April 2018

# The Validity of Validity in Debra P.: Judicial and Psychometric Perspectives on Test Consequences

Charles Olney
*University of Texas, Rio Grande Valley*

Brent Duckor
*San Jose State University*

The Validity of Validity in Debra P.: Judicial And Psychometric Perspectives on Test Consequences

**OBJECTIVES AND PURPOSE**

1. Situate the logic of validity, using comparative disciplinary perspectives, and identify continuities and discontinuities among (legal and assessment) theorists who seek to address the consequences of testing on historically disadvantaged and vulnerable K-12 student populations.
2. Assess the limits of current judicial treatments of psychometric research and expertise, with an eye toward facilitating judicial incorporation of current expert practices in those fields.
3. Examine the psychometric grounds used to reach judicial decisions in cases (*Debra P.* for example) that challenge the disparate impact of educational testing systems on historically disadvantaged students.

**CORE ARGUMENT & SIGNIFICANCE**

Over the past three decades, a rich and textured debate in the field of psychometrics has produced new theories of testing, grounded in more holistic ideas of validation. These approaches have sought to account for inequality of inputs, and have focused on incorporating the larger social consequences of test usage into the assessment of test validity. Broadly speaking, these developments have been grouped under the label 'consequential validity,' a term designed to illustrate the more complex ways in which validity should be understood to operate. In 1999, this expanded notion of validity was incorporated into the Psychological Testing Standards.

However, despite significant movement *within* these fields, very little appears to have changed in the *legal* treatment of validity as a benchmark for assessing the development of public policy. In the courts, 'validity' still means much the same as it did in the 1970s—with a focus on 'content' or 'construct' validity, and a concern for the predictive value of testing systems. This cramped approach obscures the consequences of using 'validity' as a reference point for educational outcomes.

The lack of effective communication between these two fields impoverishes the work done on both sides. For psychometrics, the lack of translation into legal outcomes exposes the limited effectuality of attempts to delineate 'best practices.' Conversely, for the law, failure to acknowledge developments in the field reveals a troubling gap between the *idea* of deference to expertise and the *reality* of selective incorporation.

Our central claim is that the failure to effectively bridge the gap between the way in which 'validity' is conceived by psychometric and legal communities significantly impedes the effort to provide fair and just educational outcomes. Practitioners on both sides of the divide tend to regard 'validity' as a neutral device, capable of generating discrete packets of information, which may then be ported across fields of knowledge. However, this act of translation often obscures the functions of power at the interstices of educational policy.

Our study analyzes this effect by looking to judicial treatments of validity, in particular the way that psychometric expertise has been integrated into legal analysis. Specifically, we ask: *How are courts thinking about validity in testing and what is ignored about validity discourse in judicial*

*renderings of landmark cases such as Debra P. Why are plaintiffs unable to convince judges about broader consequences of testing on historically disadvantaged students?*

## PHILOSOPHICAL FRAMEWORK

The underlying structure of our argument is a comparative evaluation of two debates within psychometrics and law: the Popham/Shepard debate on consequential validity, and the Hart/Dworkin debate on the nature of legal meaning. We argue that Hart (1958, 1994) and Popham (1997) pursue analogous objectives, in that both emphasize the need for defenses against conceptual leakage. They both demand a firewall between the *production* of facts and the *use* to which those facts are turned. This means treating 'validity' as a subject for experts (within communities of knowledge), which must be translated in whole across the boundaries between communities. Further, while both regard knowledge as essential to good decision-making, they reject the attempt to blur the line between factual/objective claims and moral/political ones. For them, validity is *discrete, neutral, depoliticized*.

Conversely, Dworkin (1986) and Shepard (1993; 1997) represent the inclination to regard knowledge as interpretive, collaborative, and shared across boundaries. For them, the 'validity' of a thing is wrapped up in the manner of its *consequences*. For both Shepard and Dworkin, to speak of a 'valid' system requires formulating a broader framework of justice. As such, they each focus primarily on the *usage* of concepts, and emphasize the need to regard knowledge-claims as multifaceted and permanently unsettled.

