Occasional Report No. 11

CSE
Report
No. 11

COMMENTS ON PROFESSOR BLOOM'S PAPER ENTITLED "TOWARD A THEORY OF TESTING WHICH INCLUDES MEASUREMENT-EVALUATION-ASSESSMENT"

Gene V. Glass

Center FOR THE
Study of
Evaluation
OF INSTRUCTIONAL
PROGRAMS

University of California, Los Angeles, September 1968

COMMENTS ON PROFESSOR BLOOM'S PAPER ENTITLED "TOWARD A THEORY OF TESTING WHICH INCLUDES MEASUREMENT-EVALUATION-ASSESSMENT"

Gene V. Glass

University of Colorado

From the Proceedings of the

SYMPOSIUM ON PROBLEMS IN THE EVALUATION OF INSTRUCTION

University of California, Los Angeles
December, 1967

M. C. Wittrock, Chairman

Sponsored by the Center for the
Study of Evaluation of Instructional Programs

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

CSEIP Occasional Report No. 11, September 1968
University of California, Los Angeles

COMMENTS ON PROFESSOR BLOOM'S PAPER ENTITLED "TOWARD A THEORY OF TESTING WHICH INCLUDES MEASUREMENT-EVALUATION-ASSESSMENT"

Gene V. Glass

First, I want to suggest that syntheses of the concepts Bloom identifies under reliability and norms do exist. Second, I want to offer an answer to the question of whether we ought to synthesize measurement, evaluation, and assessment into a theory of testing.

Cronbach, Gleser and Rajaratnam's liberalization of reliability theory known as "generalizability theory" has, to my mind, synthesized diverse concepts and methods of reliability assessment. Variability in performance arising from any universe of influences (whether "readers," "time," "types of instrument," or "content") can be estimated in a factorially designed G-study (generalizability-study) and used to determine the generalizability of measures taken in some practical application (a decision-study). In a G-study, lack of generalizability from an observed score to a universe score is conceived of as the interaction (in the Fisherian sense) of examinees with the factors over which generalization is sought. It is possible with present techniques to determine, for example, a lower bound to the correlation of an observed "neuroticism" score derived by averaging the ratings of two psychiatrists (reader reliability)*, on three occasions (instrument stability) over

---

* The terms in parentheses are Bloom's.

six symptoms (sampling or congruence reliability) with a universe "neuroticism" score defined to be the average score for an examinee over universes of the psychiatrists, occasions, and symptoms sampled (Gleser, Cronbach, Rajaratnam, 1965). I don't think we need to look far beyond generalizability theory for a synthesis of reliability notions.

Though I can suggest no conceptual synthesis of the five types of norms Bloom identifies, there is, I believe, a technological synthesis worth pursuing. We have long acknowledged that the validity of a test is always for a particular purpose. However, we have been slower to acknowledge that one always has a particular purpose in mind when referring a test score to a set of norms. Seldom is this purpose to determine the status of a person in an anonymous and ill-defined norm group. For example, in counseling an 18-year-old Negro drop-out on making a vocational choice, it may be irrelevant to determine the status of his mechanical aptitude score with respect to the general population in getting a job--labor unions being what they presently are.

Instead of publishing test norms as is presently done, perhaps we should record on magnetic tape test performance--on more than one occasion--along with extensive biographical, social, and psychological data for a large sample of persons. Anyone seeking normative information would need only to specify the composition of the norm group which suits his purpose. For example, one might request the norms on the Bennett Test of Mechanical Comprehension for Negro males between the ages of 16 and 35 without a high-school

diploma and living in cities of over 500,000 population. The problems of programming a computer to search the available data and produce norms for such a group seem insignificant.

A practical synthesis of the four types of validity Bloom lists seems less imminent, although Rozeboom (1966) has recently examined these issues and exposed greater unity in our diverse notions of validity than we might have imagined existed.

A synthesis of the terms Bloom lists seems clearly possible. However, the primary question he has posed in his paper is, "How do we synthesize three separate data-gathering activities into one theory of testing?" To seek an answer begs the question that they ought to be synthesized, a question we should examine carefully.

