TEST VALIDITY

Edited by
Howard Wainer
Henry I. Braun

LEA

TEST VALIDITY

Edited by

Howard Wainer

Henry I. Braun

Educational Testing Service

1988

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS

Hillsdale, New Jersey Hove and London

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

Library of Congress Cataloging-in-Publication Data

Test validity.

Papers of a conference entitled "Test validity for the 1990's and beyond," held May 28–29, 1986 at the Educational Testing Service, Princeton, N.J.
Bibliography: p.
Includes index.
1. Examinations—Validity—Congresses. 2. Educational tests and measurements—Congresses. I. Wainer, Howard. II. Braun, Henry I., 1949– . III. Educational Testing Service.
LB3060.7.T47 1988 371.2'6'013 87-8986
ISBN 0-89859-997-0

Printed in the United States of America
10 9 8 7 6 5 4 3 2

To Laurent, Ilana, and Nora,
whose validity needs no testing

List of Contributors

WILLIAM H. ANGOFF Distinguished Research Scientist, Educational Testing Service, Princeton, NJ 08541

HENRY I. BRAUN Director, Statistical and Psychometric Research and Services, Educational Testing Service, Princeton, NJ 08541-0001

LEE J. CRONBACH 16 Laburnum Rd., Atherton, CA 94025

RICHARD P. DURÁN Department of Education, University of California at Santa Barbara, Santa Barbara, CA 96016

BERT F. GREEN Department of Psychology, Johns Hopkins University, Charles and 34th Streets, Baltimore, MD 21218

LAWRENCE V. HEDGES Department of Education, University of Chicago, Chicago, IL 60637

PAUL W. HOLLAND Director, Research Statistics Group, Educational Testing Service, Princeton, NJ 08541

SAMUEL MESSICK Vice President, Research, Educational Testing Service, Princeton, NJ 08541

BENGT MUTHÉN Graduate School of Education, University of California, Los Angeles, CA 90024

JAMES W. PELLEGRINO Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106

DONALD B. RUBIN Department of Statistics, Harvard University, Cambridge, MA 02138

FRANK L. SCHMIDT School of lndustrial Relations, Phillips Hall, Room 569, University of Iowa, Iowa City, Iowa 52242

LYNN STEINBERG Department of Psychology, Indiana University, Psychology Building, Bloomington, IN 47405

ROBERT J. STERNBERG Department of Psychology, Yale University, Box 11A, Yale Station, New Haven, CN 06520

DOROTHY T. THAYER Research Statistics Group, Educational Testing Service, Princeton, NJ 08541

DAVID THISSEN Department of Psychology, University of Kansas, 426 Fraser Hall, Lawrence, KS 66045

HOWARD WAINER Principal Research Scientist Educational Testing Service, Princeton, NJ 08541

WARREN W. WILLINGHAM Assistant Vice President, Research, Educational Testing Service, Princeton, NJ 08541

Preface

On May 28–29, 1986, a conference entitled "Test Validity for the 1990s and Beyond" was held at the Educational Testing Service in Princeton, New Jersey. The purpose of this conference was to provide an assessment of the current state of research and practice in test validity and to indicate future directions. With testing expected to undergo significant changes in the years ahead, driven in part by demands of users and in part by advances in technology, we felt it would be useful to bring together some eminent scientists closely involved with test validity and ask them to contribute their views on the subject.

In our opinion, the conference was very successful with a broad range of interests and opinions expressed. This volume is intended to make the conference presentations available to a wider audience.

The chapters contained herein were prepared especially for this volume. The talks given at the conference were based upon the papers but were not necessarily identical to them. At the conference we were fortunate to have Professor Donald B. Rubin to discuss the presentations. An edited transcription of his remarks is appended to Section IV. We decided to include his comments because his wide-ranging remarks often provided both context and generality for the invited contributions.

The volume contains all the proceedings of the conference with one notable exception. One afternoon was devoted to an abbreviated mock trial based on an actual court case that focused on the validity of pre-employment tests used in screening applicants for places in a firefighter academy. Although tests have been increasingly involved in litigation, few of those attending the conference had actually participated in or viewed legal proceedings of this kind. Presenting this trial allowed a close-up view of the nature of the legal argument, as well as

contrast between the notions of scientific evidence and those of legal evidence. Posttrial comments by the presiding judge as well as the two attorneys were particularly informative. Although this volume could not include the trial, a videotape of it is available from the editors.

Acknowledgments

We owe thanks to the many individuals who played an important role in bringing the conference and this volume to fruition. Principal funding for the conference was provided by the Air Force Human Resources Laboratory (AFHRL) in San Antonio, Texas [Contract no. F41689-84-D-0002, Subcontract no. S-744-030-001].

Dr. Malcolm Ree, Senior Scientist, was not only instrumental in securing AFHRL support but also provided useful advice on the design of the program. His enthusiasm and commitment to the project is greatly appreciated. Funding for the preparation of the volume was provided by Educational Testing Service. The support of Gregory Anrig, President, Robert Solomon, Executive Vice-President and Ernest Anastasio, (then) Vice President for Research Management were most appreciated. Stanford von Mayrhauser, General Counsel, also provided financial, intellectual, and moral support for this effort. Linda DeLauro, Conference Coordinator, provided administrative support from the inception of the project through to the completion of this book. Without her extraordinary efforts the entire enterprise would surely have foundered.

H. W.
H. I. B.
Princeton, N.J.
July 1987

Introduction

Testing plays a critical role in all our lives. Beyond its formidable presence in our schools, it is widely employed both by government and industry as an aid to making decisions about people. Thus, testing affects us both directly as individuals who are tested and indirectly through its influence on the welfare of the nation. As befits a practice of such importance, the scrutiny of professionals, national policymakers and the courts has become ever more intense. Their concern centers on whether a particular test properly accomplishes its specified goal and whether it does so in a fair and equitable fashion. This is the core of test validity.

More formally, the *Joint technical standards for educational and psychological testing* (APA, AERA, NCME, 1985) states: "Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness and usefulness of *the specific inferences made from test scores.*[1] Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from test scores."

Not surprisingly the practice of testing antedates concern with its validity. One of the earliest references to testing is described in *Judges* (12:4–6). It seems

[1]Note that it is not the test that has validity, but rather *the inferences* made from the test scores. Thus before we can assess a test's validity, we must know the purposes to which it is to be put.

that the Gileadites developed a short verbal test to uncover the fleeing Ephraimites that were hiding in their midst. The test was one item long. Candidates had to pronounce the word "shibboleth"; Ephraimites apparently pronounced the initial "sh" as "s." Although the consequences of this test were quite severe (the banks of the Jordan were strewn with the bodies of the 42,000 who failed), there is no record of any validity study. Consequently, even though history records all the punished as interlopers, we really do not know how many were Gileadites who spoke with a lisp.[2]

In 1115 B.C., at the beginning of the Chan dynasty, formal testing procedures were instituted for candidates for office. This appears to be the first documented example of a rigorous mental testing program. The Chinese discovered the fundamental tenet of testing; that a relatively small sample of an individual's performance, measured under carefully controlled conditions, could yield an accurate picture of that individual's ability under much broader conditions for a longer period of time. The procedures developed by the Chinese are quite similar to many of the canons of good testing practice used today.[3]

The Chinese system developed over many centuries. By 1370 A.D. it had arrived at a reasonably mature form. A person aspiring to hold public office had to pass three competitive examinations. The first one, given annually and locally, required a day and a night. The passing rate was from 1% to 7%. Those passing were named "budding scholars." The second stage was given every 3 years and was held in each provincial capital. Here all of the "budding scholars" were assembled for three sessions of 3 days and 3 nights each. These examinations were scored with independent readers and the examinees were anonymous. Successful candidates (reported as 1% to 10%) were considered "promoted scholars" and were eligible for the third tier of examinations given the following spring in Peking. In this third set of examinations about 3% passed and became eligible for public office.

Although there is no record of any formal validity studies held as part of the Chinese program, few would doubt that those who succeeded in surmounting the third tier of the rigorous program[4] were highly qualified individuals. Why is it that the results of the biblical test seem obviously questionable, whereas the Chinese program yielded results that no one would question—at least in terms of errors of the first kind?

---

[2]One is reminded of baseball umpire Bill Clem's response when he was asked, "Bill, did you call 'em as you saw 'em? Or did you call 'em as they were?" He answered, "The way I called 'em was the way they were."

[3]The Chinese testing program was used as a model for the British system set up in 1833 to select trainees for the Indian Civil service—the precursor to the British civil service. The success of the British system influenced Senator Charles Sumner of Massachusetts and Representative Thomas Jenckes in developing the examination system they introduced into Congress in 1860. A fascinating description of this is found in Têng (1943) and the interested reader is directed to that source.

[4]There were a number of deaths recorded during the testing process.

The key is the selection ratio. Note that in the Chinese program the overall selection ratio was no greater than about 200 successes per million candidates, and could be as few as 3 per million. Thus the test needed only a very small amount of validity to assure the quality of the final selection.

Modern testing applications usually do not have this extraordinary selection ratio. Instead they are more often like the biblical situation wherein a substantial proportion of candidates will be selected. Consequently, validity becomes a crucial question. The modern shibboleth must be shown to be connected both theoretically and empirically to the criteria of interest.

In the introduction to his classic text on mental testing, Harold Gulliksen pointed out that "During the 1890s several attempts were made to utilize the new methods of measurement of individual differences in order to predict college grades. J. McKeen Cattell and his student, Clark Wissler, tried a large number of psychological tests and correlated them with grades in various subjects at Columbia University; see Cattell (1890), Cattell and Farrand (1896), and Wissler (1901). The correlations between psychological tests and grades were around zero, the highest being .19. A similar attempt by Gilbert (1894) at Yale, produced similarly disappointing results." The validity study provided a clear cautionary note.

Over the last 90 years, as the span of testing has broadened, so too has the notion of test validity. What makes validation research so exciting is not simply that there are always new tests to study. Rather, it is that we are making greater demands of our tests: While asking that they conform to an increasingly stringent set of principles of good practice, we are also asking that they perform more difficult tasks; e.g., predicting performance in complex domains, even when performance levels are not easily quantified. As exciting as the area of test validity is now, it promises to get more so. Rapid advances in cognitive psychology and computer technology will usher in a new generation of tests that may well change not only how we use tests but also how we think about them.

This book addresses some of those changes. It is divided into four sections. The first is concerned with the historical and epistemological bases of validity, as well as current and future issues. Section II, The "Changing Faces of Validity," discusses the potential impact that developments in cognitive psychology and computer technology might have on the makeup and administration of tests. These essays point to changes in the validity criteria as well as methodology of measurement. Section III, "Testing Validity in Specific Subpopulations," examines the questions of validity in the testing of particular groups of individuals. Specifically considered are people with handicaps and members of linguistic minorities. In addition there are two methodological chapters which describe the latest developments in the assessment of differential item functioning. The last section, "Statistical Innovations in Validity Assessment," describes new methods of analysis that allow investigators to look more deeply into the validity question.

We have come a long way in recognizing that we should be seeking not only the best way to separate the Ephraimites from the Gileadites, but also the best way to measure how well we seem to have done so. The contributions to this volume point to new directions for that path.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15*, 373–381.

Cattell, J. M., & Ferrand, L. (1896). Physical and mental measurements of the students at Columbia University. *Psychological Review, 3*, 618–648.

Gilbert, J. A. (1894). Researches on the mental and physical development of school children. *Studies Yale Laboratory, 2*, 40–100.

Gulliksen, H. O. (1950). *A theory of mental tests*. New York: Wiley.

Têng, Ssu-yü. (1943). Chinese influence on the Western examination system. *Harvard Journal of Asiatic Studies, 7*, 267–312.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Monographs, 3*, (16), 1–62.

Contents