The Dworkin-Shepard/Hart-Popham analogy offers significant explanatory potential. By highlighting formal structure of disputes we clarify why some validity arguments tend to be *stickier* and others *smoother* in translation between communities. In particular, expert testimony pitched in Popham's terms is more likely to be well-received by judges influenced by Hart's point of view. In the case of the law, the practical necessities of the job drive judges toward a model of interpretive deference (Posner, 1993).

## METHODS

To explore this effect, our paper employs both qualitative textual analysis and quantitative methods for case analysis using the Lexis-Nexis database.

We initiate the **qualitative** approach to textual analysis by focusing on a key benchmark case: *Debra P. v. Turlington* (644 F. 2d 397 1981), which concerned the legality of high stakes testing as a device for determine the assignation of high school diplomas. At issue in *Debra P.* was a high school exit exam intended to measure general competency of Florida high school students. The students raised both Equal Protection and Due Process claims. This test, the students argued, was unfair and biased, disproportionately affecting minority students and those already ill served by the state educational system.

The 5th Circuit's decision in this case found in favor of the students, but only in a narrow sense. The problem was not the broader equality concerns raised by such tests, but merely that this *specific* test had been insufficiently *validated*. The state had not shown a proper correlation between the content of the exam and the purposes to which they used the results it generated. In order to legitimize a test ('validate' the legality of the punitive effect), the state must show that the test has 'curricular validity.'

Notably, the structure of this "validation" approach was endorsed by expert testimony from psychometricians on both sides of the case. By close textual analysis (including review of evidence from depositions) we identify how evolving Standards that would have advanced the notion that validity as *on-going*, a matter of *degree*, and based on carefully assembled *types of evidence* to support an intended use was largely ignored. Moreover, both plaintiffs and defendants agreed that the appropriate measure of due process in this case was the validity of the purpose to which the test is employed. This agreement allowed the court to narrow its legal responsibilities, for example, by demanding deeper deliberation on test consequences and use (Messick, 1988, 1995) and instead delegated the task of *adjudicating* validity to the external experts.

*Debra P.* provides an excellent case study for the discontinuities in treatment of validity, as well as a useful framework for connecting the twin debates within these two communities. By our methodological approach, we trace how *Debra P.* signals a rejection of the Dworkinian approach, which would require treating validity as a *relational* concept that necessarily cross-pollinates with the constitutional rights questions raised by the plaintiffs.

Our **quantitative** approach takes *Debra P.* as a starting point for further judicial treatments of validity in the context of testing. In order to assess the use and interpretation of these terms, we have gathered an extensive set of judicial opinions and law reviews from the past three decades, which will be coded using a Python-based linguistic analysis tool to assess: 1) the frequency of relevant terminology over time—as concepts have developed within psychometric fields, as they are translated into judicial treatments 2) the relational usage of different terms—when validity is discussed, how it is used and 3) connected key phrases and points of reference within case law archives.

**RESULTS**

Cross-sectional analyses of these case documents yields a relatively fine-grained picture of where validity sits in current legal approaches. But even a relatively simplistic initial test (searching keywords on Lexis-Nexis over 100 years) exposes a stark discrepancy, with *hundreds* of judicial opinions employing the language of 'content' or 'construct' validity, and *no* substantive references to the newer standards.

The early results from our database review indicate a strong correlation in judicial opinions between validity concepts and procedural due process claims. This finding also supports our intuition that *Debra P.* is an exemplar of judicial attitudes toward psychometric concepts, despite their evolution. Rather than regarding validity as a contact point for substantive or equality-based concerns, courts appear to regard it as merely a technique for navigating procedural thickets.

