stipulations

Given the thesis of semantic holism — that the distinction between facts and stipulations is pointless or arbitrary for well-developed constructs — what will be more closely considered here is the degree to which holism makes sense for constructs toward "the moving front of science." In a nutshell, the focus is the relationship between the fact-stipulation distinction, on the one hand, and the fact-value distinction, on the other. These relationships, in turn, will be discussed within the context of a third distinction: the difference between what is "taken for fact" and what is "taken as problematic," when interpreting educational and psychological tests. Of interest is the relationship between practical background knowledge and being able to draw the analytic-synthetic distinction.

The example of using tests to place children in special education classes can be used to focus the discussion. When interpreting results from a test designed to make special education placements, the context for making a decision will likely include aspects that can be "taken as fact," such as the validity of the test, the promise of funding support for students placed, or the availability of curricula. The context will also include aspects that are more problematic, such as likely impact of the placement, a family background that raises some red flags, or instances of unsuccessful placements for this student in the past. Foreshadowing Messick's consequential validity, the decision whether or not to place a child in a special education program has uncertainty associated with the effect that the placement will have upon the child and with the evidence that is used to support the decision (test scores, teacher judgments, or parents' wishes, for example). We cannot predict the future, and we do not have exact knowledge of the past.

What is of interest here is how the difference between "taken for facts" and "open for interpretation" varies, depending on how much is known about a situation and how this epistemic context influences the analytic-synthetic distinction. If the test designed for assisting placement decisions in special education is particularly new, then there would seem to be a greater need to make a firm distinction between the facts and the interpretations. The difference would be drawn along the lines of the mythical "given." The problematic or interpretive would be most clearly the

result of something conventional, stipulative, and tentative. When precious little is known about a situation, a cautious attitude leading toward a clearer distinction between the reliable or factual and the unreliable or interpretive would be prudent. This distinction, in the case of novel situations with tests, will be such that the "facts" are likely to be so regarded by the largest number of untrained observers, just as a jury is ideally composed of ordinary citizens.

The case would appear to be different when looking at the everyday practice of mature science. A physicist puzzled by an electromagnetic phenomenon will also need to take something for granted. However, the justification for her conjectures which explain the electromagnetic phenomenon may appeal to a "taken as unproblematic" background that would be only regarded as "fact" by other trained specialists. The reason for this is that the value-laden dimensions associated with constructs such as "electron" are not explicit parts of the nomological network of electromagnetic phenomena. But the same is not true for constructs such as "reading readiness" used for special education placements. This point will be expanded upon below.

More generally, in knowledge-poor situations in the social sciences, stipulations can be separated from facts. Here, one is trying out hunches regarding data and instrumentation. However, the moral importance of these stipulations needs to be always problematic, which is why "the facts" need to be so regarded by specialists and laypersons alike. If the moral importance of a stipulation (say, as it pertains to a potential special education placement based on a "new" instrument) is temporarily bracketed when making interpretations based on scores from an instrument, then these moral considerations had better be unbracketed before making any decisions based on these scores. Stipulations can be separated from facts, but in knowledge-poor situations the moral and the epistemological aspects of the stipulations should be difficult to separate. To not consider the moral implications of a stipulation in a case where practical knowledge is weak would be ethically risky.

To take the other extreme, in knowledge-rich situations in the social sciences, stipulations are more difficult to separate from facts. Here, the data again and again lead some credence to a causal hypothesis, such as the use of "advance organizers" to facilitate student reading comprehension. Evidence impinges upon the nomological network in a holistic fashion. Because it is so difficult to separate stipulations from facts in these "mature" theories, evidence needs to be run through both the nomological network and the "moral network." Parsimony would dictate not drawing a distinction between nomological and moral contexts. However, failing to draw the distinction runs the risk of either confusing moral issues with technical issues or of confusing moral issues with aesthetic issues.

In short, the interpretation of standardized tests may require a moral reductionism. This is due, in part, to differences between present-day circumstances, as compared to those when construct validity was first created. In opposition to the radical operationism of the early operational psychologists, what one often finds in present-day public policy is the insistence that standardized tests be trusted for

making important decisions about people, such as the decision as to who is to be placed in special education. In opposition to the position of the positivist empiricists, who insisted on a firm distinction between analytic and synthetic statements, what one finds is a holistic pragmatism resting on a naive causal realism. Collapsing the distinction between "meaning" and "evidence" can reinforce a holism that tries to include both scientific and moral aspects of testing. As the next section will argue, this risks confusing persons and things.

Messick's Distinction

Messick's consequential validity helps to work against these excesses. His distinction between evidential and consequential validity underscores a difference between two ways in which practical problems in testing can be problematic. There is not enough evidence — not enough is known — about a given phenomenon. Or, what evidence there is may be interpreted differently within different moral outlooks. The difference between these two cases can be illustrated using contrasts. In one sense, there are few practical problems associated with "electrons," in the context of current physical theory, which is not to say that everything is known about the associated phenomena that can be known. However, there are many practical problems associated with "electrons," in the context of legal theory, when these particles are used to terminate the life of a convicted murderer. Electricity may not be problematic in the practical scientific sense, but it is problematic in the practical moral sense.

