ASSESSMENT

ASSESSMENT

Volume 9 • Number 1 • March 2002

Assessment publishes articles derived from psychometric research, clinical comparisons, or theoretical formulations and literature reviews that fall within the broad domain of clinical and applied assessment. The emphasis of this journal is on publication of information of direct relevance to the use of assessment measures, including the practical applications of measurement methods, test development and interpretation practices, and advances in the description and prediction of human behavior. The scope of the journal extends from the evaluation of individuals and groups in clinical, counseling, health, forensic, organizational, industrial, and educational settings to the assessment of treatment efficacy, program evaluation, job performance or the study of behavior outcomes. Papers considered appropriate for publication encompass investigations related to the development, validation, and evaluation of new or established assessment instruments; the application of computer-based technologies to assessment issues; measurement of human characteristics, resources, aptitudes, abilities, or traits; and description of abnormal behavior and diagnosis of mental disorders. Papers that focus on cognitive and neuropsychological assessment, the interface of cognitive, personality, and psychopathology domains, and the definition of traits and behaviors that comprise aspects of normal behavior and the realm of psychopathology are within journal purview.

Subscriptions: Annual subscription rates for institutions and individuals are based on the current frequency of publication. Prices quoted are in U.S. dollars and are subject to change without notice. Canadian subscribers add 7% GST (and HST as appropriate). Outside U.S. subscription rates include shipping via air-speeded delivery. Institutions: $249 (within the U.S.) / $265 (outside the U.S.) / single issue: $73 (worldwide). Individuals: $65 (within the U.S.) / $81 (outside the U.S.) / single issue: $26 (worldwide). Orders with ship-to addresses in the United Kingdom, Europe, the Middle East, and Africa should be sent to the London address (below). Write to the London office for sterling prices. Orders with ship-to addresses in India and South Asia should be sent to the New Delhi address (below). Noninstitutional orders must be paid by personal check, Discover, VISA, MasterCard, or American Express.

This journal is abstracted or indexed in **Abstract Journal of the Educational Resources Information Center, Current Contents: Social & Behavioral Sciences, Current Index to Journals in Education (CIJE), Index Medicus, MEDLINE, PsychINFO, Psychological Abstracts,** and **Social Sciences Citation Index**.

Assessment (ISSN 1073-1911) is published quarterly—March, June, September, and December—by Sage Publications, 2455 Teller Road, Thousand Oaks, CA 91320. Telephone: (800) 818-SAGE (7243) and (805) 499-9774; fax/order line: (805) 375-1700; e-mail: order@sagepub.com; http://www.sagepub.com. Copyright © 2002 by Sage Publications. All rights reserved. No portion of the contents may be reproduced in any form without written permission of the publisher.

Back Issues: Information about availability and prices of back issues may be obtained from the publisher's order department (address below). Single-issue orders for 5 or more copies are eligible for discounts. Contact the order department for details.

Inquiries: All subscription inquiries, orders, and renewals with ship-to addresses in North America, South America, Australia, China, Indonesia, Japan, Korea, New Zealand, and the Philippines must be addressed to Sage Publications, 2455 Teller Road, Thousand Oaks, California 91320, U.S.A.; telephone: (800) 818-SAGE(7243) and (805) 499-9774; fax: (805) 375-1700; e-mail: info@sagepub.com. All subscription inquiries, orders, and renewals with ship-to addresses in the United Kingdom, Europe, the Middle East, and Africa must be addressed to Sage Publications Ltd, 6 Bonhill Street, London EC2A 4PU, England; telephone: +44 (0)20 7374-0645; fax: +44 (0)207374-8741. All subscription inquiries, orders, and renewals with ship-to addresses in India and South Asia must be addressed to Sage Publications Private Ltd, P.O. Box 4215, New Delhi 110 048 India; telephone:(91-11) 641-9884; fax: (91-11) 647-2426. Address all permissions requests to the Thousand Oaks office.

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Sage Publications for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of 50¢ per copy, plus 10¢ per copy page, is paid directly to CCC, 21 Congress St., Salem, MA 01970. 1073-1911/2002 $.50 + .10.

Advertising: Current rates and specifications may be obtained by writing to the Advertising Manager at the Thousand Oaks office (address above).

Claims: Claims for undelivered copies must be made no later than six months following month of publication. The publisher will supply missing copies when losses have been sustained in transit and when the reserve stockwill permit.

Change of Address: Six weeks' advance notice must be given when notifying of change of address. Please send old address label along with the new address to ensure proper identification. Please specify name of journal. POSTMASTER: Send address changes to the Assessment, c/o 2455 Teller Road, Thousand Oaks, CA 91320.

Printed on acid-free paper

A Letter to Our Readers

With the distribution of Volume 8, Issue 4, in 2001, *Assessment* has completed 8 years of publishing articles covering a wide variety of topics including objective and projective methods, personality assessment, as well as intellectual and neuropsychological assessment, and advances in computer administration and interpretation techniques. The journal has become an important component of the assessment literature and is indexed in all of the central services including MEDLINE, Psychological Abstracts, and Social Science Citation Index. The journal has been able to establish and maintain high standards for acceptance of articles, comparable to those of long-established journals operational for many decades. This level of success has been a direct result of several factors, most certainly including the outstanding efforts of the editorial board and associate editors in providing a high-quality review process for manuscripts submitted to the journal, and also the vision and support of Bob Smith, CEO of Psychological Assessment Resources (PAR). Bob began working with me nearly a decade ago to establish a high-quality journal to provide an important service to the assessment field. As a result of his commitment and support, the journal has matured from a new and relatively unknown publication to a well-established journal that has made numerous important contributions to the scientific literature in the assessment area.

This letter is to announce a "changing of the guard" in terms of the publication home for *Assessment*. After nearly a decade, PAR and Bob Smith will step aside from their founding role as owner and publisher of *Assessment*, and Sage Publications will become the new home and publisher of the journal, starting with our first issue in 2002. Bob Smith has my sincerest thanks and gratitude for his crucial efforts in initiating and funding the journal and guiding it through its initial development to its current status. He has been a constant source of encouragement for the journal, providing all of the resources necessary for the journal's success while remaining completely nonintrusive in terms of editorial decisions on manuscript publication. No journal editor could have asked for a better owner/publisher, and it has been a rare privilege to work with Bob to help create *Assessment*.

The numerous tasks involved in the transition of the journal to new ownership are greatly eased by our ability to work with Sage Publications, a leading international publisher of books, journals, and electronic media. Sage has an international reputation for its publication of high-quality scholarship and for its ability to work closely with authors and editors to produce outstanding journals across a wide variety of fields. The publication experience, expertise, and commitment to quality demonstrated by Sage over the past 35 years will be instrumental in moving *Assessment* to the next level of the development of the journal. Although there will undoubtedly be a few challenges along the way, I am confident that this transition will essentially be "seamless" to our readership and that our review process will continue to provide the highest quality feedback to the authors submitting their work to *Assessment*.

Once again, I would like to express my deepest appreciation to Bob Smith for his crucial role in the founding of this journal, and I look forward to working with Sage in the publication of *Assessment*, beginning with this issue.