Our findings suggest that judges are using their own conceptions of validity (in the legal sense) as bridges for understanding and incorporating treatments of validity produced by external fields. Absent a well-developed theory to situate an idea like consequential validity, judges will naturally employ their existing analytic architecture (developing through centuries of debate over the nature of law, and generally framed through the language of 'due process'). In the instance of *Debra P.* the consequences of this choice became apparent when the case was reheard in 1984.

In effect, the court utilized a narrow concept of validity to shield itself from the obligations of directly adjudicating the thorny constitutional rights issues at stake in this case.

**IMPLICATIONS**

When it comes to issues like those raised in *Debra P.* and its progeny—the outcomes of educational policy, high-stakes assessment, constitutional protections of due process and equal protection—the existence of a parallel language of validity creates an easy outlet. By reducing constitutional protections to questions of 'test validity,' courts shift the burden of judgment to outside experts. In doing so, these courts import a technique of validation, and thereby shield themselves from the troublesome necessity of adjudicating *justice* claims. The tests are valid or not (as determined by experts), and their uses are good or bad (as determined by political actors). On either side, courts are freed from the responsibility of final judgment. They have effectively re-distributed the "validity burden" back onto society (Superfine, 2004).

Left out of this juridical reasoning process, however, is a recognition that these external psychometric experts are only articulating *one specific idea* of validity. Namely: the "Popham model," which is far more translatable to settled case law precisely because it shares the same presumption that final judgments can be separated from discrete packets of knowledge. In effect, the convergence of (Popham's explicit and Hart's implicit) theories of validity shield practitioners of law from a serious reckoning with the problem that any given theory of validity must *itself* be subjected to scrutiny (Messick, 1989, 1994; Kane, 1992, 2013; Haertel & Herman, 2005).

One crucial effect of this tendency is to devalue perspectives that emphasize broader and more interpretive models of validity. We note that Shepard seems to have 'won' her debate with Popham—as reflected in new standards for validity which include *consequences* and *fairness* (AERA, APA, NCME, 1999/2014). But where testing policy encounters legal judgment—where the rubber meets the road, and where most test construction and implementation challenges actually manifest—Popham's logic still dominates. Courts tend to side mostly with Hart: seeking to decide narrow questions of legal meaning, and to avoid responsibility for determining the *political* implications of their decisions.

However, this should not be treated as a *fait accompli*. While it is beyond the scope of our paper to propose alternative models for legal treatment, one key implication of this argument is that the process is a two-way street. Recognizing the parallel structures of jurisprudential reasoning with the law and psychometric reasoning with Testing Standards might allow for a productive reconfiguration of *both* sides, where the impulse toward interpretive, collaborative, expansive knowledge-frameworks can better attach itself to the systems of legal judgment.

**REFERENCES:**

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing (2nd ed.). Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing (1st ed.). Washington, DC: Author.

Dworkin, R. (1986). *Law's Empire*. Cambridge: Belknap Press.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity argument for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (The 104th Yearbook of the National Society for the Study of Education, Part 2) (pp. 1-34). Malden, MA: Blackwell Synergy.

Hart, H.L.A. (1958). Positivism and the separation of law and morals. *Harvard Law Review*, 71(4), 593-629.
Hart, H.L.A. (1994). *The Concept of Law* (2nd edition). Oxford: Clarendon Press.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527-535.

Kane, M. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, 50(1), 115-122.

Messick, S. (1988). Meaning and values in test validation: The science and ethics of assessment. *ETS Research Report Series*, *1988*(2).

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Messick, Samuel. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50 (9), 741-749.

Popham, W. J. (1997). Consequential Validity: Right Concern—Wrong Concept. *Educational Measurement: Issues and Practice*, 16, 9-13.

Posner, R. (1993). *The Problems of Jurisprudence*. Cambridge: Harvard University Press.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling- Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

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

Superfine, B.S. (2004). Education, Legislation, Law & Social Science Research: At the Intersection of Law and Psychometrics: Explaining the Validity Clause of No Child Left Behind. *Journal of Law & Education*, 33, 475-513.