To take an "opposite" example: There are few practical problems associated with giving praise to school children. From a human point of view, giving praise is associated with the natural trait of caring and nurturing, and though praise could be abused, most would agree that children with healthy self-images are desirable. "Praise" is not seriously problematic, from a moral point of view. However, from a scientific point of view, there are many practical problems associated with giving praise to school children. The relationship between positive feedback and test scores is not well-understood. There are instances where self-doubt may encourage greater effort and achievement. There are others where a mistrust about what one knows can lead to a lack of confidence, which a healthy self-concept would correct. Here, a morally unproblematic notion is technically problematic.

One way to interpret this asymmetry is that, within the context of physical theory, "electron" occurs within a well-developed network of laws and theories. In the context of moral theory, ending a criminal's life by electrocution is interpreted within different theories which may be, to some degree, incommensurable. By contrast, within the context of moral theory, giving praise to school children occurs within well-accepted norms of parenting and caring. In the context of scientific theory, however, different psychological frameworks interpret the relationship between praising children and student achievement in different ways.

The point here is that the moral network is different from the nomological network. Though well-accepted epistemic aspects may engender holism in the nomological network, and widely-accepted beliefs may lead to holism in the moral network, there are few instances in interpreting standardized tests where both moral

outlook and scientific knowledge together can warrant a holism that embraces both networks. Or, put another way, the ways in which practical aspects of testing can be problematic are at least two: not enough is known, or there is a confusion of values. Though these two types of confusion can arise in tandem, it is useful to separate them.

The reason for suggesting this separation parallels Messick's distinction between evidential and consequential aspects of construct validity. Namely, different types of evidence are needed to address whatever type of confusion is present. Or, put in holistic-pragmatic terms, different language games are played, depending on whether the muddle is of an epistemic or a moral sort. Just as one would not do an "experiment" to determine whether or not the Golden Rule is valid, so one would not engage in a moral debate to determine whether a particular scientific interpretation of "electron" is valid.

Isolating the moral stipulation from whatever non-moral or epistemological stipulation one is making can lead to greater clarity as to just where one is standing. Morally, test interpreters needs to stand somewhere. Being as clear as one possibly can be about this stand is what is being argued for. Expressed as an hypothesis: Clarity with regards to a moral issue in interpreting standardized tests scores is different in kind, or qualitatively different, from clarity with regard to an epistemological issue.

Some support for this hypothesis can be found in Kantian moral theory. One of the formulations of the Categorical Imperative is that we always treat others as ends in themselves rather than as means to ends.[14] For example, when interpreting test scores related to special education placement, there would be a difference between those aspects of the interpretation where means to ends thinking is appropriate and those where this thinking is not appropriate. Broadly, the former would be the theoretical or epistemic aspects of the situation, whereas the latter the practical or decision making aspect. If one mixes these two up, one may risk acting in such a way that a moral course (children need to be treated as ends) becomes contaminated by a technical course (children are treated as means to ends).

If this distinction between persons and things is accepted, then Messick's distinction between the evidential and the consequential aspects of validity can be seen as a reminder of the Categorical Imperative. Broadly put, the evidential aspects of test validity are epistemic matters, whereas the consequential aspects are axiological matters. The evidential aspects of construct validity pertain either to empirical facts or logical stipulations, or a combination of both in a normal scientific paradigm. The consequential aspects of construct validity pertain to either empirical facts or moral prescriptions, or to some combination of both in a value-laden social context. However—and this is the main point — logical stipulations are not to be confused with moral prescriptions. Though both are to be evaluated using a mixture of empirical and conceptual criteria, the former (logical stipulations) are pragmatically instrumental in the sense that they can safely move between persons and things. The latter do not have this characteristic, in that moral stipulations must always preserve the categorical distinction between persons and things.

Put another way, when creating theory in social science, we are sometimes on safe grounds to blur the distinction between persons and things, treating people as objects to be manipulated in techniques such as people as numbers and then manipulate these numbers without a lot of concern for the fact that these numbers represent people. Often, the ends of people are better served by this manipulation, such as when doing so leads to better surgical techniques, cancer drugs, or cautions about safe driving with seat belts on Labor Day. But, particularly when we move from the creation of theory to the making of decisions that affect people, treating people as objects is a violation of the Categorical Imperative. This is not to say that we never break the rule. But we ought not to break it, and when we do, we should feel bad about ourselves. At least, following Kant, this is the law that we ought to want to live under, as free, rational beings.

To summarize: An epistemic muddle is different from a moral muddle. Unless virtually everything must be always viewed from a moral perspective, which seems difficult, a distinction between moral and epistemic language games is useful. Unless it is immoral to be anything but difficult, a distinction between facts and values is helpful. Legal rules are played by different strategies or tactics than scientific rules. Keeping these rules distinct is a cautious strategy, in that it helps to preserve the distinction between "persons" and "things."