Sincerely,

—Robert P. Archer, Ph.D.
Editor
January 4, 2002

A Structure-Based Approach to Psychological Assessment

Matching Measurement Models to Latent Structure

John Ruscio
Elizabethtown College

Ayelet Meron Ruscio
The Pennsylvania State University

The present article sets forth the argument that psychological assessment should be based on a construct's latent structure. The authors differentiate dimensional (continuous) and taxonic (categorical) structures at the latent and manifest levels and describe the advantages of matching the assessment approach to the latent structure of a construct. A proper match will decrease measurement error, increase statistical power, clarify statistical relationships, and facilitate the location of an efficient cutting score when applicable. Thus, individuals will be placed along a continuum or assigned to classes more accurately. The authors briefly review the methods by which latent structure can be determined and outline a structure-based approach to assessment that builds on dimensional scaling models, such as item response theory, while incorporating classification methods as appropriate. Finally, the authors empirically demonstrate the utility of their approach and discuss its compatibility with traditional assessment methods and with computerized adaptive testing.

Keywords: taxometrics, latent structure, measurement, classification, scaling

The latent structure of a psychological construct may be either taxonic (categorical, discrete, qualitative, latent class), dimensional (continuous, quantitative, latent factor, latent trait), or some combination of both. Although the importance of latent structure for measurement has been noted in the assessment literature (e.g., Meehl, 1992; Smith & McCarthy, 1995), there has not yet been a systematic effort to present the full range of possible latent structures, discuss how latent structure can inform the choice of measurement models, or articulate the implications of this choice for assessment. In the present article, we assert that the match—or mismatch—between the latent structure of a construct and the model by which that construct is measured affects the accuracy with which individuals are placed along a continuum or assigned to classes. We explore the consequences of this structure-model match for measurement error, statistical power, the search for an efficient cutting score, and statistical relations among constructs.

In what follows, we develop and illustrate the value of a comprehensive, structure-based approach to assessment by highlighting the critical role of latent structure in measurement. First, we explore differences between the latent and manifest levels of analysis and describe the possible types of latent structures. Next, we describe why it is important to match one's measurement approach to the latent

Portions of this article were presented at the annual meeting of the American Psychological Association, Washington, D.C., August 2000. This work was supported in part by National Research Service Award MH12675-01 to Ayelet Meron Ruscio from the National Institute of Mental Health. We would like to express our thanks to Paul E. Meehl and Leslie J. Yonce for access to the Hathaway Data Bank. Correspondence concerning this article should be addressed to John Ruscio, Department of Psychology, Elizabethtown College, One Alpha Drive, Elizabethtown, PA 17022; e-mail: rusciojp@etown.edu.

Assessment, Volume 9, No. 1, March 2002 4-16

structure of the construct under investigation. We then briefly review various methods for empirically evaluating latent structure and suggest a generalized strategy for applying these methods. Finally, we outline a structure-based approach to assessment that builds on knowledge of latent structure by incorporating dimensional scaling and categorical classification methods as appropriate.

DISTINGUISHING MANIFEST AND LATENT STRUCTURE

The critical distinction between latent and manifest levels of analysis is seldom discussed in the assessment literature. Latent structure refers to the fundamental nature of a construct, the underlying structure that exists regardless of how one might choose to conceptualize or measure it. Manifest structure, in contrast, refers to characteristics associated with observable indicators of a construct, the surface structure that depends—among other things—on how one chooses to conceptualize and assess the construct. For a given construct, latent and manifest structure can differ (Grayson, 1987; Murphy, 1964). Meehl's (1962, 1990) theory of schizophrenia provides one example of how a latent category can give rise to manifest continua. The theory posits the existence of a single dominant gene that causes central nervous system deficits specific to schizophrenia. Those who inherit this gene develop schizotypy, a condition characterized by psychological and behavioral features such as cognitive slippage, social aversiveness, anhedonia, and ambivalence. Though signs such as these are distributed continuously at the manifest level, they have been found to correspond to a class of schizotypes at the latent level (Golden & Meehl, 1979; Korfine & Lenzenweger, 1995; Lenzenweger, 1999; Lenzenweger & Korfine, 1992; Tyrka et al., 1995).

In contrast, any construct that is continuous at the latent level can be made to appear categorical at the manifest level. One way in which this is often done is by applying a median split to a distribution of scores for the purpose of analytic convenience. For example, the continuous scores yielded by Rotter's (1966) Internal-External Locus of Control Scale are typically divided at the median to create groups with an internal or external locus of control, despite the possibility that this construct is continuous at the latent level. Another approach to categorization is demonstrated by the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1994), which depicts all psychological disorders as latent taxa. The *DSM-IV* organizes mental disorders within diagnostic categories, each associated with a specific set of criteria that determine whether an individual is, or is not, disordered. Although these diagnostic categories may accurately reflect the taxonic structure of some psychological disorders, there is evidence that at least some of these categories mask underlying continua (A. M. Ruscio, Ruscio, & Keane, 2001; J. Ruscio & Ruscio, 2000). Thus, both statistical and conceptual categorization at the manifest level may correspond to continua at the latent level.

As these examples illustrate, a given manifest structure need not match the underlying latent structure of a construct—latent categories may give rise to an observed continuum, and a latent continuum may give rise to categorical measurements. We focus on the latent level of analysis in the present article due to the importance of understanding the true nature of a construct, regardless of how people have chosen to measure it.

For clarity and consistency with the literature on latent structure, we use the term *taxonic* to refer to a construct in which individuals or objects are separated into nonarbitrary classes, or taxa, at the latent level. That is, one or more qualitative boundaries "carve nature at its joints": Objects either do or do not belong to these taxa regardless of an observer's beliefs or preferences. By contrast, we use the term *dimensional* to refer to a construct along which individuals or objects differ only quantitatively, such that any classes that might be formed are arbitrary. Incontrovertible examples of latent taxa and dimensions exist in many sciences, though comparatively little research has explored the latent structure of psychological variables. Clear-cut examples of taxonic constructs include biological species, chemical elements, and subatomic particles, with representative taxa being the blue-ring octopus, magnesium, and the proton, respectively. Definitive examples of dimensional constructs include body mass, barometric pressure, and temperature, which are scaled along the continua of kilograms, millimeters of mercury, and degrees centigrade, respectively. "Obese" people, "high-pressure" weather, and "hot" objects are not naturally occurring categories but rather distinctions superimposed on dimensions for pragmatic purposes. Within psychology, preliminary evidence suggests that constructs such as psychopathy (Harris, Rice, & Quinsey, 1994), pathological dissociation (Waller, Putnam, & Carlson, 1996; Waller & Ross, 1997), and Type A personality (Strube, 1989) may be taxonic, whereas constructs such as adult attachment (Fraley & Waller, 1998), depression (A. M. Ruscio & Ruscio, 2001; J. Ruscio & Ruscio, 2000), and worry (A. M. Ruscio, Borkovec, & Ruscio, 2001) may be dimensional. Although this is only a partial listing of psychological constructs whose latent structure has been investigated, the overwhelming majority of variables of interest to psychologists have not been studied. Thus, there is an acute need for research that empirically evaluates latent structure using powerful analytic techniques designed expressly for this purpose.

