OXFORD

J. Patrick Meyer

Reliability

RELIABILITY

SERIES IN UNDERSTANDING STATISTICS

NATASHA BERETVAS Series Editor-in-Chief
PATRICIA LEAVY Qualitative Editor

Quantitative Statistics

Confirmatory Factor Analysis
Timothy J. Brown

Effect Sizes
Steve Olejnik

Exploratory Factor Analysis
Leandre Fabrigar and
Duane Wegener

Multilevel Modeling
Joop Hox

Structural Equation Modeling
Marilyn Thompson

Tests of Mediation
Keenan Pituch

Measurement

Item Response Theory
Christine DeMars

Reliability
Patrick Meyer

Validity
Catherine Taylor

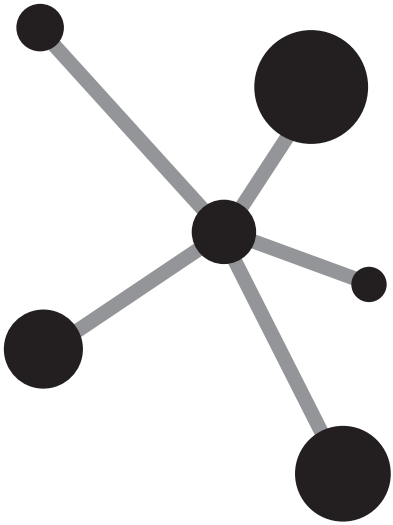
Qualitative

Oral History
Patricia Leavy

The Fundamentals
Johnny Saldaña

J. PATRICK MEYER

RELIABILITY



OXFORD
UNIVERSITY PRESS
2010

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further Oxford University's objective of excellence in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Copyright © 2010 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of Oxford University Press

Library of Congress Cataloging-in-Publication Data

Meyer, J. Patrick.
Reliability / J. Patrick Meyer.
p. cm. — (Series in understanding statistics)
Includes bibliographical references.
ISBN 978-0-19-538036-1
1. Psychometrics. 2. Psychological tests—Evaluation. 3. Educational tests and measurements—Evaluation. 4. Examinations—Scoring. I. Title.
BF39.M487 2010
150.28′7—dc22
2009026618

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

In loving memory of my mother, Dottie Meyer

This page intentionally left blank

ACKNOWLEDGMENTS

I owe many thanks to my wife Christina and son Aidan for their loving support and encouragement. You were very patient with me as I spent many nights and weekends writing this manuscript. I love you both, and I am forever grateful for your support.

I would like to thank the South Carolina Department of Education for the use of the Benchmark Assessment and the Palmetto Achievement Challenge Test data. Special thanks are extended to Teri Siskind, Robin Rivers, Elizabeth Jones, Imelda Go, and Dawn Mazzie. I also thank Joe Saunders for his help with the PACT data files and Christy Schneider for discussing parts of the analysis with me.

I am very grateful for the help and support of colleagues at the University of Virginia. In particular, I thank Billie-Jo Grant for providing feedback on earlier drafts. Your input was invaluable and produced needed improvements to the manuscript. I also thank Sara Rimm-Kaufman and Temple Walkowiak for writing a description of the Responsive Classroom Efficacy Study and allowing me to use the MSCAN data.

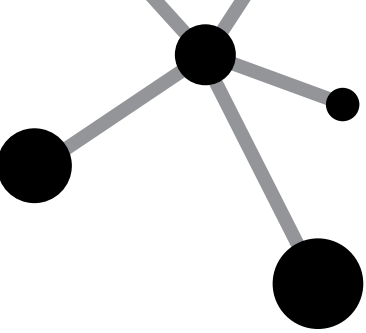
This page intentionally left blank

CONTENTS

This page intentionally left blank

RELIABILITY

This page intentionally left blank



1

INTRODUCTION

Context and Overview

SOCIAL SCIENTISTS frequently measure unobservable characteristics of people such as mathematics achievement or musical aptitude. These unobservable characteristics are also referred to as *constructs* or *latent traits*. To accomplish this task, educational and psychological tests are designed to elicit observable behaviors that are hypothesized to be due to the underlying construct. For example, math achievement manifests in an examinee's ability to select the correct answer to mathematical questions, and a flautist's musical aptitude manifests in the ratings of a music performance task. Points are awarded for certain behaviors, and an examinee's *observed score* is the sum of these points. For example, each item on a 60-item multiple-choice test may be awarded 1 point for a correct response and 0 points for an incorrect response. An examinee's observed score is the sum of the points awarded. In this manner, a score is assigned to an observable behavior that is posited to be due to some underlying construct.