Conclusions

It is a commonplace that "the practical" is ambiguous between "the technical" and "the moral." As Elliot Eisner puts it, there is a difference between installing a refrigerator and nurturing a friendship."[15] Joseph Schwab's eclecticism may be one of the best ways to muddle through a mix of art and life.[16] However, not all situations are best interpreted in this blended fashion. In his self-proclaimed "unartistic" style, Kant distinguishes between hypothetical and categorical thinking.[17] The first is associated with reasoning from means to ends and is more closely linked to art. The latter is bound up with reasoning about ends themselves and is more associated with morals. A similar distinction is explained by Aristotle between the "productive" and the "practical" aspects of thinking."[18] The first are always done for the sake of a further end. The latter are always done for their own sake.

It has been argued that Messick's distinction between evidential and consequential aspects of test validity helps preserve this distinction, guarding against treating persons as things when making high stakes testing decisions. In this way, Messick's distinction works against a holism supported by naive causal realism. Though semantic holism may best explain mature scientific practice and everyday common sense, it fares less well in the case of interpreting and making decisions based on educational and psychological measurements. The following "heuristics" summarize the argument: If the analytic-synthetic distinction is easy to draw, then the values-analytic distinction should be hard to draw. We do not know enough, at least in the epistemic sense, and we should rely on our moral intuitions. If the analytic-synthetic distinction is hard to draw, then the distinction between values and the analytic-synthetic merge should be easier to draw. We know more, but this knowledge would ideally serve a higher moral purpose, or it should be instrumental

for moral purposes. People, not electrons, should decide conceptions of justice. Finally, if we cannot tell whether or not it is easy or hard to draw the analytic-synthetic distinction, then we need to be doubly cautious, or perhaps hyper-pessimistic in Michel Foucault's sense.[19] This last case appears to be very common when educational and psychological tests are used to make decisions about people.

The "testing movement" shows little evidence of abating. The surest path to progress may be to slow the train down. However, no one wants to stand in front of a moving train without a real way out. A pragmatically negotiated, context-dependent nominalist distinction that could at any moment evaporate before one's eyes may not be enough. What has been argued is that Messick's evidential-consequential distinction provides a way to help insure that the train is on the right tracks.

---

1. L.J. Cronbach and Paul E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin* 52 (1955): 281-302.

2. Samuel Messick, "The Standard Problem: Meaning and Values in Measurement and Evaluation," A*merican Psychologist* 30 (1975): 955-66; Samuel Messick, "Evidence and Ethics in the Evaluation of Tests," *Educational Researcher* 10, no. 9: 9-20; and Samuel Messick, "Validity," in *Educational Measurement*, 3d ed., ed. R.L. Linn (New York: American Council of Education), 13-103.

3. W.V.O. Quine, *Word and Object* (Cambridge: MIT Press, 1960); W.V.O. Quine, *From a Logical Point of View* (Cambridge: Harvard University Press. 1953).

4. Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans. H.J. Paton (1785; reprint, New York: Harper and Row, 1964).

5. See Kenneth R. Howe, "Two Dogmas of Educational Research," *Educational Researcher* 14, no. 8 (1985):10-18.

6. Quine, *Word and Object*.

7. Quine, *From a Logical Point of View*, 51-57.

8. Bertrand Russell, *Mysticism and Logic* (New York: W.W. Norton, 1929).

9. D. Davidson, "On The Very Idea of a Conceptual Scheme," *Proceedings of the American Philosophical Association* 47, (1973-74).

10. Kenneth MacCorquodale and Paul E. Meehl, "On a Distinction between Hypothetical Constructs and Intervening Variables," *Psychological Review* 55 (1948).

11. G. Bergmann, "Theoretical Psychology," *Annual Review of Psychology* 4 (1953): 439.

12. W.V.O. Quine, in *The Philosophy of Rudolph Camap*, ed. P. Schlipp (LaSalle, Ill: Open Court, 1963), 395.

13. Hilary Putnam, "The Analytic and the Synthetic," in *Minnesota Studies in the Philosophy of Science. Volume 111*, ed. H. Feigl and G. Maxwell (Minneapolis: University of Minnesota Press, 1962).

14. Kant. G*roundwork of the Metaphysics of Morals*.

15. Elliot Eisner, *The Educational Imagination* (New York: Macmillan. 1985).

16. Joseph Schwab, "The Practical: A Language for Curriculum" *School Review* 78, no. 5 (1969): 1-24.

17. Kant, *Groundwork of the Metaphysics of Morals*.

18. Aristotle, *Nicomachean Ethics*, trans. M. Ostwald (New York: Bobbs-Merrill, 1962).

19. Hubert Dreyfus and Paul Rabinow, *Michael Foucault: Beyond Structuralism and Hermeneutics* (Chicago: University of Chicago Press, 1982).