BASIC PSYCHOLOGICAL MEASUREMENT, RESEARCH DESIGNS, AND STATISTICS WITHOUT MATH

ABOUT THE AUTHOR

Dr. Marty Sapp is a Professor of Educational Psychology at the University of Wisconsin-Milwaukee. He earned his Ed.D. in Educational Psychology at the University of Cincinnati, and his substantive interests are cognitive-behavioral therapies, hypnosis, anxiety disorders, research methods and designs, measurement, and statistics.

Dr. Sapp is a licensed psychologist and Fellow of the American Psychological Association. He is a distinguished Alumnus of the University of Cincinnati College of Education, Criminal Justice, and Human Services, and the recipient of the School of Education Award at the University of Wisconsin-Milwaukee, 2005. In addition, he has published seven books and authored more than fifty journal articles. Moreover, he is an editorial board member of the *Journal of Rational-Emotive and Cognitive-Behavior Therapy,* and he has served on the following editorial boards: *Multiple Voices for Ethnically Diverse Exceptional Learners, Cognitive and Behavioral Practice, Journal of Counseling of Development,* and *Journal of Mental Health Counseling*. Finally, he has served as a reviewer for the following journals: *The Journal of Counseling Psychology, The Counseling Psychologist, Educational Foundations, The International Journal of Clinical and Experimental Hypnosis, Professional Psychology: Research and Practice, Measurement and Evaluation in Counseling and Development.*

BASIC PSYCHOLOGICAL MEASUREMENT, RESEARCH DESIGNS, AND STATISTICS WITHOUT MATH

By

MARTY SAPP

Professor
Department of Educational Psychology
(Counseling Area)
University of Wisconsin-Milwaukee

CHARLES C THOMAS • PUBLISHER, LTD.
Springfield • Illinois • U.S.A.

Published and Distributed Throughout the World by

CHARLES C THOMAS • PUBLISHER, LTD.
2600 South First Street
Springfield, Illinois 62704

This book is protected by copyright. No part of
it may be reproduced in any manner without written
permission from the publisher. All rights reserved.

© 2006 by CHARLES C THOMAS • PUBLISHER, LTD.

ISBN 0-398-07614-6 (hard)
ISBN 0-398-07615-4 (paper)

Library of Congress Catalog Card Number: 2005050661

With THOMAS BOOKS *careful attention is given to all details of manufacturing and design. It is the Publisher's desire to present books that are satisfactory as to their physical qualities and artistic possibilities and appropriate for their particular use.* THOMAS BOOKS *will be true to those laws of quality that assure a good name and good will.*

Printed in the United States of America
SM-R-3

Library of Congress Cataloging-in-Publication Data

Sapp, Marty, 1958–
Basic psychological measurement, research designs, and statistics without math / Marty Sapp.
p. cm.
Includes bibliographical references and index.
ISBN 0-398-07614-6 -- ISBN 0-398-07615-4 (pbk.)
1. Psychometrics. 2. Psychology--Methodology. 3. Psychology--Research. 4. Psychology--Research--Methodology. I. Title.

BF39.S265 2005
150'.72--dc22

2005050661

To my students

PREFACE

Basic Psychological Measurement, Research Designs, and Statistics Without Math is designed for students who are taking an introductory statistics class within the social sciences or a research methods, research design course, or measurement course. Within these quantitative areas, often students are forced to interpret psychological measurement results, research designs, and statistics. Often these three areas are presented as separate areas and students can have strength in statistics but have difficulty with measurement or research design. There are many introductory books on each of these areas, but most books tend to focus on math and calculations or tend to be cookbooks on quantitative methods; nevertheless, few books integrate all three areas.

This text is designed to give students confidence to understand theoretically issues like reliability and validity and through a calculator and statistical packages such as Microsoft Excel, SPSS, SAS and EQS, and students will be shown that they can easily find reliability and validity measures without mathematics. With a few key strokes of a calculator, or a few commands on a statistical package, students can easily calculate reliability point estimates and confidence intervals around reliability estimates. Within the modern era of psychological measurement, mathematical ability is no longer a prerequisite for understanding psychological measurement concepts.