Although our definitions of taxa and dimensions are standard, it is misleading to imply that a construct must be either taxonic or dimensional because structural combinations are possible. This point has been alluded to elsewhere (e.g., Waller & Meehl, 1998) but not elaborated in a rigorous way. In addition to the relatively simple latent structures of pure taxa (latent classes containing individuals whose manifest scores differ only due to measurement error) and pure dimensions (latent continua along which there are no qualitative boundaries), many hybrid latent structures are theoretically possible. This occurs in cases where a latent class or dimension can itself be broken down into additional latent classes and/or dimensions. Thus, the assessment of latent structure can be conceived as a hierarchical, iterative process in which constituent taxa or dimensions are sought until no further subdivisions of any kind are possible. Ultimately, stopping points will be reached whenever a pure dimension (one that is indivisible into constituents) or a pure taxon (one with no reliable residual variation) is uncovered.

To illustrate some of the possibilities, several hypothetical latent structures for the construct of depression are depicted in Figure 1. Panel A depicts depression as having no qualitative boundaries whatsoever, a pure dimension. Panel B shows depression divided into two latent classes, with no reliable residual variation within either latent class. Panel C shows a simple combination: There is one pure latent class, whereas the other class contains reliable residual variation and is thus a dimension. Panel D depicts a more complex combination: One latent class consists of three subtypes, and the other is a dimension. There are, of course, far more possibilities than the small sampling presented here.

Although a recent literature review (Flett, Vrendenburg, & Krames, 1997) and empirical investigations (A. M. Ruscio & Ruscio, 2001; J. Ruscio & Ruscio, 2000) suggest that depression may best be represented by a single dimension, the nature and complexity of most psychological constructs remain an unexplored empirical question. Waller et al. (1996) noted that as psychologists, "we too often presuppose that our data are unquestionably scaleable along latent dimensions or latent traits (factors or continua)" (p. 317). Dahlstrom (1995), Gangestad and Snyder (1985), and Meehl (1992, 1995) also discussed strong biases against latent taxa. Although structure is often presumed to match the manifest measurement scale employed or the presupposition of the researchers, the determination of a construct's true latent structure poses an empirical question that can be addressed using appropriate methods.

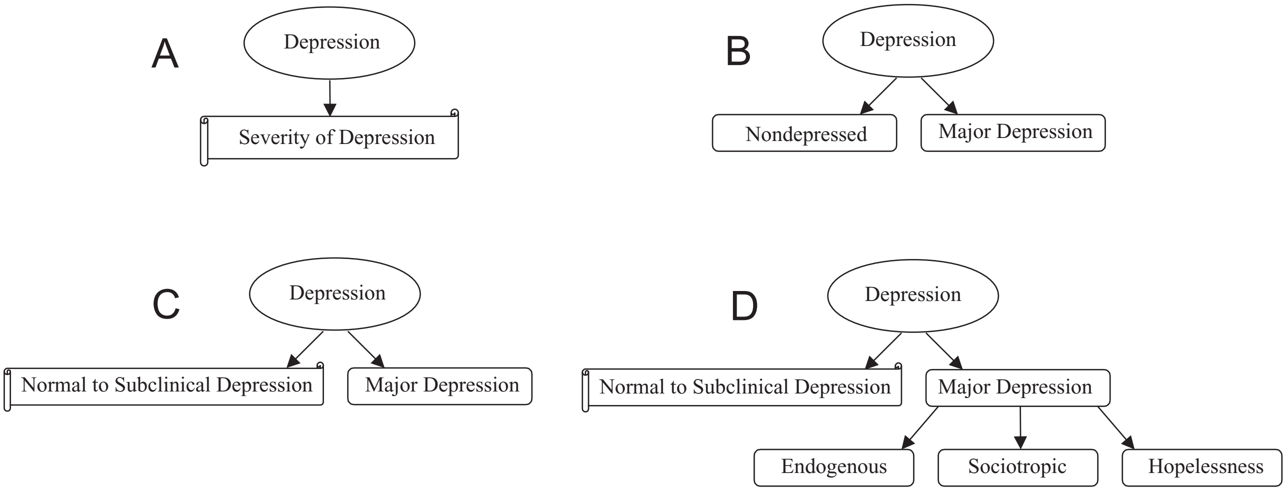
LATENT STRUCTURE AND PSYCHOLOGICAL ASSESSMENT

Understanding latent structure has significant implications for psychological assessment. A measurement model based on dimensional scaling will best locate an individual's position along a continuum, whereas a measurement model based on classification into taxa will best assign individuals to groups. As Meehl (1992) has noted, these disparate measurement goals lead to considerably different assessment guidelines and approaches, making an appropriate match between latent structure and measurement model particularly important.

When assessing a latent dimension, the goal of measurement is to most precisely determine the value of each individual's true score (in classical test theory) (Guilford, 1954; Gulliksen, 1950) or latent trait (in item response theory [IRT]) (Embretson, 1996; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). In this context, any model that classifies individuals into groups is inappropriate. Dimensional measurement of a dimensional construct results in maximal measurement precision and statistical power, whereas spurious classification may have devastating consequences. Cohen (1983) has shown that when computing the statistical association between two continuous variables, dichotomizing one of them throws away 36% of the systematic variance, whereas dichotomizing both of them throws away nearly 60% of the systematic variance. In this way, research employing categorical diagnoses to study the comorbidity among psychological disorders may dramatically underestimate the co-occurrence of conditions that exist along a latent continuum. Indeed, preliminary evidence of dimensional structure for several depressive and anxiety disorders (A. M. Ruscio, Ruscio, & Keana, in press; J. Ruscio & Ruscio, 2000) suggests that this weakening of statistical power may systematically distort our understanding of the controversial relationship between these constructs (e.g., Clark & Watson, 1991; Foa & Foa, 1982; Maser & Cloninger, 1990). The rise in Type II errors associated with decreased power led Fraley and Waller (1998) to argue that spurious classification can cripple a field of research in the long run.

Another deleterious effect of spurious classification is that it may alter—not just weaken—statistical relations and inferred theoretical links between constructs. For example, the common practice of measuring adult attachment styles by the popular three-group scheme (secure, insecure-avoidant, and anxious-ambivalent) (Ainsworth, Blehar, Waters, & Wall, 1978), rather than by the two dimensions suggested by an examination of latent structure (anxiety and avoidance) (Fraley & Waller, 1998), may account for the alleged temporal instability of attachment

FIGURE 1
Four Hypothetical Ways in Which the Construct of Depression Might Be Broken Down Into Its Latent Structure (with scrolls representing dimensions and boxes representing taxa)



NOTE: Panel A shows depression as one dimension. Panel B shows depression as two latent classes. Panel C shows depression as one dimension plus a latent class. Panel D shows depression as one dimension plus a latent class with three subtypes.

styles (Baldwin & Fehr, 1995). That is, the standard error of difference scores will cause some individuals to be classified differently over time. Moreover, forcing latent dimensions into taxa will greatly increase error by throwing away meaningful variation in scores. Because a majority of published studies on adult attachment have superimposed a typological measurement scheme on the data, this literature may be in need of both reanalysis and reconceptualization (Fraley & Waller, 1998). Similarly, arbitrary categorization of data guided by communicative convenience or preference for a particular analytic strategy (e.g., ANOVA rather than regression) has likely weakened the strength and even distorted the form of statistical relations in many other domains of psychological research.

