

Test Validity

Edited by
Howard Wainer and
Henry I. Braun



TEST VALIDITY

Edited by
Howard Wainer
Henry I. Braun

 **Routledge**
Taylor & Francis Group

TEST VALIDITY

Edited by

Howard Wainer

Henry I. Braun

Educational Testing Service

First Published by

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

Transferred to Digital Printing 2009 by Routledge
270 Madison Ave, New York NY 10016
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Copyright © 1988 by Lawrence Erlbaum Associates, Inc.

All rights reserved. No part of this book may be reproduced in any form, by photostat, microform, retrieval system, or any other means, without the prior written permission of the publisher.

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

Publisher's Note

The publisher has gone to great lengths to ensure the quality of this reprint but points out that some imperfections in the original may be apparent.

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

Contents

List of Contributors	xi
Preface	xiii
Acknowledgments	xv
Introduction	xvii

Section 1 **1**
HISTORICAL AND EPISTEMOLOGICAL BASES OF VALIDITY

<i>Chapter 1</i>	3
Five Perspectives on the Validity Argument <i>Lee J. Cronbach</i>	
Validation as Evaluation Argument, 4	
The Functional Perspective, 5	
The Political Perspective, 6	
The Operationist Perspective, 8	
The Economic Perspective, 9	
The Explanatory Perspective, 12	
The Key to Progress, 14	
References, 15	

<i>Chapter 2</i>	19
Validity: An Evolving Concept <i>William H. Angoff</i>	
Conceptions of Validity, 19	

Construct Validity, 25

Summary, 29

References, 30

Chapter 3

33

The Once and Future Issues of Validity:
Assessing the Meaning and Consequences of Measurement

Samuel Messick

Tension Between Ideal Principles
and Real-World Practices, 34

Evidential and Consequential Bases of Validity, 41

Keynote for Future Validity, 43

Acknowledgments, 44

References, 44

Section II

47

THE CHANGING FACES OF VALIDITY

Chapter 4

49

Mental Models and Mental Tests

James W. Pellegrino

Introduction and Background, 49

Construct Validity, 51

Criterion Validity, 54

Practical Validity, 56

References, 58

Chapter 5

61

GENECES: A Rationale for the Construct Validation
of Theories and Tests of Intelligence

Robert J. Sternberg

The GENECES Framework, 62

Application of one GENECES Framework
to Current Theories and Tests, 67

Conclusions, 74

References, 74

Chapter 6

77

Construct Validity of Computer-Based Tests

Bert F. Green

Conventional and Computer Tests, 77

Examining Construct Validity, 83
 New Tests and New Criteria, 86
 References, 86

Section III 87
 TESTING VALIDITY IN SPECIFIC SUBPOPULATIONS

Chapter 7 89
 Testing Handicapped People—The Validity Issue
Warren W. Willingham
 Introduction, 89
 Admissions Testing and the 504 Regulation, 90
 Special Circumstances, 93
 Defining the Validity Problem, 96
 References, 101

Chapter 8 105
 Validity and Language Skills Assessment:
 Non-English Background Students
Richard P. Durán
 Introduction, 105
 Language Assessment Research, 107
 Insights from Comprehension Research, 112
 Insights from Discourse Analysis, 118
 Summary, 119
 References, 120
 Appendix A, 122

Chapter 9 129
 Differential Item Performance
 and the Mantel–Haenszel Procedure
Paul W. Holland and Dorothy T. Thayer
 Introduction and Notation, 129
 Previous Chi-Square Procedures, 131
 The Mantel–Haenszel Procedure, 133
 The MH Procedure and IRT Models, 137
 Discussion, 142
 References, 143

<i>Chapter 10</i>	147
Use of Item Response Theory in the Study of Group Differences in Trace Lines <i>David Thissen, Lynne Steinberg, Howard Wainer</i> The Three Parameter Logistic Model, 150 Methods of Assessing <i>dif</i> Based on Tests of Equality of Item Parameters, 151 Some Results with Simulated Data, 160 Conclusions, 167 Acknowledgments, 167 References, 168	
<hr/>	
<i>Section IV</i>	171
STATISTICAL INNOVATIONS IN VALIDITY ASSESSMENT	
<i>Chapter 11</i>	173
Validity Generalization and the Future of Criterion-Related Validity <i>Frank L. Schmidt</i> Applications and Impact, 176 Future Trends and Implications, 181 Acknowledgments, 185 References, 185	
<i>Chapter 12</i>	191
The Meta-Analysis of Test Validity Studies: Some New Approaches <i>Larry V. Hedges</i> Notation and Statistical Models, 193 Case I: Random Effects Models with No Corrections for the Effects of Unreliability or Restriction of Range, 196 Case II: Procedures with Corrections for the Effects of Unreliability or Restriction of Range with No Missing Data, 201 Case III: Procedures with Corrections for the Effects of Restriction of Range and Unreliability with Missing Data, 205 Example of Empirical Bayes Methods, 207 Conclusion, 210 Acknowledgment, 210 References, 210	

Chapter 13	213
Some Uses of Structural Equation Modeling in Validity Studies: Extending IRT to External Variables <i>Bengt Muthén</i>	
1. Introduction, 213	
2. Latent Variable Measurement Modeling, 214	
3. A Structural Model, 216	
4. The SIMS Data, 219	
5. Analysis by a Structural Model, 224	
6. Conclusions, 236	
Acknowledgments, 236	
References, 237	
Editors' Introduction to the Discussion	239
Chapter 14	241
Discussion <i>Donald B. Rubin</i>	
On Cronbach and Modeling Self-Selection, 241	
On Messick and the Validity of Inferences, 242	
On the Representativeness of Tests, 242	
On Checking on the Validity of Models and Appropriateness Measurement, 243	
On Holland and Thissen and the Meaning of "Robust," 244	
On Thissen and Detecting Differences in Small Samples Using Large Sample Test Statistics, 245	
On Holland and Extending the <i>dif</i> Model, 246	
On Willingham and Using Low-Dose Technology to Test Handicapped People, 247	
On Bert Green and Computer-Aided Testing, 249	
On Muthén and Structural Modeling, 249	
On Larry Hedges, Test Validity, and Meta-Analysis, 252	
On Schmidt and Validity Generalization, 253	
Summary, 256	
Author Index	257
Subject Index	264

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

xii LIST OF CONTRIBUTORS

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 Industrial 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

¹Note 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 program⁴ 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?

²One 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.”

³The 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.

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