After psychological measurement, research design is the next most important area within quantitative methods. Within this area, conceptually students need to know the definitions of independent and dependent variables and how to design and measure such variables. In addition, students need to understand threats to interval validity, which are factors that can prevent one from concluding that an independent variable caused a change on a dependent variable.

Once students understand internal validity, the next issue is generalization of results or external validity. There are a variety of factors that affect external validity such as social characteristics like the Hawthorne effect, demand characteristics, placebo effects, social desirability, and evaluation apprehension.

After students grasp external validity, common research designs are presented. This section will start with the simplest research design–the one-group case, followed by two-group designs, multiple treatment designs, factorial designs, quasi-experimental designs, and nested designs.

With a firm foundation in psychological measurement and research design, students are first introduced to measures of central tendency and measures of variability. Like the material presented on psychological measurement, again exercises and examples are connected to a calculator and statistical packages. In addition, the general univariate statistics such as t-tests, analysis of variance, simple regression, and analysis of covariance are covered.

After univariate statistics, students are introduced to the univariate and multivariate approach to repeated measures, multiple regression, log linear regression, multilevel regression, multivariate analysis of variance, discriminant analysis, multivariate analysis of covariance, multivariate factorial analysis of variance, step-down analysis, canonical correlation, factor analysis, structural equations analyses, path analysis, and log linear analysis. Finally, a nonmathematical treatment of psychological measurement, research designs, and statistics are presented within one book.

ACKNOWLEDGMENTS

It took several individuals to bring this text into press. First, I would like to thank the students on my research team, especially Mrs. Daun Kihslinger who proofread this entire manuscript. Second, I offer thanks to the University of Wisconsin–Milwaukee School of Education word processing pool for typing this entire manuscript. Moreover, I thank Dr. Walter Farrell at the University of North Carolina at Chapel Hill. In addition, I offer thanks to my University of Cincinnati connections: Dr. Patricia O'Reilly, Dr. Judith Frankel, Dr. Marvin Berlowitz, Dr. Purcell Taylor, the late Dr. David L. Johnson, and Dr. James Stevens. I offer special thanks to Dr. Stevens, because he taught me how to embrace the quantitative aspects of being a psychologist. Moreover, I thank Dr. Festus Obiakor for encouragement and support. Finally, thanks also go to my wife, Mariana Gómez, and also thanks to Mariana for running many of the exercises in this book. I am grateful to the Library Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., to Longman Group, Ltd., London, and to Oliver and Boyd, Ltd,. Edinburgh for permission to reproduce statistical tables from their book, Statistical Tables for Biological, Agricultural, and Medical Research. In closing, comments or discussions concerning this text–both positive and negative–are encouraged. My address is The University of Wisconsin–Milwaukee, Department of Educational Psychology, 2400 E. Hartford Avenue, Milwaukee, Wisconsin 53211. My telephone number is (414) 229–6247, my e-mail address is sapp@uwm.edu, and my fax number is (414) 229–4939.

CONTENTS

BASIC PSYCHOLOGICAL MEASUREMENT, RESEARCH DESIGNS, AND STATISTICS WITHOUT MATH

Chapter 1

RELIABILITY

THEORIES OF RELIABILITY

The material in this chapter was adapted from Sapp (2002). **Classical theory**, also referred to as **weak true score theory**, or **true score theory**, **generalizability theory**, and **item response theory** are the **three** major educational and psychological theories that dominate theories of reliability and validity.

Classical theory is the model often presented within introductory educational and psychological measurement courses. And psychologists have used this theory since the earlier 1900s (Crocker & Algina, 1986; Sapp, 1999). For small-scale projects, classical theory is very simple and useful, and it states that a person's observed score (we call X) is an addition of the person's true score (we call T) and some error term (we call E). E, in this case, is not a mistake, but it is a theoretical construct or concept that takes into account the inconsistency or the lack of perfect ability to measure concepts. Several factors can contribute to error such as the way test items are selected, the way tests are administered, the way tests are scored, and error of measurement due to a theoretical model in which a test has been constructed upon (Allen & Yen, 1979; Mehrens & Lehmann, 1987).

Currently, classical test theory is the dominant viewpoint within psychology and education. Symbolically, the formula for the true score is simply:

$$\mathrm{X} = \mathrm{T} + \mathrm{E} \tag{1.1}$$

X = the person's observed score.
T = the person's true score.
E = error score or the error of measurement.

There are seven assumptions of true-score theory. Classical theory describes how two factors (T and E) affect observed scores. Allen and Yen (1979) reported seven assumptions that are necessary for this model or theory to be tenable. **First**, as we stated earlier, a person's observed score is the addition of two parts a true score and error score or error of measurement. Theoretically, a person's true score is assumed to be fixed, and only the observed score (X) and the error score (E) can vary. For example, if Mark's true IQ test score is 115, and his observed score is 112 and the error of measurement for his score is 3. We can employ Equation 1.1, which is $X = T + E$. If we substitute Mark's true IQ test score $X = 115$ into the equation and substitute his observed score 112, the result is 3. In summary, within this model, Mark's true score and error score are assumed to be an additive, rather than multiplicative or some other mathematical operation. This is referred to as the additivity assumption and it underlies many statistical techniques that are used in measurement such as analysis of variance and factor analysis (Sapp, 1999).

Second, the expected value (population mean) of a person's raw score (X) is the person's true score. Essentially, a person's true score is the mean of the theoretical distributions of raw scores (Xs) that would be obtained from an infinite number of repeated independent testings of the same person with the same test (Allen & Yen, 1979). Independence suggests each testing is unrelated or not affected by another testing; however, in actual practice, an infinite number of testings is not possible; therefore, a person's true score is a theoretical construct.

Third, the correlation among a person's error scores and true scores equals zero; hence, they are statistically uncorrelated. **Fourth**, if the testings are not affected by usual factors, such as the person being fatigued, practice effects, mood, and the person's environment and so on, the errors obtained from two administrations of a test to the same individual equals zero. **Fifth**, a person's error scores on one test and his or her true scores on another test are uncorrelated; nevertheless, this assumption can be violated by personality tests and ability dimensions that affect errors (Allen & Yen, 1979). Finally, assumptions 1 through 5 define error within the classical test score theory; therefore, errors of measurement are random, unsystematic variations of an examinee's observed score from a theoretically expected observed score (Allen & Yen, 1979).

Assumption **six** deals with parallel test and it states that if two tests satisfy assumptions 1 through 5, and for every population of examinees the true scores from test one equals the true scores of test two, and the error variance of test one equals the error variance of test two, then the two tests are called parallel tests. The reader should note that parallel tests are not necessarily perfectly correlated, because, in practice, there is always error variance within test scores (Allen & Yen, 1979; Anastasi & Urbina, 1997; Kaplan &

Saccuzzo, 2001).

Assumption **seven** defines tau equivalent tests. Tests that are **tau equivalent** have true scores that are the same, but the tests differ by a constant. Hence, if two tau equivalent tests satisfy assumptions 1 through 5, and for every population of examinees the true scores of test one equal the true scores of test two plus a constant, the tests are said to be tau equivalent. The reader should note that parallel tests meet stronger restrictions than tau equivalent tests, and parallel tests meet the requirements or assumptions of tau equivalent tests (Embretson & Hershberger, 1999).

There are at least three conclusions that can be drawn from the classical theory. **First**, the observed score variance of a group of examinees equals the examinees' true score variance plus the examinees' error variance. Symbolically, the relationship is as follows:

$$\text{(1.2)} \qquad S_x^2 = S_t^2 + S_e^2$$

where S_x^2 = observed score variance
S_t^2 = true score variance
S_e^2 = error variance

Second, equation 1.2 leads to a theoretical definition of reliability, which states that reliability is the ratio of true score variance divided by observed score variance. If we symbolize reliability as r_{xx}, notice that the subscript "xx" indicates that reliability is a square or squared area.

$$\text{(1.3)} \qquad r_{xx} = \frac{S_t^2}{S_x^2} = \frac{\text{true score variance}}{\text{observed score variance}}$$

Third, if we let S_e denote the standard error of measurement, or the intraindividual variability, S_e is:

$$\text{(1.4)} \qquad S_e = S_x\sqrt{1\text{-}r_{xx}}$$

Anastasi and Urbina (1997) reported that the Weschsler Adult Intelligence Scale-Revised (WAIS-R), a common measure of intelligence, has a standard error of measure (SEM) of 5. How can we use equation 1.4 to arrive at this value? First, items from the WAIS-R have a reliability coefficient of approximately .89 and the items have a standard deviation of 15 (note: see the section on standard scores and the normal curve in Chapter 4). If we substitute into equation 1.4, we get:

$\text{SEM} = S_e = 15\sqrt{1\text{-}.89} = 15\sqrt{.11} = 5$ rounded to a whole number. Because the SEM is analogous to the standard deviation for true scores and the WAIS-R is a standard score, the SEM can be interpreted in terms of the normal curve and confidence intervals. For example, if a client obtained an IQ score of 100 on the WAIS-R, the IQ score of 100 plus and minus $1(S_e)$–the standard error approximates the 68% confidence interval, level, or limit. The

IQ score of 100 minus S_e, or 95 is the lower limit, and the IQ score of 100 plus 5, or 105 is the upper limit. We can expect the client's true IQ score to fall between 95 and 105 68% of the time. Likewise, 100 plus and minus 1.96 times the standard error of measure (5) represents the 95% confidence interval. Where 100 – 9.8 or 90.2 equals the lower limit, and 100 + 9.8 or 109.8 equals the upper limit. Finally, 100 plus and minus 2.58 (5) or 12.9 forms the 99% confidence interval, so the lower limit equals 87.1 and the upper limit equals 112.9. Therefore, with the 95% confidence interval, one can be 95% confident that the client's true IQ score falls within the lower and upper limit (90.2 and 109.8) 95 % of the time. Furthermore, the 99% confidence limit suggests that one is 99% confident that the client's true IQ score falls within the interval width of (87.1 and 112.9) 99% of the time. A point to be noted with the standard error of measurement or with any statistic is that one is never 100% confident or certain; hence, statistics and measurement are based on probability. Sapp (2004b) defined a confidence interval as an interval among an infinitely large set of intervals for a given parameter (population value) in which a certain percentage of the intervals would capture the population parameter. When zero is within the interval, statistical significance is not achieved. The reader who wants more information on confidence intervals can view the following Website:

http://exploringdata.cqu.edu.au/conf_int.htm.

Standard Error Exercises

Suppose a client obtained an IQ score of 110 on the WAIS-R. Establish 68%, 95%, and 99% confidence intervals around the client's score.

Standard Error Answers

68% interval	105–115
95% interval	100.2–119.8
99% interval	97.1–122.9

The major difficulty with reliability is that is can be expressed in at least six ways, and as Thompson (1994, 2003) has noted, tests are not reliable or valid, but it is test scores or items that are reliable and/or valid. **First**, reliability can be defined as the correlation between observed scores on parallel tests. **Second**, reliability always refers to **squared area**. A squared correlation is often referred to as variance accounted for from one variable onto another. For example, a correlation coefficient of .3 square is .09, and this indicates that 9% of the variance is explained; hence, the more variance explained, the greater the relationship or correlation. **Third**, as previously stated, reliability is the ratio of true score variance to observed score variance. The reader should note that as reliability increases, error score variance decreases; therefore, when error variance is small, the observed score is close to the true score. Conversely, large error variance results in poor estimates of true scores and smaller reliability estimates.

Fourth, reliability is the squared correlation between observed scores and true scores. Test scores cannot correlate higher with other variables than with its own true scores; hence, there are limits on reliability, and there is a relationship between reliability and validity. For example, if we used a general anxiety test to predict speech anxiety-a criterion, the correlation between the general anxiety measure and the speech anxiety scores is called a **validity coefficient**. This validity coefficient cannot be larger than the correlation of observed general anxiety test scores with the true general anxiety test scores; therefore, the **valid coefficient cannot be larger than the square root of the reliability coefficient** (Sapp, 1997). Clearly, reliability is a necessary condition for validity. For example, if a test had a reliability coefficient of .90, the validity coefficient cannot be greater than the square root of .90 or .95 rounded to two decimal places. In summary, when the results of a test are said to be consistent, the test scores are reliable, and when test scores measure what they are designed to measure, the test scores are valid. Finally, reliability places upper bounds on validity. And, as Thompson (1994, 2003) noted, it is incorrect to refer to tests as reliable or valid, since it is the test scores or items that are possibly reliable and valid.

Fifth, the reliability coefficient is one minus the squared correlation between observed scores and error scores. **Finally**, the reliability coefficient can be defined as one minus error score variance divided by observed score variance.

Generalizability Theory

Cronbach, Gleser, Nanda, and Rajaratnam (1972) broadened measurement theory by showing that reliability did not have to be restricted to the two-component linear model true scores and error scores (classical theory of reliability). Generalizability (G) theory suggests that several components of error variation can be found, and generalizability theory subsumes and extends classical theory (Brennan, 1998; Shavelson & Webb, 1991; Rajarathun, 1972).

Brennan (1983) developed a program that can simultaneously estimate several sources of main effects (rows or columns from a factorial design) variance and interactions among variance sources, and the program is called generalized analysis of variance (GENOVA).

G theory is concerned with the reliability of generalizing from a client's observed score on a test to his or her average measure that would occur under all possible conditions that are acceptable, and implicit in this assumption is that the client's measured attributes are in a steady state, changes in the client's scores are not the result of maturation, learning, or development, and changes in the client's attributes are the result of multiple sources of error such as occasions, different test forms, different test administrators, and so on. Classical test theory can only estimate one source of reliability at a time. For example, test-retest reliability can estimate variability of scores across time (Shavelson & Webb, 1991). Thus, the strength of G theory is that several sources of error can be estimated within a single analysis. Moreover, G theory allows a clinician to determine how many occasions, test forms, and test administrators are needed to obtain generalizable or reliable scores. Finally, G theory provides a reliability coefficient that is analogous to the classical theory of reliability; therefore, clearly, classical theory can be subsumed under the G theory.

The previous discussion of G theory is often referred to as **G studies**. In contrast, **D studies** use information from G studies to make relative and absolute decisions. **Relative decisions** refer to the rank order of a client in reference to a group. For example, "Mary scored higher than 3/4 of her normed reference group on a standardized science test," could be an example of a relative decision. Mary is compared to other students who took the science test. In contrast, if a client's performance is based on the number of items answered correctly, this could be an example of an absolute decision or a set standard for success. For example, within the United States, many licensure boards for the practice of psychology have established a cut-off score for passing the national psychology exam. An examinee's performance is not based on other psychologists taking the exam, but on the success of the examinee's answering enough items correctly to pass the psychology licensure exam.

In summary, G theory allows a clinician to generalize from a sample of an examinee's behavior to some domain or universe of interest. Clearly, G theory's universe score is analogous to the classical theory's true; however, G theory can estimate several sources of error and several universes for generalization. Finally, G studies can contribute to construct validity by showing the sources of error that are large (Thompson & Cronbach, 1994).

Reliability Coefficients Within Generalizability Theory

The interpretation of reliability coefficients within generalizability theory are similar to that of classical theory in that they represented squared area. For example, the **coefficient G** or **G coefficient** of .5231 represents the