Encyclopedia of Evaluation

Validity

Contributors: Nigel Norris
Editors: Sandra Mathison
Book Title: Encyclopedia of Evaluation
Chapter Title: "Validity"
Pub. Date: 2005
Access Date: October 15, 2013
Publishing Company: Sage Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780761926092
Online ISBN: 9781412950558
DOI: http://dx.doi.org/10.4135/9781412950558.n567
Print pages: 440-443

http://dx.doi.org/10.4135/9781412950558.n567

Writing a general statement about validity in evaluation is a hazardous business. The field of evaluation is so diverse and complex and it has such an array of models, approaches, forms, and disciplinary homes that generalizing about evaluation is invariably a potentially foolhardy enterprise. Moreover, this is historically disputed territory. Validity is related to truth. They are members of the same family, so to speak. Truth, however you cut it, is an essentially contested concept with little agreement about what constitutes the right basis for truth-seeking activities. Of course, much has already been said about validity. There are various classic explications of validity from some of evaluation's most notable theorists: Donald Campbell, Thomas Cook, Lee J. Cronbach, Egon Guba, Ernest House, Yvonna Lincoln, Michael Patton, Michael Scriven, and Robert Stake, to name a few.

Much of the justification for doing evaluation is that it can at least offer approximations to the truth and help discriminate between good and bad, better and worse, desirable and less desirable courses of action. It is not surprising, therefore, that one of the defining problems of the field of evaluation remains the validation of evaluative judgments. There are three main issues that have beset discussions about validity in evaluation. The first issue has to do with the nature and importance of generalization and the ways in which evaluation can and should support social decision making. In turn, this issue depends on the assumptions made about the objects of evaluation (practices, projects, programs, and policies), how they are theorized, and the political context of evaluation. The second issue is the extent to which nonmethodological considerations, such as fairness, social responsibility, and social consequence, should inform discussions about validity. The third issue, and seemingly the most intractable, is the extent to which it is possible to have a unified conception of validity in evaluation. Given the methodological and substantive diversity that now characterizes the transdisciplinary field of evaluation, is it possible to have common standards for and a shared discourse about validity? This issue is acutely felt in debates about whether the traditional discourse of scientific validity in quantitative research is relevant to qualitative approaches to evaluation.

The publication in 1963 of Donald Campbell and Julian Stanley's chapter “Experimental and Quasiexperimental Designs for Research on Teaching” is probably the single most significant landmark in the conceptualization of validity. This and Campbell's later work with Thomas Cook, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, published in 1979, were the touchstones for most, if not all, discussions about validity in evaluation and applied research more generally. Campbell and his colleagues introduced a nomenclature and framework for thinking about validity and constructing the research conditions needed to probe causal relationships. Central to this experimental tradition has been the development of strategies (designs) for controlling error and bias and eliminating plausible rival hypotheses or explanations.

Discussions about validity in the experimental tradition have resulted in a shared language about the threats to validity associated with different research designs and types of validity (e.g., internal validity, external validity, statistical conclusion validity, **[p. 440 ↓]** construct validity). The distinction between internal and external validity has been of particular importance in discussions about validity. Internal validity usually refers to the validity of inference from and confined to a particular study or domain of investigation. It is about the validity of statements or judgments about the case or cases under investigation. It addresses the question: Did the treatment make a difference in this experimental instance? By contrast, external validity refers to whether inferences or judgments from a study or domain of investigation apply to other populations, settings, or times. It is about whether findings generalize. In classic theories of experimental design, internal validity was taken to be the sine qua non of external validity. This was because establishing internal validity was regarded as the basic minimum for the interpretability of experiments. Thus generalization depended first and foremost on establishing that the findings of the study were true for practical and methodological purposes. The codification of various threats to validity has been central to the experimental tradition and has proved useful in the evaluation of social and educational intervention programs. Threats to validity indicate some of the prototypical rival hypotheses or alternative explanations for whether the program is in fact responsible for changes in outcomes and whether it will generalize. A list of threats to validity would often include the following:

A major strength of experimentalism—its focus on creating the conditions to test “causal” claims and infer the “causal” relationships between specific variables in a domain of investigation—is also its weakness, at least in the context of program and policy evaluation. The precedence given to internal validity undermines the usefulness or applicability of an evaluation. Threats to internal validity are reduced by rigorous control over the experimental conditions. Rigorous control, however, tends to limit the application of evaluative conclusions because the context of the experiment is unrepresentative of the contexts of application. Put simply, there is often a trade-off to be made between validity-enhancing and utility-enhancing evaluation designs. The experimentalist tradition places too much emphasis on establishing causal relationships between variables, thus limiting the extent to which it is possible to confidently extrapolate from an evaluative study to other implementation strategies, settings, populations, and times. This is significant. When a program or social practice is developed in a new setting, it is adapted by different people at different times for different populations. It is not the same program or practice. The difference between the original and the intended copy may be substantial or slight, but important differences there will be. The issue here is whether evaluation should strive to support formal generalization or whether it should be more concerned with the potential for wider transferability and with supporting the capacity to learn from the experience of others. A related limitation of experimentalism is lack of attention to the reasons that programs work (the so called black-box problem). Advocates of “realist” or theory-driven approaches to evaluation have argued that the prominence given to internal validity has been at the expense of developing theories about the underlying causal mechanisms that generate outcomes in particular contexts.

The question of whether validity is solely a technical matter has become a critical issue in evaluation, as it has in educational measurement. In his prescient and influential book *Evaluating With Validity*, Ernest House argued that technical considerations alone cannot fully address the problem of bias in evaluation. Rejecting the primacy of operational definitions of validity, he said that the validity of an evaluation depends on whether it is true, credible, and normatively correct. In short, validity in evaluation is not only concerned with technical standards but with fairness and social consequences. Developments in educational measurement theory have similarly stressed the importance of taking into account value implications and action outcomes when considering the validity of test scores for particular uses.

Involving stakeholders in the validation of an evaluation is one way to move beyond technical validity issues. Participant validation involves different interest groups, particularly intended users and those whose work and lives are represented in an evaluation, in commenting on matters such as the accuracy, fairness, relevance, comprehensiveness, truthfulness, and credibility of an evaluative account and its conclusions and recommendations. By its nature, participant validation poses practical constraints, notably for large-scale, systemwide evaluations. There are risks associated with disagreement and competition among stakeholders that have to be managed. Sponsors or executive decision makers, for example, may object that disseminating draft evaluation reports to others runs counter to their own proprietorial rights and is a breach of the customer-contractor relationship that typifies much commissioned evaluation. Nevertheless, participant validation has an important role to play in strengthening confidence in an evaluation and building consensus around its conclusions and recommendations. Perhaps most important, it widens the social basis of validation and helps protect against evaluation being seen as little more than a bureaucratic service to the politically powerful.

Validity is a contested concept. It is contested because there are disputes about its meaning, how it should be pursued, its relevance to evaluative inquiry, and its intelligibility in the face of postmodern, antifoundational critiques of traditional research. Whether it is because of a crisis of representation or disenchantment with science or, more mundanely, a consequence of the political economy of academic disciplines, validity has become a vexing and confusing concept.

Some want to give validity up altogether: It is too modern, too prescriptive, too concerned with demarcating the acceptable from the unacceptable; also, it is part of an epistemologically repressive or coercive regime of truth, authoritarian and distasteful. This is not a position that can be sustained by those doing evaluation, however. In practice, evaluation can fulfill its promise of being useful only if the knowledge it produces is credible, and such credibility as it can muster depends in part on claims to truth. There are other demands to change the meaning of validity and redefine its various methodological expressions, personalize and politicize its meanings, and develop new criteria for quality in evaluation: trustworthiness, goodness, credibility, authenticity, efficacy, and generative or transformational power, for example. What lies behind some of these attempts to recast validity is a belief that the traditional paradigms of social science and classical conceptions of validity have failed applied research and evaluation because they are limited and conservative, binding us to traditional values and a singular version of reality and blinding us to alternative perspectives and multiple truths. These arguments and ideas about validity have largely been explored and elaborated in the context of debates about the fundamental differences between quantitative and qualitative research. Quantitative methodologies have been associated with positivist or scientific epistemology. By contrast, qualitative methodologies have been associated with naturalistic or constructivist epistemologies. Traditional concepts of validity have been rejected by some evaluation theorists of a qualitative persuasion because of a belief that quantitative and qualitative methods of knowing belong to different and incommensurate paradigms. It is hardly surprising that these issues are unresolved. Still, the assumption that methodology is inextricably linked to coherent metaphysical paradigms is probably overstated. Similarly, the belief that qualitative and quantitative ways of knowing and validating knowledge are incompatible because they originate from different and incommensurate worldviews is, in some measure, an unsubstantiated empirical claim. The danger with paradigmatic representations of methodology is that they can easily become caricatures, unanchored from the experience and practice of evaluation. However, even if there are no fundamental reasons to conclude that qualitative and quantitative ways of knowing and validating are incommensurate, there remains a feeling that the concepts of validity associated with **[p. 442 ↓]** experimentalism are not as meaningful or productive for evaluations that draw on ethnography, case study, or naturalistic inquiry for their modus operandi. In these evaluations, the judgments of participants and users as to the quality of an evaluation are often a crucial element in the process of validation.

Evaluation is difficult to do. This is because its purpose is to inform decision making and social understanding. Evaluation is meant to be useful, and mostly this means useful in the short run, matching the timing of organizational and political decision making. Inevitably this requires striking a balance between timeliness, utility, and methodological ideals. What makes this hard is that the context of social valuation and decision making is often overtly political. Evaluation can impinge on interests and values in unpredictable ways. Political standing, reputations, and resource allocations may be at stake. Negative evaluations can be used by some at the expense of others associated with a policy, program, project, or practice. Evaluation findings are a possible resource for securing advantage, and correspondingly, they are a possible threat. Evaluation may intrude into autonomous spheres of action and semiprivate workspace, and the knowledge it produces provides opportunities for control. Evaluation is best seen as an activity that can affect the distribution of power and resources. That evaluation is deeply imbued with politics has become something of an accepted truism. Issues of territory, competition, hierarchy, reputation, and privacy impinge on the work of evaluators to a greater extent than is usual in research. The political context can undermine an evaluation, affecting the kind and quality of evidence available to it, the validity of findings, how they are received, and what can be learned. As a consequence, specific measures must be taken to foster the credibility and openness necessary for the successful conduct of evaluation. In this context, validity has an important job to do. It helps stand guarantor for the independence and impartiality needed for trust in an evaluation. It is essential in the defense against the robust criticism that can be expected when evaluators report unwelcome news.

Validity is not a property of approaches to or methods of evaluation. At its most rudimentary, validity refers to the reasons we have for believing truth claims, what Dewey called “warranted assertibility.” These truth claims may take many forms: statements of fact, conclusions, representations, descriptions, accounts, theories, propositions, generalizations, inferences, interpretations, and judgments. Irrespective of their form, what is important is the warrant evaluators have for the claims they make. *Warrant* here refers to justifying conditions, reasons, and arguments. Validity speaks to why we should trust a representation of reality or an evaluative account. One way to think about validity is not so much cutting the difference between truth and its pretenders, but providing us with a sense of the limitations of knowledge claims. Despite its objective connotations, validity is important in evaluation precisely because of the socially constructed, fallible, provisional, and incomplete nature of knowledge and the personal and political propensity to act as if this were not the case.

Nigel Norris

http://dx.doi.org/10.4135/9781412950558.n567

Further Reading

Bickman, L. (Ed.). (2000) Validity and social experimentation . Thousand Oaks, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1963) Experimental and quasi-experimental designs for research on teaching . In N. L. Gage (Ed.), Handbook of research on teaching (pp. 171–246) . Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979) Quasi-experimentation: Design and analysis issues for field settings . Boston: Houghton Mifflin.

House, E. R. (1980) Evaluating with validity . Beverly Hills, CA: Sage.