Whereas the classification of latent dimensions into groups results in a considerable loss of information, proper classification of latent taxa has been suggested (Meehl, 1992) and twice demonstrated (Gangestad & Snyder, 1985; Strube, 1989) to yield stronger relationships between taxa and other variables than measurement using dimensional scaling. This is because, for pure taxa, any variance in observed scores around the true scores of the taxa must be measurement error. Thus, applying a dimensional measurement model can increase error when taxa exist. The general conditions under which categorical classifications outpredict dimensional scales in the presence of latent taxa remain an important open question (see Grove, 1991b).

Finally, there is an additional advantage to classification models that is seldom addressed in the assessment literature: They assist users in locating an efficient cutting score for classifying cases into taxa. Even in the presence of latent taxa, dimensional scaling models typically yield unimodal distributions of manifest scores. Without any natural breaks in such a distribution, it is quite challenging to determine an appropriate cutting score for separating individuals into groups, and the efficiency of classification drops off rapidly with suboptimal choices. Classification models, on the other hand, yield a strongly bimodal distribution of manifest scores for latent taxa. With such a distribution, one can clearly identify an efficient cutting score by locating the low point toward the center of the distribution. Moreover, with so few cases in nearby regions of the distribution, the efficiency of classification is highly robust to the selection of suboptimal cutting scores.

In sum, there are a number of practical advantages associated with a structure-based approach to psychological assessment. Therefore, we turn now to the first step of such an approach: determining the latent structure of the construct of interest.

METHOD FOR DETERMINING LATENT STRUCTURE

Because measurement models presume a latent structure that is not directly observable at the manifest level, it is critical that latent structure be evaluated using methods expressly designed for this purpose. Among the presently available techniques, we believe that Meehl's (1995, 1999) taxometric method is the most promising. Below, we briefly outline the logic of this taxometric method, provide an overview of several procedures that constitute the method, and compare the method to the available alternative approaches.

Logic of the Taxometric Method

Meehl (1973, 1995, 1999) and his colleagues (Golden & Meehl, 1979; Grove & Meehl, 1993; Meehl & Golden, 1982; Meehl & Yonce, 1994, 1996; Waller & Meehl, 1998) have pioneered the development of a family of taxometric procedures based on the principles of coherent cut kinetics. Most procedures within the method search for orderly statistical relations between one or more variables along sliding intervals, or cuts, of another. Each procedure uses manifest indicators to search for a qualitative boundary between two latent taxa, traditionally referred to as the "taxon" and "complement." Meehl's taxometric method relies on the convergence of evidence obtained from multiple, quasi-independent analytic procedures—rather than on traditional null hypothesis significance tests—to provide clues to latent structure. Each procedure serves as a consistency check for the results provided by the others, with confidence in a structural solution increasing as each additional consistency test is passed.

Although procedures in the taxometric method directly test only a two-group latent class model, investigators can resolve more complex latent structures by combining these procedures with psychometric analyses (e.g., evaluating unidimensionality, homogeneity, and internal consistency) and applying them in an iterative fashion. For example, consider the latent structure depicted in Panel D of Figure 1. With the proper selection of indicator variables, an initial taxometric analysis would indicate that there was a qualitative boundary between the taxon (major depression) and the complement (normal to subclinical depression). Subsequent taxometric analyses within the complement class would fail to reveal any additional taxa; psychometric analyses would reveal the reliable residual variation of a single dimension. However, a series of subsequent taxometric analyses within the taxon, using new sets of indicators specific to the conjectured subtypes, would uncover three subtypes of major depression, and psychometrics would show no reliable residual variation within them. Taken together, these steps represent an idealization of a careful program of systematic research essential for the comprehensive understanding and appropriate assessment of any psychological construct.

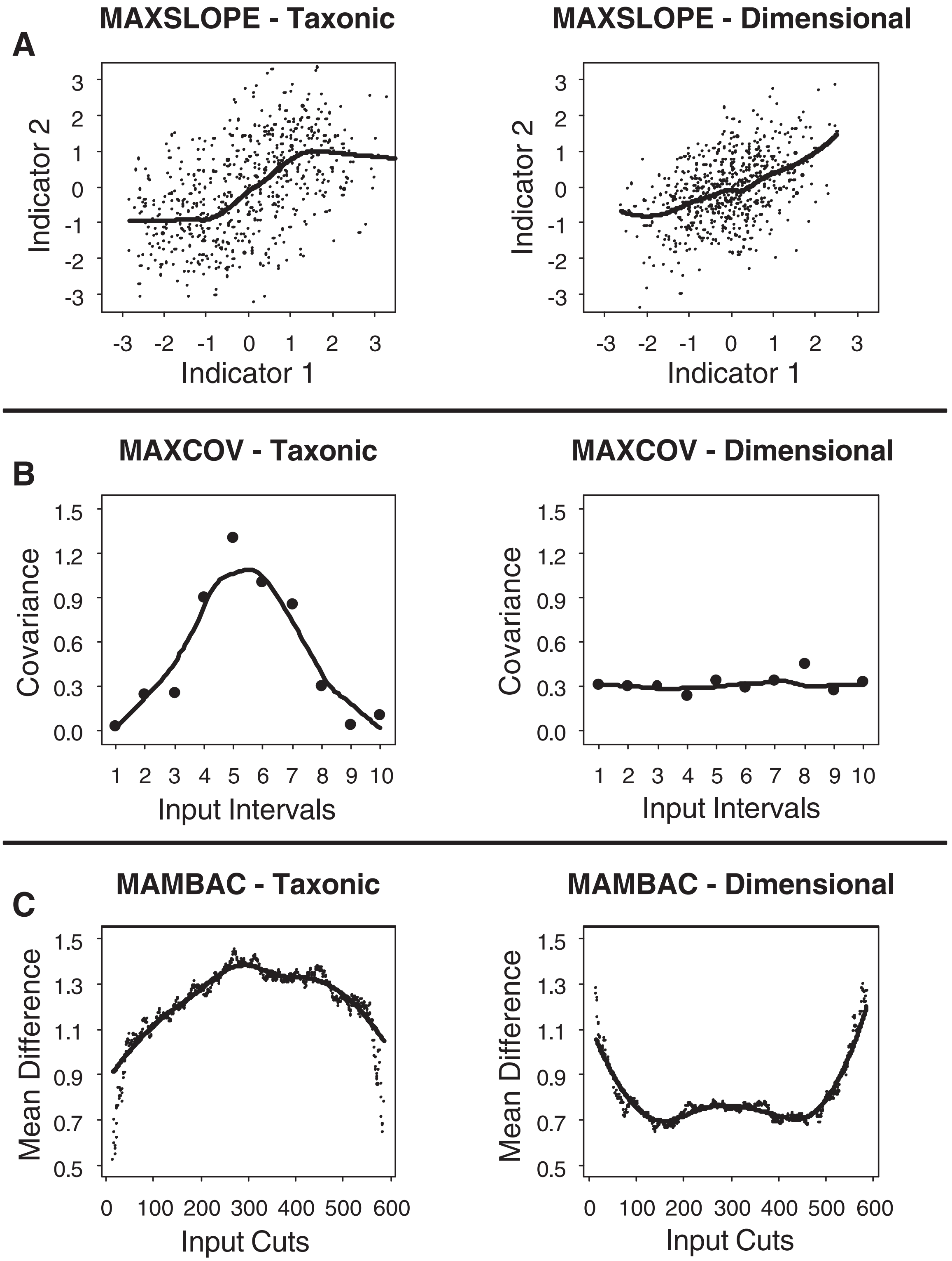
Procedures in the Taxometric Method

Although many taxometric procedures have been developed and validated, only a few of the conceptually simplest techniques will be presented here in the interest of conserving space. Whereas each of the procedures described below can be used alone to provide a structural solution, they are more appropriately used in tandem, with each procedure serving as a consistency test that checks the conclusions of the others. We focus on a conceptual presentation of the method so that it can be compared to better known alternatives, and we illustrate each taxometric procedure using three continuously distributed indicator variables—competitiveness, time urgency, and hostility—that have been suggested to distinguish individuals with Type A personality from those with Type B personality (Friedman & Rosenman, 1974), which research suggests is taxonic (Strube, 1989). Readers interested in more detailed treatments of Meehl's taxometrics, including descriptions of powerful multivariate procedures in the method, should consult Meehl and Golden (1982), J. Ruscio and Ruscio (2001), Waller and Meehl (1998), and the references cited below.

Maximum slope (MAXSLOPE). MAXSLOPE (Grove & Meehl, 1993) examines the slope of a local regression across a scatterplot of two indicators of the conjectured latent taxa. For example, suppose that two manifest indicators, such as time urgency and hostility, were plotted for a sample containing 50% Type A and 50% Type B individuals. Because these indicators are unrelated to one another within personality types, the local regression will be fairly flat in both regions dominated by one particular type—the upper right (composed mostly of Type A individuals) and the lower left (composed mostly of Type B individuals). There will be a positive slope toward the center of the scatterplot due to the mixture of personality types in that region. This slope will reach a maximum where the groups intersect. Hence, taxonic latent structure yields a steplike or S-shaped curve. For dimensional latent structure, there would be a fairly constant positive slope across the entire scatterplot, yielding a comparably straight line. Panel A of Figure 2 presents sample MAXSLOPE plots for both latent structures.

Maximum covariance (MAXCOV). MAXCOV (Meehl, 1973; Meehl & Yonce, 1996) is also based on the statistical behavior of indicators in the vicinity of group mixture.[1] Suppose that a sample of Type A and Type B individuals is

FIGURE 2
Sample Curves



NOTE: MAXSLOPE = maximum slope; MAXCOV = maximum covariance; MAMBAC = mean above minus below a cut. Taxonic (n = 600, base rate = .50, 2.00σ separation) and dimensional (n = 600, $\bar{r}_{ij}$ = .50) latent structures analyzed using three different taxometric procedures. Each graph contains a smoothed line generated by the locally weighted scatterplot smoother method.

divided into successive intervals according to their level of competitiveness (referred to as the input indicator) and that the covariance of the two remaining indicators—time urgency and hostility (referred to as output indicators)—is calculated within each interval. Intervals demarcating relatively low levels of competitiveness will contain mostly Type B individuals, whereas those demarcating relatively high levels of competitiveness will contain mostly Type A individuals. Thus, at either extreme of the input scale, the covariance between time urgency and hostility will approach zero. Covariance values will be higher within more centrally located input intervals that correspond to moderate competitiveness, reaching a maximum in the input interval containing an equal mixture of Type A and Type B individuals. Hence, in the MAXCOV procedure, taxonic latent structure yields a peaked curve. For dimensional structure, relatively constant positive covariances would be observed across all intervals of the input indicator, yielding a comparably flat line. Panel B of Figure 2 presents sample MAXCOV curves.

Mean above minus below a cut (MAMBAC). MAMBAC (Meehl & Yonce, 1994) is based on the fact that if latent taxa exist, there will be an optimal cutting score on any valid indicator for classifying individuals into these taxa. Suppose that cases are sorted along an input indicator, such as competitiveness, and that the efficiency of all possible cutting scores on this indicator is examined. To do this, means are computed on an output indicator, such as hostility, separately for cases falling above and below each cut. A MAMBAC curve is constructed by plotting the difference between hostility means above and below each cut on the competitiveness indicator. Latent taxa generate a curve that is peaked near the cutting score that best distinguishes the classes (e.g., with Type A individuals falling above the cut and Type B individuals falling below the cut), whereas latent dimensions generate comparably dish-shaped curves. Panel C of Figure 2 presents sample MAMBAC curves.

Two additional features of taxometric procedures are worthy of note. First, each procedure can be conducted using available indicators in all possible input and output combinations, permitting examination of the consistency of results across combinations. For example, if four indicators are available, each of these taxometric procedures can be performed 12 times, and a panel of graphs can be plotted for interpretation.[2] Second, each procedure can be used to estimate latent parameters such as the base rate of taxon membership (e.g., the proportion of Type A individuals) in the sample under investigation. These estimates can then be compared for consistency within and between procedures as further tests of the existence of taxa.

Comparison With Alternative Procedures

There exist other procedures for examining latent structure, most notably distributional analyses (e.g., inspection for bimodality or negative kurtosis, admixture analysis, commingling analysis), cluster analysis, and approaches that model the relationship between manifest and latent variables (latent class analysis, latent profile analysis, latent trait analysis, and factor analysis). A number of important limitations, however, render each of these procedure less effective than Meehl's taxometric method for empirically distinguishing latent taxa from dimensions.

Distributional analysis. There is a variety of ways in which a manifest distribution can be examined for clues to latent structure. One method is to look for bimodal or multimodal distributions (e.g., Harding, 1949), which are suggestive of latent taxa. However, even in the clearest case (two equal-sized groups), the individual distributions must differ by at least two within-group standard deviations before a visible dip emerges toward the center of the joint distribution and two modes become apparent (Murphy, 1964). Groups that are separated by lesser amounts might instead form a unimodal distribution that is flattened relative to the normal curve, yielding a negative kurtosis. Other methods, such as admixture or commingling analyses (e.g., MacLean, Morton, Elston, & Yee, 1976), use trial and error to determine the parameters of hypothetical subgroup distributions that, when combined, would generate the observed distribution.

The primary difficulty with each of these procedures is that as noted earlier, manifest structure need not—and often does not—correspond to latent structure. For example, a scale containing items of equal difficulty and steep discrimination will tend to yield a manifest bimodal distribution, regardless of the latent structure of the construct being assessed. By contrast, a scale containing items of widely varying difficulties will tend to yield a unimodal distribution regardless of latent structure (Grayson, 1987). Many other factors can also alter the relationship between latent and manifest structure, thereby undermining the results of any procedure that simply analyzes a manifest distribution (see Grayson, 1987, and Murphy, 1964, for extended discussions of these limitations). Finally, these approaches do not provide an independent means of checking the structural conclusions that they produce, as do the consistency tests of the taxometric method.

Cluster analysis. The procedures in this large analytical family seek to determine whether cases tend to cluster together in a multidimensional hyperspace (e.g., Sneath & Sokal, 1973; Sokal & Sneath, 1963). There are a tremendous number of clustering algorithms available, all sharing two common characteristics. First, some measure of similarity (or distance) is chosen to quantify the relations between all cases in a sample. Second, some mathematical rule is applied to parse these similarity values into clusters.

Several factors limit the ability of cluster analysis to distinguish taxonic from dimensional latent structure. For example, there is often no reliable way to determine the appropriate number of clusters (Grove, 1991a). This problem is compounded by an even greater concern: Most algorithms will always uncover clusters in the data, even if the latent structure is dimensional (see Grove & Andreasen, 1989; Meehl, 1979, 1992; and references contained therein for more detailed treatments of these and related issues). Even simply rearranging the rows in a data set can substantially alter the clusters produced by the many algorithms in which the order of cases determines how clusters are initialized. Moreover, in contrast to the role of independently derived consistency tests in the taxometric method, researchers seldom employ multiple clustering algorithms, and the handful of algorithms that predominate in psychological research (see Blashfield, 1976, 1984) seldom yield results that are consistent with one another (Golden & Meehl, 1980). Thus, although cluster analyses may be useful for classifying cases within a validated taxonomy, there is insufficient support for their use as tools to determine latent structure.

Latent class analysis and related approaches. A final family of four conceptually related analytic techniques models the association between manifest and latent variables: latent class analysis (e.g., Green, 1951; Lazarsfeld & Henry, 1968), latent profile analysis, latent trait analysis (e.g., IRT) (Embretson, 1996; Hambleton et al., 1991; Lord, 1980), and factor analysis (e.g., Gorsuch, 1983; Thurstone, 1935, 1947). These procedures differ according to the structure of the manifest variables that they analyze and the presumed structure of the inferred latent variable(s). Factor analysis, for example, is typically used to reduce a large number of continuously distributed items to a smaller number of latent factors that are nearly always presumed to be dimensional in nature. Latent class analysis is a categorical analogue of factor analysis, reducing a large number of manifest categories to a smaller set of latent categories. Latent profile analysis and latent trait analysis are, in a sense, hybrid procedures: The former uses manifest continua to infer latent categories, whereas the latter uses manifest categories to infer latent continua.

Although each of these procedures can provide valuable information when used for either exploratory (data reduction) or confirmatory (testing a conjectured latent structure) purposes, none is ordinarily employed to test the competing hypotheses of taxonic and dimensional latent structure. For example, Waller and Meehl (1998) noted that despite passages in Thurstone's (1935, 1947) classic treatises on factor analysis dealing with the possibility of categorical factors, it is usually presumed that factors represent latent dimensions.[3] Moreover, none of these methods makes use of multiple consistency tests to help identify faulty conclusions. Thus, like cluster analysis, these four procedures may be of greater value once latent structure has been established as either taxonic or dimensional in nature.

Conclusions

Each of the procedures described above has characteristics that limit its ability to distinguish taxa from dimensions at the latent level. The real test of any method, however, lies in empirical evaluations of its efficacy. A considerable body of research using Monte Carlo simulations (Cleland & Haslam, 1996; Cleland, Rothschild, & Haslam, 2000; Haslam & Cleland, 1996; Meehl, 1973; Meehl & Golden, 1982; Meehl & Yonce, 1994, 1996; J. Ruscio, 2000) and "pseudo problems" (e.g., evaluating known latent structures such as that of biological sex using empirical data) (Gangestad & Snyder, 1985; Korfine & Lenzenweger, 1995; Meehl & Golden, 1982; Trull, Widiger, & Guthrie, 1990) has demonstrated the ability of taxometric procedures to accurately distinguish taxonic from dimensional latent structure. Despite decades of research, none of the alternative methods developed for evaluating latent structure has achieved this level of success in Monte Carlo or pseudoproblem trials (cf. Meehl & Golden, 1982).

CLASSIFICATION USING BAYESIAN PROBABILITIES OF TAXON MEMBERSHIP

Once the latent structure of a construct has been established, the next step is to use this knowledge to determine which measurement model is most appropriate for the construct. The most widely used measurement models in psychological assessment, particularly the sophisticated IRT models of recent vintage, are premised on the existence of latent dimensions. However, if a construct is taxonic, it makes little sense to plot item characteristic curves along the values of a latent dimension. Instead, methods are needed to classify individuals into taxa, a task for which classification models such as Bayes's theorem are eminently well suited.

Armed with an estimate of the taxon base rate in one's sample, as well as the valid and false positive rates of available indicators, one can calculate the probability that an individual belongs to a taxon given his or her response pattern—$Pr(t|RP)$—using the following formula (Waller & Meehl, 1998, p. 29):

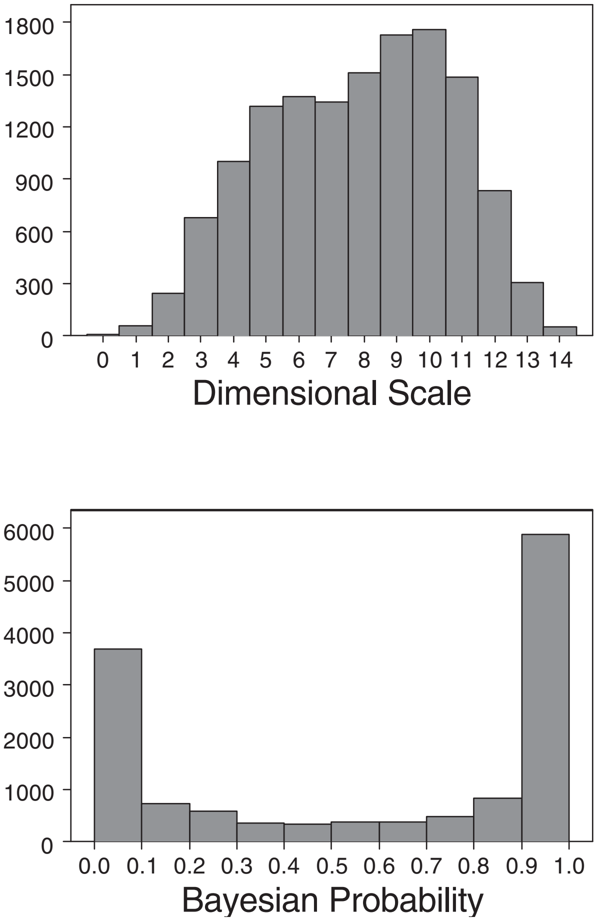
$$Pr\left(t \middle| RP\right) = \frac{P \prod_{i=1}^{v} pt_i^{\theta} qt_i^{1-\theta}}{P \prod_{i=1}^{v} pt_i^{\theta} qt_i^{1-\theta} + Q \prod_{i=1}^{v} pc_i^{\theta} qc_i^{1-\theta}} \tag{1}$$

where Π is the cumulative product operator, P is the base rate of taxon membership in the relevant population, $Q = 1 - P$, v is the number of indicators, pt_i is the valid positive rate achieved by each indicator, $qt_i = 1 - pt_i$, pc_i is the false positive rate achieved by each indicator, $qc_i = 1 - pc_i$, and $\theta = 1$ for a positively keyed response on an indicator, 0 otherwise.

To illustrate the power of this Bayesian model, consider its application to the pseudo problem of classifying the sexes. We use this as our example because the latent structure of biological sex is indisputable: It consists of two latent taxa, men and women. At the same time, the availability of a large data set with multiple indicators of biological sex allowed us to compare the efficacy of IRT and Bayesian models for classification. Using data from the Hathaway Data Bank (see J. Ruscio & Ruscio, 2000, or Waller, 1999, for descriptions of this database), 14 Minnesota Multiphasic Personality Inventory (MMPI) items from the Masculinity-Femininity Scale (Mf) having high corrected item-total correlations and varying difficulty levels were used to classify 13,684 adults—8,056 women (keyed as the taxon) and 5,628 men (keyed as the complement)—according to their sex. The MMPI Mf items were summed to yield a 15-point (0 to 14) dimensional scale on which each individual received a score. In addition, each individual's probability of taxon membership was calculated according to Bayes's theorem using the formula above. Thus, the traditional approach of dimensional scaling was compared to a classification model using the same set of indicator variables (MMPI items) in handling a construct with taxonic latent structure.

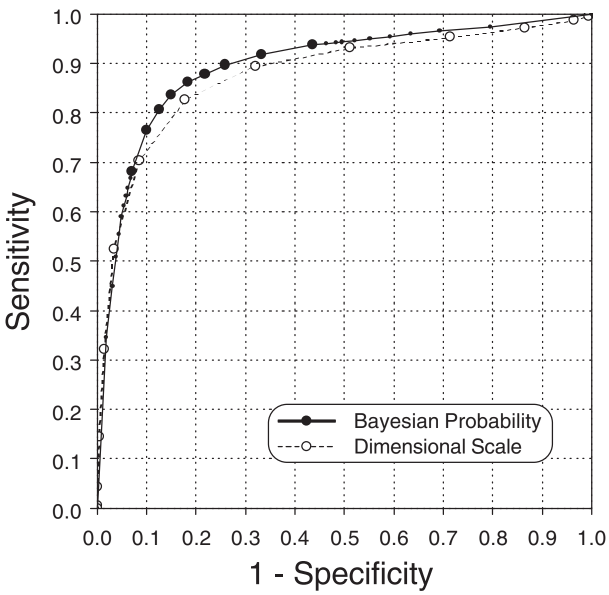
As can be seen in Figure 3, dimensional scale scores were unimodally distributed, with latent taxonicity almost completely obscured at the manifest level. The distribution of Bayesian probabilities, on the other hand, displayed a striking bimodality. To examine the efficiency with which individuals' sex could be classified by various cutting scores along the distributions, receiver operating characteristic (ROC) curves were plotted (see Figure 4). Bayesian probabilities achieved slightly greater accuracy (area under ROC curve = .897, confidence interval [CI] [95%] = .891 to .902) than did the dimensional scale (area = .879, CI [95%] = .873 to .885). Expressed as hit rates, the optimal cutting score along the distribution of Bayesian probabilities correctly classified 84.3% of all cases, whereas the optimal cut along the dimensional scale correctly classified 82.5% of all cases.

FIGURE 3
Frequency Distributions for the Dimensional Scale and Bayesian Probabilities



Although Bayesian classification yielded slightly better accuracy, the primary advantage of this approach was the extent to which it facilitated the location of an efficient cutting score to separate the latent classes. The pileup of cases at intermediate values along the dimensional scale made it difficult to choose an efficient cutting score. Moreover, this scale offered little tolerance for a suboptimal choice, as is evidenced by the wide spacing of successive cuts (open circles) on the ROC curve. In sharp contrast, it made relatively little difference where the bimodal distribution of Bayesian probabilities was cut. Cuts made anywhere from .10 to .90 (large dark circles) resulted in closely adjacent points on the ROC curve. In fact, all cutting scores between .30 and .70 on the Bayesian distribution achieved greater hit rates than did the optimal cut along the dimensional scale. This clearly illustrates that the selection of a cutting score is greatly simplified by use of a classification model when latent taxa exist.

FIGURE 4
ROC Curves for the Dimensional Scale and Bayesian Probabilities



NOTE: ROC = receiver operating characteristic. Open circles (dashed lines) represent the accuracy achieved through all cutting scores along the dimensional scale. Solid circles (solid lines) represent the accuracy achieved through cutting the distribution of Bayesian probabilities, with the nine cutting scores of .10 through .90 plotted as large points and cutting scores in increments of .01 out to the extremes of 0 and 1 plotted as small points.

The above demonstration indicates that when taxa are present, calculating Bayesian probabilities of taxon membership affords the simultaneous advantages of distributional continuity and bimodality. Continuity is useful in the event that situational demands call for the optimization of an index other than the overall hit rate of classification, allowing the selection ratio to be altered as desired to trade sensitivity for specificity or vice versa (see Meehl & Rosen, 1955). In this case, the Bayesian model provided much finer discriminations than did the summed scale scores, which yielded only 15 unique scores (0, 1, 2, . . . , 14). At the same time, bimodality simplifies the selection of an efficient cutting score and protects against a suboptimal choice.

Thus, we contend that the conventional preference for dimensional measurement in psychological research and the oft-heard claim that dimensions "retain more information" may be overly simplistic. An accurate classification of cases is much harder to achieve when items are combined using a dimensional scaling technique than when they are combined using a categorical measurement model. In the absence of empirical evidence regarding the latent structure of a given construct, it remains an open question whether a dimensional scaling or a classification approach would afford greater utility for psychological assessment, making the evaluation of latent structure particularly important.

IMPLEMENTATION OF A STRUCTURE-BASED APPROACH TO PSYCHOLOGICAL ASSESSMENT

Because latent structure is so important, we envision the rigorous development of psychological assessment devices beginning with a careful examination of the latent structure of each construct to be assessed. This would include delineation of all taxa (types and subtypes) through iterative applications of the taxometric method and evaluation of all dimensions through more conventional psychometric methods. We encourage readers to combine measurement models as suggested by empirical analysis of their constructs' latent structures, incorporating dimensional scaling and classification models as appropriate into a comprehensive assessment package. Whenever a distinction between taxa must be made, the relevant Bayesian probabilities can be calculated and used for classification. Whenever a continuum is encountered, a dimensional scaling model can be used to estimate individuals' scores along the continuum. This approach avoids the pitfalls stemming from the mismatch of latent structures and measurement models.

In recent years, technological developments have facilitated the computerized administration and scoring of psychological tests. Especially noteworthy is the rapidly expanding area of computerized adaptive testing (CAT), a highly desirable assessment method for reasons of brevity, reduced fatigue, elimination of hand-scoring errors, and immediacy of results (Embretson & Herschberger, 1999; Wainer et al., 1990). As has often been noted (e.g., Embretson & Herschberger, 1999), IRT models for dimensional scaling lend themselves well to implementation via CAT. Using CAT to implement an IRT model allows all scores on the latent trait to be estimated with equal precision through the administration of a custom-tailored subset of available items to each individual. However, although they are frequently paired in the literature (e.g., Embretson, 1996), IRT models and CAT interfaces are separable: Many measurement models can be implemented using the general strategies of CAT. Unfortunately, in keeping with psychologists' pervasive presumption of latent dimensionality, the use of CAT for classification is seldom discussed in the psychological assessment literature.[4]

Despite any apparent dissimilarity, there is a straightforward conceptual analogy between Bayesian classification and IRT models. To calculate an individual's score along a latent trait, an IRT model begins with an initial estimate of the trait score and refines it through responses to a set of items with known item characteristic curves. Bayesian classification begins by using the taxon base rate as an initial estimate of the probability of taxon membership and updates it through responses to a set of items with known valid and false positive rates. When IRT models are implemented using CAT, programmed algorithms guide item selection according to criteria such as content coverage, often involving the administration of a minimal number of items and/or the achievement of a certain standard error of measurement. Using a CAT interface, Bayesian classification could also proceed by selecting items according to content coverage, administering items of minimal redundancy until a threshold of high or low probability of taxon membership is crossed. At this point, the individual case would be classified into the taxon or complement, respectively. Thus, although our structure-based approach to assessment can be implemented using traditional paper-and-pencil methods, as can IRT or Bayesian models alone, CAT can be used to perform both scaling and classification functions. Regardless of the mode of implementation that is judged most feasible in any given assessment context, we suggest that measurement models should be chosen to best match the latent structure of the psychological construct being assessed.

NOTES

1. This procedure is conceptually quite similar to maximum slope (MAXSLOPE), which has been recommended as a "MAXCOV [maximum covariance] surrogate" when only two indicators are available (P. E. Meehl, personal communication, October 26, 1998). Another similar procedure is maximum eigenvalue (MAXEIG) (Waller & Meehl, 1998), a multivariate extension of the MAXCOV procedure.

2. MAXSLOPE and mean above minus below a cut (MAMBAC) can be performed twice using each pairwise combination of indicators by swapping each pair of indicators on the *x* and *y* axis for MAXSLOPE and switching the input and output for MAMBAC. For $k \geq 2$ indicators, one can calculate $k(k-1)$ MAXSLOPE or MAMBAC curves. MAXCOV can be performed thrice using each three-way combination of indicators by treating each member of the triplet as the input in turn. For $k \geq 3$ indicators, one can calculate $k(k-1)(k-2) / 2$ MAXCOV curves.

3. Waller and Meehl (1998) have developed L-Mode, a taxometric procedure that uses elements of factor analysis to distinguish taxonic from dimensional latent structure. The procedure works by examining the number of latent modes in the distribution of true scores on the first principal factor derived from a factor analysis of all available indicator variables. Unimodality is suggestive of dimensionality, whereas bimodality is suggestive of latent taxa.

4. For example, Embretson (2000) described 10 methodological frontiers in current research on psychological testing, all of which involved refinements of computerized adaptive testing and/or item response theory (IRT) models. Only 1 of the 10 refinements touched on qualitative distinctions between individuals: IRT models are being developed to incorporate categorical data as input. Nonetheless, even these models presume that the underlying construct is structured as a latent dimension.

REFERENCES

Ainsworth, M.D.S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Hillsdale, NJ: Lawrence Erlbaum.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Baldwin, M. W., & Fehr, B. (1995). On the instability of attachment style ratings. *Personal Relationships*, *2*, 247-261.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four hierarchical agglomerative methods. *Psychological Bulletin*, *83*, 377-388.

Blashfield, R. K. (1984). *The classification of psychopathology: Neo-Kraeplinian and quantitative approaches*. New York: Plenum.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, *100*, 316-336.

Cleland, C., & Haslam, N. (1996). Robustness of taxometric analysis with skewed indicators: I. A Monte Carlo study of the MAMBAC procedure. *Psychological Reports*, *79*, 243-248.

Cleland, C. M., Rothschild, L., & Haslam, N. (2000). Detecting latent taxa: Monte Carlo comparison of taxometric, mixture model, and clustering procedures. *Psychological Reports*, *87*, 37-47.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.

Dahlstrom, W. G. (1995). Pigeons, people, and pigeon-holes. *Journal of Personality Assessment*, *64*, 2-20.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*, 341-349.

Embretson, S. E. (2000, August). *Psychometric methods in the second century of testing*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Embretson, S. E., & Herschberger, S. (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum.

Flett, G. L., Vrendenburg, K., & Krames, L. (1997). The continuity of depression in clinical and nonclinical samples. *Psychological Bulletin*, *121*, 395-416.

Foa, E. B., & Foa, U. G. (1982). Differentiating depression and anxiety: Is it possible? Is it useful? *Psychopharmacology Bulletin*, *18*, 62-68.

Fraley, R. C., & Waller, N. G. (1998). Adult attachment patterns: A test of the typological model. In J. A. Simpson & W. S. Rholes (Eds.), *Attachment theory and close relationships* (pp. 77-114). New York: Guilford.

Friedman, M., & Rosenman, R. H. (1974). *Type A behavior and your heart*. New York: Knopf.

Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review*, *92*, 317-349.

Golden, R. R., & Meehl, P. E. (1979). Detection of the schizoid taxon with MMPI indicators. *Journal of Abnormal Psychology*, *88*, 217-233.

Golden, R. R., & Meehl, P. E. (1980). Detection of biological sex: An empirical test of cluster methods. *Multivariate Behavioral Research*, *15*, 475-496.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Grayson, D. A. (1987). Can categorical and dimensional views of psychiatric illness be distinguished? *British Journal of Psychiatry*, *151*, 355-361.

Green, B. F. (1951). A general solution for the latent class model of latent structure analysis. *Psychometrika*, *16*, 151-166.

Grove, W. M. (1991a). The validity of cluster analysis stopping rules as detectors of taxa. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology* (Vol. 1, pp. 313-329). Minneapolis: University of Minnesota Press.

Grove, W. M. (1991b). When is a diagnosis worth making? A statistical comparison of two prediction strategies. *Psychological Reports*, *68*, 3-17.

Grove, W. M., & Andreasen, N. C. (1989). Quantitative and qualitative distinctions between psychiatric disorders. In L. N. Robins & J. E. Barrett (Eds.), *The validity of psychiatric diagnoses* (pp. 127-139). New York: Raven Press.

Grove, W. M., & Meehl, P. E. (1993). Simple regression-based procedures for taxometric investigations. *Psychological Reports*, *73*, 707-737.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.

Hambleton, R. K., Swaminathan, H. S., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harding, J. P. (1949). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association*, *28*, 141-153.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1994). Psychopathy as a taxon: Evidence that psychopaths are a discrete class. *Journal of Consulting and Clinical Psychology*, *62*, 387-397.

Haslam, N., & