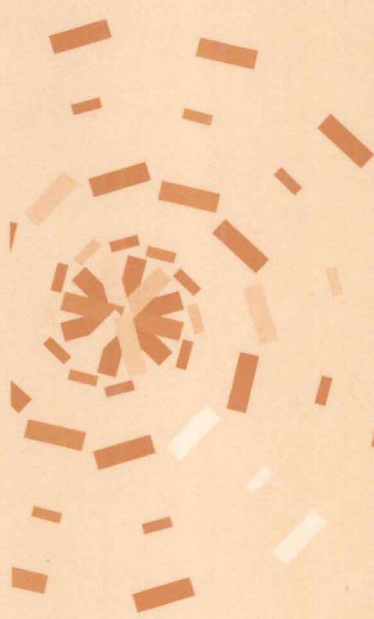
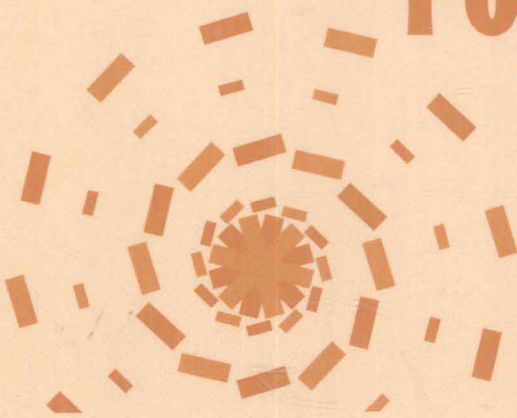


Multicultural Measurement in Older Populations



John H. Skinner
Jeanne A. Teresi
Douglas Holmes
Sidney M. Stahl
Anita L. Stewart
Editors



Springer Publishing Company

Multicultural Measurement in Older Populations

John H. Skinner, EdD
Jeanne A. Teresi, EdD, PhD
Douglas Holmes, PhD
Sidney M. Stahl, PhD
Anita L. Stewart, PhD
Editors



Springer Publishing Company

Copyright © 2002 by Springer Publishing Company, 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 Springer Publishing Company, Inc.

Springer Publishing Company, Inc.
536 Broadway
New York, NY 10012-3955

Acquisitions Editor: Helvi Gold
Production Editor: Sara Yoo
Cover design by Joanne Honigman

01 02 03 04 05 / 5 4 3 2 1

Library of Congress Cataloging-in-Publication-Data

Multicultural measurement in older populations / John H. Skinner, Jeanne A. Teresi,
Douglas Holmes, editors.

p. cm.

Published also as v. 7, no. 1, spring 2001, of the Journal of mental health and aging.
Includes bibliographical references and index.

ISBN 0-8261-2246-9

I. Minority aged—Research—Methodology. 2. Health and race—Research—
Methodology. 3. Gerontology—Statistical methods. I. Skinner, John H., Ed. D.
II. Holmes, Douglas. III. Stahl, Sidney M. IV. Journal of mental health and aging.

HQ1061.M816 2003
305.26'07'2—dc21

2002075791

Printed in the United States of America by Maple-Vail Book Manufacturing Group.

Preface

Measurement in Older Ethnically Diverse Populations

Without culture-fair measures, advancements in health and social services research will be limited. Data that are not reliable and valid can result in biased multivariate results and biased estimates of the prevalence and of the magnitude of risk factors in epidemiological research. In multivariate studies that attempt to identify determinants of health, biases in self-reported measures of the determinants and health outcomes can lead to erroneous conclusions. For example, observed differences between two cultural groups in self-reported health may reflect true differences, or may instead reflect cultural bias in measures used to measure health. Similarly, screening measures used to estimate prevalence of disorder across groups may indicate differing risk factors and etiology; however, one cannot assume without question that such differences reflect true differences in the prevalence of the underlying attribute measured by the screen. If different prevalence ratios are artifacts of the assessment methods, or if different determinants of health outcomes are the result of cultural bias in measures, erroneous conclusions drawn from the results can hinder identification of the disease process, understanding of the effects of environmental stressors on illness, and/or the refinement of interventions to improve outcomes.

The Resource Centers for Minority Aging Research (RCMARS) constitute an effort by the National Institute on Aging (NIA) to redress disparities in health outcomes and delivery. Within the RCMARs, measurement is considered so important that one of the four RCMAR Cores at each site is devoted to this topic. Many of the authors or associate editors of the book are RCMAR investigators or associates. Stahl (see afterword) reviews the history, goals and early achievements of the RCMARs, pointing out that “valid measurement is a prerequisite to accurate assessment of medical needs and outcomes.”

We are fortunate with this book to have as authors or as associate editors, many experts in measurement among ethnically diverse populations. Thus, we are grateful to Bob DeVellis for acting as associate editor of the methods section, to Ron Abeles, Bill Haley and Jennifer Manly for oversight of the section on measurement of acculturation, ethnic identity, socioeconomic status, and social support; and to Ana Abraido-Lanza, Steve Albert, Neal Krause, Jennifer Manly, Harold Neighbors and Al Siu for serving as associate editors for the sections on health, mental health, cognition and religion.

Methodological Issues in Cross-Cultural Assessment

Over the past decade advances have been made in the methodologies used to evaluate measures; moreover, methods long used for detection of item bias in educational testing have become more widely applied to measures of health, mental health and cognition. At the same time, some of the well-known caveats regarding use of standard or traditional measurement statistics have been forgotten. Three chapters in the methods section provide a background for the evaluation of the performance of measures. Teresi and Holmes provide some guidelines and caveats regarding the comparison of summary statistics such as corrected item-total correlations, alphas, sensitivities and specificities. The chapter by Liang discusses cross-cultural invariance in the context of structural equation modeling. He discusses the hierarchy of invariance from conceptual to metric to structural. The lowest level of evidence for factorial invariance is whether the factor structure (number of factors and item loadings on the factors) is equivalent. A second level refers to whether the loadings and measurement error variances and factor means are equivalent. An important issue is whether exact metric invariance is scientifically reasonable or whether configural invariance is the most that can be expected. Configural invariance implies that the pattern of zero and non-zero loadings are the same across groups but are not necessarily equal. Liang argues that metric equivalence remains the standard for valid comparisons.

The chapter by Teresi reviews the different methods for examining differential item functioning (DIF), focusing on studies of the elderly and health and mental health constructs. Both item response theory (IRT) and other methods are reviewed; while IRT-methods are generally preferred theoretically, some other methods can be useful first steps in identifying items that may show DIF.

Acculturation, Ethnic Identity, Socioeconomic Status and Social Support

Several chapters address the issues of acculturation, ethnic identification, socioeconomic status, and social support. The chapter by Skinner reviews acculturation measures developed over the past two decades; of these only 11 provided evidence for both reliability and validity. He underlines the point that scoring high on a measure of adherence to the native culture does not necessarily mean rejection of

the dominant culture. Rather, he suggests that a bicultural approach measuring the degree of adoption of elements of both cultures should be the goal. He also calls for a multidimensional approach to the measurement of acculturation and suggests that more attention be given to developing a measure of the dominant American culture. Skinner cautions against inclusion of items that relate to childhood practices that do not therefore change over time, as such items compromise the ability to examine the process of acculturation over time.

Socioeconomic status (SES) has been linked to health outcomes in numerous studies. Rudkin and Markides describe the ways in which SES has been measured (a) over the years (b) in different epidemiological studies and (c) among different minority groups. They recommend as minimum requirements for adequate measurement of SES among ethnically diverse groups the inclusion of years of schooling, occupational status, income and financial status. However, also recommended for some groups is measurement of literacy, sources of current income and assets. Because of the sensitive nature of questions about income, the authors recommend reducing non-response by using the "bracketing" technique, where those who refuse are asked to indicate either a category of income or whether their income is above or below a certain amount. They recommend oversampling of higher SES minority elderly and of lower SES majority members because of the large SES disparities between ethnic groups that render statistical control inadequate. Rudkin and Markides conclude by emphasizing that ethnicity and minority status are primary determinants of SES differences across groups, and therefore of health outcomes associated with SES differences.

Social support has been linked to positive health and mental health outcomes. However, Mutran, Reed and Sudha point out that differences in definition of social support constructs pose an obstacle to measurement research in this area. They found few studies that examined the properties of social support measures among elderly minority group members, and call for studies examining scales that already exist in terms of performance across different ethnic and racial groups.

Cognitive Function

Two chapters review the findings with respect to possible item and test bias in cognitive functioning measures. Teresi and colleagues examine the few studies using modern psychometric methods to examine the properties of cognitive screening scales and neuropsychological tests. While most scales contain items that show some differential item functioning (DIF), there are scales that contain items that exhibit relatively greater magnitudes of DIF. Because of the growing body of consistent and compelling evidence of DIF with respect to some items, experts are beginning to agree that they are probably culturally biased. Several new scales show promise for use in cross-cultural comparisons.

While item bias is important, the criterion validity of cognitive scales is also important. Several important caveats in evaluating criterion validity are reviewed,

including the fact that the diagnostic criterion can also be biased. Ramírez and colleagues review 18 studies, nearly all published in the 1990's, examining the sensitivity, specificity and predictive values of screening tests against a diagnostic "gold standard." Although summary statistics are variously affected by factors such as base rate, study design and sampling, some guidelines are provided for selecting scales that may perform better in a multicultural environment.

Health, Mental Health and Quality of Life

While self-report and performance measures of functional capacity generally agree, the level of their association is only moderate. However, both are related to poor outcomes, for example, morbidity. The attenuated mutual association may reflect the fact that they are measuring different constructs and contributing different unique information, or that self-report measures are influenced by extraneous factors such as depression, older age, culture or language. As reviewed by Angel and Frisco, some studies show that self-report measures overestimate functional capacity, for example, one large study of Mexican-Americans found that some individuals who reported that they could walk across a small room were unable to actually perform the task. Angel and Frisco conclude that there is little work that examines either cultural invariance in health and ADL measures or how "culture, language, and social class interact with factors that influence responses to survey probes and performance on tests of functional capacity."

Mui, Burnette and Chen point out that sociocultural factors such as "differences in perception, interpretation, valuation, expression, and tolerance of symptoms" may contribute to bias in measures of depression. Somatic symptoms in particular have been found to be problematic, either inflating estimates among older people in general or exhibiting differential item response among different ethnic/racial groups. The authors review two widely-used depression measures that have been applied cross-culturally among the elderly. While one of the measures developed among younger groups for use in epidemiological studies has been subjected to extensive confirmatory factor analyses among ethnically diverse populations, little evidence of cross-cultural criterion validity is available. The other measure, developed for clinical screening for geriatric depression, has been subjected to numerous analyses of the sensitivity, specificity and predictive value across different racial/ethnic groups, but few factor analyses. Cross-cultural criterion validity coefficients were only modest for this latter scale. While depression is an area that has been (relatively) better studied in terms of cross-cultural invariance, more work is still needed.

Health-related quality-of-life constructs are frequently measured as outcomes in studies of chronic disease. However, Nápoles-Springer and Stewart found only 16 studies that examined health-related quality of life measures among elderly ethnically diverse groups, and report that many suffered from one or more deficiencies in measurement adequacy. The authors discuss the fact that there are no clear guidelines

regarding explicit methods for dealing with differences in measurement characteristics across groups when they are observed.

Coping and Religiosity

Coping and health locus of control are two constructs that may be linked with health-promoting behavior and outcomes. Ford, Hill, Butler and Havstad review measures of these two constructs, noting that despite the fact that the John Henry Active Coping Theory grew out of studies of hypertension among African Americans, the scale has not been evaluated systematically among large samples of African Americans. The authors call for more research examining the conceptual equivalence and factorial invariance of the measure.

Increasing attention has been focused on the role of religion as a coping mechanism that can influence health outcomes. Measurement of religious constructs was a focus of a workgroup of experts in measurement of religion sponsored by the Fetzer Institute and the National Institute on Aging. The result was a multidimensional measure of religiousness and spirituality; however, few large epidemiological studies include more than one or two items measuring religion. Chatters, Taylor and Lincoln discuss the lack of research in the mental health literature that deals with religious coping, a potentially important mediating variable related to health outcomes. They conclude that, despite the long-investigated role of religion in the lives of African Americans, few studies include sufficient numbers of African Americans to investigate adequately the measures of religion among this group.

Conclusions

In conclusion, we feel that this book constitutes an important compilation of chapters addressing the state-of-the art in multicultural measurement. The authors (and the associate editors) are well-established researchers who collectively represent the best thinking in this field. What is apparent from their chapters is that in many substantive areas, there is a dearth of cross-cultural measurement research. The most studied areas are cognition and depression. If disparities in health, mental health and service delivery are to be reduced, an important first step is their accurate assessment. Hopefully, this book sets the conceptual stage for more work toward this goal.

John H. Skinner, EdD
Jeanne A. Teresi, EdD, PhD
Douglas Holmes, PhD
Sidney M. Stahl, PhD
Anita L. Stewart, PhD

Contents

Contributors
Preface

vii
ix

Part I: Methodological Issues in Cross-Cultural Assessment

- 1** Some Methodological Guidelines for Cross-Cultural Comparisons 3
Jeanne A. Teresi and Douglas Holmes
- 2** Assessing Cross-Cultural Comparability in Mental Health Among Older Adults 11
Jersey Liang
- 3** Statistical Methods for Examination of Differential Item Functioning (DIF) With Applications to Cross-Cultural Measurement of Functional, Physical and Mental Health 23
Jeanne A. Teresi

Part II: Acculturation, Ethnic Identity, Socioeconomic Status and Social Support

- 4** Acculturation: Measures of Ethnic Accommodation to the Dominant American Culture 37
John H. Skinner
- 5** Measuring the Socioeconomic Status of Elderly People in Health Studies With Special Focus on Minority Elderly 53
Laura Rudkin and Kyriakos S. Markides
- 6** Social Support: Clarifying the Construct With Applications for Minority Populations 69
Elizabeth J. Mutran, Peter S. Reed, and S. Sudha

Part III: Cognitive Function Measures and Cross-Cultural Variation

- 7 Performance of Cognitive Tests Among Different Racial/Ethnic
and Education Groups: Findings of Differential Item Functioning
and Possible Item Bias 85
*Jeanne A. Teresi, Douglas Holmes, Mildred Ramírez,
Barry J. Gurland, and Rafael Lantigua*
- 8 Cognitive Assessment Among Minority Elderly: Possible Test Bias 97
*Mildred Ramírez, Jeanne A. Teresi, Stephanie Silver, Douglas Holmes,
Barry Gurland, and Rafael Lantigua*

Part IV: Measurement of Health, Mental Health and Quality of Life

- 9 Self-Assessments of Health and Functional Capacity Among Older
Adults 129
Ronald J. Angel and Michelle L. Frisco
- 10 Cross-Cultural Assessment of Geriatric Depression: A Review of the
CES-D and GDS 147
Ada C. Mui, Denise Burnette, and Li Mei Chen
- 11 Applying Health Locus of Control and John Henryism Active Coping
Theories to Older African American Adults 179
*Marvella E. Ford, Deanna Hill, Ameera Butler,
and Suzanne Havstad*
- 12 Use of Health-Related Quality of Life Measures in Older and
Ethnically Diverse U.S. Populations 189
Anna M. Nápoles-Springer and Anita L. Stewart

Part V: Religiosity and Ethnicity

- 13 Advances in the Measurement of Religiosity Among Older African
Americans: Implications for Health and Mental Health Researchers 199
Linda M. Chatters, Robert Joseph Taylor, and Karen D. Lincoln
- Afterword A Long-Range Innovative Approach to Reducing Health
Disparities 221
Sidney M. Stahl
- Index* 225

Contributors

Ronald J. Angel, PhD

University of Texas at Austin
Austin, Texas

Denise Burnette, PhD

Columbia University School of Social
Work
New York, New York

Ameera Butler, BS

Resource Center for African American
Aging Research
Detroit, Michigan

Linda M. Chatters, PhD

University of Michigan
Ann Arbor, Michigan

Li Mei Chen, MSW

Columbia University School of Social
Work
New York, New York

Marvella E. Ford, PhD

Resource Center for African American
Aging Research
Detroit, Michigan

Michelle L. Frisco, PhD

University of Texas at Austin
Austin, Texas

Barry J. Gurland, MD

Columbia University Stroud Center
New York, New York

Suzanne Havstad, MA

Resource Center for African American
Aging Research
Detroit, Michigan

Deanna Hill, MPH

Resource Center for African American
Aging Research
Detroit, Michigan

Rafael Lantigua, MD

Columbia University Stroud Center
Department of General Medicine
New York, New York

Jersey Liang, PhD

School of Public Health and Institute
of Gerontology
University of Michigan
Ann Arbor, Michigan

Karen D. Lincoln, MSW, MA

University of Michigan
Ann Arbor, Michigan

Kyriakos S. Markides, PhD

Department of Preventive Medicine &
Community Health
University of Texas Medical Branch
Galveston, Texas

Ada C. Mui, PhD

Columbia University School of Social
Work
New York, New York

Elizabeth J. Mutran, PhD

Center on Minority Aging
University of North Carolina at
Chapel Hill
Chapel Hill, North Carolina

Anna M. Nápoles-Springer, PhD

University of California
Center for Aging in Diverse
Communities
San Francisco, California

Mildred Ramírez, PhD

Hebrew Home for the Aged
Riverdale, New York

Peter S. Reed, MPH

Center on Minority Aging
University of North Carolina at
Chapel Hill
Chapel Hill, North Carolina

Laura Rudkin, PhD

Department of Preventive Medicine &
Community Health
University of Texas Medical Branch
Galveston, Texas

Stephanie Silver, MPH

Hebrew Home for the Aged
Riverdale, New York

S. Sudha, PhD

Center on Minority Aging
University of North Carolina at
Chapel Hill
Chapel Hill, North Carolina

Robert Joseph Taylor, PhD, MSW

University of Michigan
Ann Arbor, Michigan

Part I

**Methodological Issues
in
Cross-Cultural
Assessment**

1

Some Methodological Guidelines for Cross-Cultural Comparisons

Jeanne A. Teresi, Douglas Holmes

The purpose of this chapter is to review and revisit some methodological issues of relevance to cross-cultural research. Companion chapters by Liang and Teresi in this book discuss statistical invariance issues and the role of Confirmatory Factor Analyses and Item Response Theory in such research. Below are several guidelines related to interpretation of reliability and validity coefficients that are well addressed in the measurement and biostatistics literature, but that are frequently ignored in common practice. Some of these guidelines are based on conclusions presented in the psychometric literature over 70 years ago, but have since been forgotten (or were never learned).

1. Avoid omission and translation bias.

Measurement error will result from inadequate representation of a construct. Bias can result from differential definition and operationalization of relevant constructs; poor item structure; idiosyncratic patterns of item selection across administrations and raters; and/or inadequate criteria. The use of different items or wordings and problems in the interpretation of wording when used in translations will lead to error. An example of this is provided by the cognitive test item, “repeating no if’s, and’s, or but’s”: the item has been found to be easier for Latinos interviewed in Spanish than for those interviewed in English. This may be because there is no literal Spanish translation for the item and the alternatives do not measure adequately the intended construct, language (see Teresi, Holmes, Ramirez, Gurland & Lantigua in this book.)

2. Avoid phrases like “the test is reliable”; “the measure has proven reliability for _____(subgroup).”

A measure should be described in terms of the type of reliability estimated, and the context in which the measurement occurred. Reliability must be reexamined for every subsample and, if possible, for different score (disability) groups. Omnibus statistics (one coefficient for the entire group of individuals) do not present a complete picture of the performance of a test because an assumption (usually unrealistic) is that equal errors of measurement are made across all individuals, regardless of their standing on the latent attribute (see Hambleton, Swaminathan, & Rogers, 1991). (Modern psychometric theory, in contrast to classical test theory, allows estimates of precision across the disability continuum.)

Reliability theory was developed under the assumption of normally distributed variables. Moreover, it is assumed that a measure is developed using a random sample of individuals representing the target population. In educational testing, these assumptions are often met. Under such circumstances, and assuming that new samples are similar, one might compare the reliability coefficient obtained in a sample to the normative sample using a well-known test [$1-\alpha_1/1-\alpha_2$], distributed as an F statistic with $N1-1$ and $N2-1$ degrees of freedom (see Feldt, 1969; Feldt and Ankenmann, 1998). However, the case is very different for most measures used in health science research. First, normally distributed variables are not the rule, and the development samples may or may not have been large random samples. Therefore, it is not meaningful to talk in terms of the "reliability" of a measure; a more meaningful description would be "the reliability estimate for this sample using Cronbach's alpha (Cronbach, 1951) was .70. This sample differs from other samples in that the average item prevalence was .15." There is an unfortunate practice among some to report the reliability of a measure as that which is reported in the literature, rather than that which is the case for the current sample. The arguments related to how reliability coefficients should be discussed recently have been revisited (e.g., Sawilowsky, 2000; Thompson & Vacha-Haase, 2000).

3. Avoid comparing Cronbach's alpha across different racial/ethnic groups; never use comparison of means and corrected item-total correlations as evidence for (or against) differential item performance across subgroups. The same caution applies to other measures of reliability as well.

Reliant as they are on average interitem correlations and thus on degree of heterogeneity in a particular sample, reliability coefficients such as Cronbach's alpha, a measure of internal consistency, will vary across populations varying in item prevalences, rendering problematic comparisons across samples drawn from these populations. Inspection of the definition of classical test theory reliability ($1-(\sigma_e^2/\sigma_x^2)$, where σ_e^2 is the error variance, and σ_x^2 is the variance of the measure or item sum, shows that the more homogeneous the sample, the lower the reliability estimate (see Lord and Novick, 1968). Health-related scales are often comprised of dichotomous items or symptoms with low occurrence. Examination of the formula for the KR-20 (equivalent to Cronbach's alpha) applied to binary items shows

explicitly the influence of base-rate or item prevalence on the reliability estimate. $KR-20 = n / n-1 (1 - \sum p_i q_i / \sigma_x^2)$, where p_i is the proportion with the health symptom and q_i equals $(1-p_i)$ and σ_x^2 is as defined above. If item difficulties are equivalent, the numerator reduces to npq , where n is the number of items, p and q are the average p and q values for the item set, and $p_i q_i$ is the item variance. The denominator is comprised of the sum of the item variances added to the sum of the item covariances. As is obvious from these formulas, the smaller the item p values, the smaller the variances and covariances and the smaller the correlation (analogous to Cronbach's alpha). Values for coefficients will vary from sample to sample and from subgroup to subgroup without necessarily reflecting true differences in reliability. The issue of comparing reliabilities becomes especially problematic in the health sciences where items with low prevalences, indicative of greater severity, for example, thoughts of suicide, blurred vision, high systolic blood pressure can be the norm. For example, in a clinic population, prevalences for a suicide item will be greater than in the general population. Similarly, the item prevalences for hypertension-related items will be higher in samples of African Americans than of Latinos, resulting in lower item variances, covariances, corrected item-total correlations and alpha coefficients for the group with lower item prevalences. Comparing these coefficients as evidence of differential item functioning will be meaningless. This is not to say that they should not be computed for each sample. The byproducts of these analyses can help to ensure that there are no coding errors. For example, if negative corrected-item total correlations are observed, this may be the result of coding errors, and as such provide a good diagnostic test to make sure that items are scored correctly. Or varying alphas can indicate that the item prevalences are different between groups.

An example of such differences is provided by Teresi and Holmes (1994) using two samples of older community-resident elderly of different age cohorts. The observed alphas for several scales in the first community sample were .87, .84, .95, while in the other sample, the comparable alphas were .74, .51, .84. Coefficient alpha is a function of the number of items and of the average interitem correlation; given a constant number of items, differences in item base rates can result in different alphas. In the first study the estimated prevalence ratios for depression, cognitive impairment and activity limitation were .10, .05, .20, respectively. In the other study the prevalences for these scales were .03, .02, .07. The observed estimates of the average interitem correlations for the first samples were .19, .35, .35. Given the lower item prevalences, the average interitem correlation in the second study is about one half of the original estimate, resulting in lower possible (maximum) observed alphas. Carmine and Zeller (1979) provide a table showing that, given an average interitem correlation of .2, a four item scale will have an alpha of .50, a 10-item scale a value of .71. Similarly, if the average interitem correlation is .4, the 4-item scale will have an alpha of .73, the 10-item scale .87.

While in educational testing items with floor or ceiling effects in a normative sample will be considered poor items, offering little information about a construct,