Simply eliciting a certain type of behavior is not sufficient for educational and psychological measurement. Rather, the scores ascribed to these behaviors should exhibit certain properties: the scores should be consistent and lead to the proper interpretation of the construct. The former property is a matter of test score

reliability, whereas the latter concerns test score validation (Kane, 2006). Test score *reliability* refers to the degree of test score consistency over many replications of a test or performance task. It is inversely related to the concept of *measurement error*, which reflects the discrepancy of an examinee's scores over many replications. Reliability and measurement error are the focus of this text. The extent to which test scores lead to proper interpretation of the construct is a matter of test validity and is the subject of another volume in this series.

The Importance of Test Score Reliability

Spearman (1904) recognized that measuring unobservable characteristics, such as mathematics achievement or musical aptitude, is not as deterministic as measuring physical attributes, such as the length of someone's arm or leg. Indeed, he acknowledged that measurement error contributed to random variation among repeated measurements of the same unobservable entity. For example, an examinee may be distracted during one administration of a math test but not during another, causing a fluctuation in test scores. Similarly, a flautist may perform one set of excerpts better than another set, producing slight variations in the ratings of musical aptitude. These random variations are due to measurement error and are undesirable characteristics of scores from a test or performance assessment. Therefore, one task in measurement is to quantify the impact on observed test scores of one or more sources of measurement error. Understanding the impact of measurement error is important because it affects (a) statistics computed from observed scores, (b) decisions made about examinees, and (c) test score inferences.

Spearman (1904, 1910) showed that measurement error attenuates the correlation between two measures, but other statistics are affected as well (see Ree & Carretta, 2006). Test statistics, such as the independent samples t-test, involve observed score variance in their computation, and measurement error increases observed score variance. Consequently, measurement error causes test statistics and effect size to be smaller, confidence intervals to be wider, and statistical power to be lower than they should be (Kopriva & Shaw, 1991). For example, Cohen's d is the effect size for an experimental design suitable for a independent-samples t-test.

An effect size of $d = 0.67$ that is obtained when reliability is 1.0 notably decreases as reliability decreases; decreasing reliability to .8 attenuates the effect size to .60, and decreasing reliability to .5 attenuates effect size to .47. Figure 1.1 demonstrates the impact of this effect on statistical power for an independent-samples t-test. The horizontal line marks the statistical power of 0.8. The curved lines represent power as a function of sample size per group for score reliabilities of 1.0, 0.8, and 0.5. Notice that as reliability decreases, more examinees are needed per group to maintain a power of 0.8. Indeed, a dramatic difference exists between scores that are perfectly, but unrealistically, reliable and scores that are not reliable. Given the influence of reliability on statistics, the conclusions and inferences based on these statistics may be erroneous and misleading if scores are presumed to be perfectly reliable.

Although biased statistics are of concern, some of the greatest consequences of measurement error are found in applications that

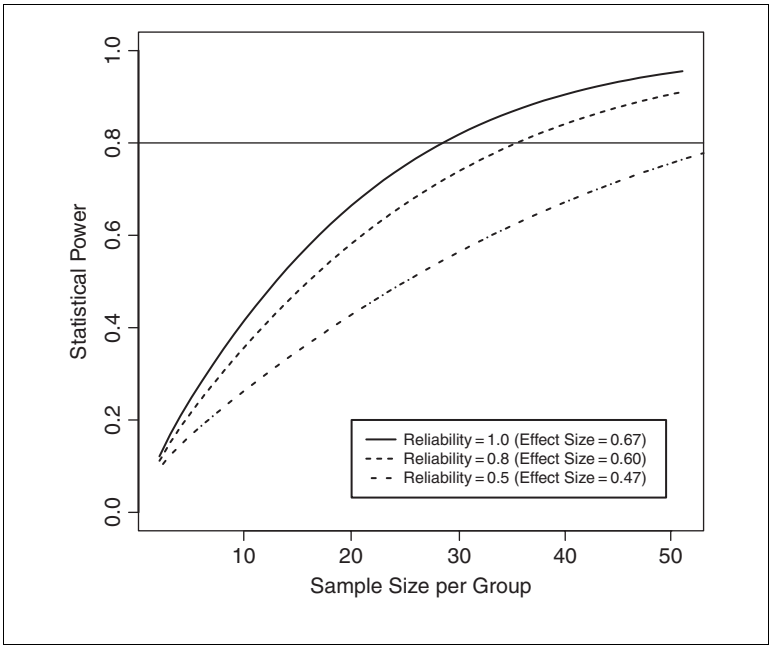


Figure 1.1. The Influence of Reliability on the Statistical Power of a Two-sample t-Test