Should one attempt to synthesize measurement, evaluation, and assessment into a theory of testing, a theory of "gathering and processing evidence about human behavior...for purposes of understanding, predicting, and controlling future human behavior?" I think not. Achieving the synthesis would misdirect the development of one of the constituents, namely, evaluation, and further subvert the already abused goal of that activity. A synthesis would redirect the development of measurement and assessment as well; but in the context of Bloom's paper and this symposium, I wish to deal only with what such a merger would mean to evolving strategies of evaluation.

Bloom and I have roughly the same thing in mind when we think of evaluation. It deals with gathering evidence about the effects of instruction; it has to do with whether a curriculum is

doing its job, etc. But we define evaluation differently; we ascribe different roles to it; and we see it developing along different lines. To Bloom, evaluation is the appraisal of change in students due to instruction, and its major quest is the "identification of learning experiences...which produce significant changes in individuals..." I am more sympathetic with Scriven's declaration that it "consists...of the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings, and in the justification of (a) the data-gathering instruments, (b) the weightings, and (c) the selection of goals" (Scriven, 1966a, p. 40). Tyler has defined evaluation similarly: "'Evaluation' designates a process of appraisal which involves the acceptance of specific values and the use of a variety of instruments of observation, including measurement, as bases for value-judgments" (Tyler, 1951, p. 48).

The current meaning of the term "evaluation" in several recent writings and in federal legislation is that it is the gathering of empirical evidence for decision-making and the justification of the decision-making policies and the values upon which they are based. (See Stake, 1967b, and Stufflebeam, 1966.) Evaluation can contribute to the construction of a curriculum, the prediction of academic success, or the improvement of an existing course. But these are <u>roles</u> it can play and not its <u>goal</u>. The goal of evaluation must be to answer questions of selection, adoption, support, and worth of educational materials and activities. It must be directed toward answering questions like, "Are the benefits of this curriculum worth its cost?" or, "Is this textbook superior to its competitors?"

The contrast between the view of evaluation as gathering test data about how well a curriculum accomplishes its objectives and the above view is the contrast between the first and second dictionary definitions of "evaluation": "to find the amount or numerical value of something (e.g., evaluate $f(x) = \log(x)$ at $x = 2$)" versus "to appraise the worth or value of something." The latter function corresponds to the goal of educational evaluation; the former function corresponds to one of many roles it can play. In the past, we have avoided the goal of evaluation with its inherent threat to teachers, administrators, and curriculum developers and have concentrated on one or more of the non-threatening roles evaluation can play. We have measured performance without questioning merit. Scriven claimed that "if we do not know that (and usually how)... performance bears on merit, it is a travesty to refer to the measurement of it as evaluation: and exactly this travesty is involved in a great deal of curriculum evaluation where no defensible conclusions about merit can be drawn from the kind of data that is so earnestly gathered" (Scriven, 1966b, pp. 6-7). "Evaluation," which is no more than the measurement of whether a curriculum attains its stated objectives, is guilty of values-relativism, i.e., the acceptance of the idea that any objective is as valuable as any other.

Aren't empirical facts irrelevant to questions of value? The answer to the last question is no. We cannot force knowledge into a facts-values dichotomy. "Our image of value and our image of fact are symbiotic. They are part of a single knowledge structure, and it is naive in the extreme to suppose that they are independent"

(Boulding, 1967, p. 886). In view of what cultural anthropologists and physiologists have learned about man, it is no longer possible rationally to value the "pure Aryan race." In education there would exist no basis for valuing recall of isolated facts about history if it could be demonstrated empirically that such recall is unrelated to scholarship in history of intelligent citizenship.

Now what part do our empirical methodologies have to play in a theory of evaluation that assesses value? What do testing, psychometrics, experimental design, survey research, etc., have to do with questions of merit or worth? First, the objectives of instruction need justification. Evaluation should seek to determine their comparative merit or worth empirically while philosophy seeks logical justification. Occasionally, both empirical science and rational thought will have to seek justification for the very values upon which course objectives are based. Second, if we are to approach rationality in decision-making, we must be able to measure the values of the decision-makers and the value-weights they ascribe to the outcomes of instruction. It is here that psychometric methods of scaling and factor analysis can make a contribution (Taylor, 1966; Maguire, 1967). Third, once the objectives of instruction and value-weights for criteria are determined and justified, comparative experimentation will arbitrate questions of relative merit and worth.

It may seem that too much is being made of an idiosyncratic definition of "evaluation," especially since Bloom expressed little interest in the "accuracy or the meaningfulness of the terms

measurement, evaluation, and assessment." I cannot help being greatly concerned with the meanings of words and, more importantly, with how they influence action. I frequently come into contact with educationists whose most energetic efforts are victimized by a semantic confusion. By happenstance, habit, or methodological bias, they may, say, label the trial and investigation of a new curriculum or organizational plan with the epithet "experiment" instead of "evaluation." I am convinced that the inquiry they conduct is different for their having chosen to call it an "experiment" and not an "evaluation." Their choice predisposes the literature they read (it will deal with experimental design), the consultants they call in (only acknowledged experts in designing experiments), and how they report the results (always in the best tradition of the Journal of Experimental Psychology). In some instances, none of these paths will lead to relevant data or promote rational decision-making. The crucial data may be "soft" instead of "hard;" they may deal with the instructional materials or parents' reactions to them instead of students' behavioral outcomes.

To be sure, many psychometricians, measurement specialists, and methodologists will want to have nothing to do with evaluation as depicted here. Some will dismiss it because it attempts to deal with "philosophical questions" and questions of value (Carroll, 1965, p. 253), as indeed it does. Others will maintain that they are free to investigate whatever they please and that broadening the definition of evaluation to include telling Congress how to spend money or administrators how to choose curricula does not

please them. For them, evaluation will stop with the attribution of behavioral change to instructional experience.

Others will want to consider seriously whether we can afford to go on merely playing one of the roles of evaluation instead of trying to fulfill its goals. The curriculum movement, the entry of industry into the production of educational materials, the development of new organizational plans, etc., are confronting educationists with choices, with decisions. The entire rapidly changing nature of education is pressing the goal of evaluation upon us. Of course, we are free not to play any part at all in this revolution. We who identify with "measurement" and "psychometrics" seem to feel (with poorly disguised satisfaction) that we stand among the angels. But my own self-satisfaction has been disturbed by my perception that increasingly educationists see us as standing among the Philistines. We react with conventional solutions (regression analysis, reliability and validity estimation, factor analysis, item analysis, etc.) to their earnest requests for help in facing unprecedented problems in decision-making; and we are even rather smug about how we think we have helped them. I know of an instance of a corporation contracting with a curriculum development project for a handsome sum to perform an "evaluation" which eventuated in nothing more than an item analysis of an achievement test.

Should we attempt to synthesize measurement, evaluation, and assessment into a theory of testing? It would seem preferable to allow evaluation (with its emphasis upon judging the overall merit of an educational enterprise) to develop along its-own lines

independent of the development of testing (with its emphasis on measuring and predicting human characteristics). Instead of thinking of measurement, evaluation, and assessment as elements of some unitary, higher-order theory, we might profit more from thinking of them as existing and growing individually in a relationship of mutual assistance and support. Evaluation will borrow from measurement and assessment to suit its needs as it will borrow from all the social sciences (Stake, 1967a); the other disciplines can be expected to do likewise. It has been a struggle to establish educational evaluation with an identity apart from educational testing; it would be a shame to see it engulfed again by the more mature discipline of testing.

REFERENCES

Boulding, K. E. Dare we take the social sciences seriously? American Psychologist. 1967, 22, 879-887.

Carroll, J. B. School learning over the long haul. In J. D. Krumboltz (Ed.), Learning and the educational process. Chicago: Rand McNally, 1965.

Cronbach, L. J., Rajaratnam, N. , & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Maguire, T. O. Value components of teachers' judgments of educational objectives. Unpublished doctoral dissertation, University of Illinois, 1967.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. Generalizability of stratified-parallel tests. Psychometrika, 1965, 30, 39-56.

Rozeboom, W. W. Foundations of the theory of prediction. Homewood, Illinois: Dorsey Press, 1966.

Scriven, M. The methodology of evaluation. In R. E. Stake, (Ed.), AERA monograph series on curriculum evaluation. Chicago: Rand McNally, 1966a.

Scriven, M. Value claims in the social sciences. Publication No. 123. Lafayette, Indiana: Social Sciences Education Consortium, 1966b.

Stake, R. E. An emerging theory of evaluation--borrowings from many methodologies. Paper presented at the Annual Meeting of the American Educational Research Association, New York, February 1967a.

Stake, R. E. The countenance of educational evaluation. Teachers College Record, 1967b, 68, 523-540.

Stufflebeam, D. L. Evaluation under Title I of the Elementary and Secondary Act of 1965. Paper read at the Evaluation Conference sponsored by the Michigan State Department of Education, East Lansing, Michigan: January, 1966.

Taylor, P. A. The mapping of concepts. Unpublished doctoral dissertation, University of Illinois, 1967.

Tyler, R. W. The Functions of measurement in improving instruction. Chapter 2 in E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951.

PAPERS FROM THE SYMPOSIUM ON PROBLEMS IN THE EVALUATION OF INSTRUCTION

M. C. Wittrock, Chairman

THEORY OF EVALUATION OF INSTRUCTION

Benjamin Bloom, Principal Speaker University of Chicago	*Toward a Theory of Testing Which Includes Measurement-Evaluation-Assessment,* Occasional Report No. 9
Michael Scriven, Discussant University of California, Berkeley	*Comments on Professor Bloom's Paper,* Occasional Report No. 10
Gene Glass, Discussant University of Colorado	*Comments on Professor Bloom's Paper,* Occasional Report No. 11
J. P. Guilford, Discussant University of Southern California	*Comments on Professor Bloom's Paper,* Occasional Report No. 12
Robert Glaser, Principal Speaker University of Pittsburg	*Theory of Evaluation of Instruction,* Occasional Report No. 13
Robert Stake, Discussant University of Illinois	*Comments on Professor Glaser's Paper,* Occasional Report No. 14
Arthur Lumsdaine, Discussant University of Washington	*Comments on Professor Glaser's Paper,* Occasional Report No. 15

INSTRUCTIONAL VARIABLES

Robert Gagné, Principal Speaker University of California, Berkeley	*Instructional Variables and Learning Outcomes,* Occasional Report No. 16
Richard Anderson, Discussant University of Illinois	*Comments on Professor Gagné's Paper,* Occasional Report No. 17
Leo Postman, Discussant University of California, Berkeley	*Comments on Professor Gagné's Paper,* Occasional Report No. 18

CONTEXTUAL VARIABLES

Dan Lortie, Principal Speaker University of Chicago	*The Cracked Cake of Educational Customs and Emerging Issues in Evaluation*, Occasional Report No. 19
C. Wayne Gordon, Discussant University of California, Los Angeles	*Comments on Professor Lortie's Paper*, Occasional Report No. 20
N. L. Gage, Discussant Stanford University	*Comments on Professor Lortie's Paper*, Occasional Report No. 21

CRITERION VARIABLES

Samuel Messick, Principal Speaker Educational Testing Service	*The Criterion Problem in the Evaluation of Instruction: Assessing Possible Not Just Probable Intended Outcomes*, Occasional Report No. 22
Paul Blommers, Discussant University of Iowa	*Comments on Dr. Messick's Paper*, Occasional Report No. 23
Leonard Cahen, Discussant Educational Testing Service	*Comments on Dr. Messick's Paper*, Occasional Report No. 24
Marvin Alkin, Principal Speaker University of California, Los Angeles	*Evaluating Cost-effectiveness of Instructional Programs*, Occasional Report No. 25
Marvin Hoffenberg, Discussant University of California, Los Angeles	*Comments on Professor Alkin's Paper*, Occasional Report No. 26
John Bormuth, Discussant University of Chicago	*Comments on Professor Alkin's Paper*, Occasional Report No. 27

METHODOLOGICAL ISSUES

David Wiley, Principal Speaker University of Chicago	*Design and Analysis of Evaluation Studies*, Occasional Report No. 28
Chester Harris, Discussant University of Wisconsin	*Comments on Professor Wiley's Paper*, Occasional Report No. 29

Theodore Husek, Discussant
University of California,
Los Angeles

Comments on Professor Wiley's Paper, Occasional Report No. 30

Martin Trow, Principal Speaker
University of California,
Berkeley

Methodological Problems in the Evaluation of Innovation, Occasional Report No. 31

Eugene Litwak, Discussant
University of Michigan

Comments on Professor Trow's Paper, Occasional Report No. 32

David Nasatir, Discussant
University of California,
Berkeley

Comments on Professor Trow's Paper, Occasional Report No. 33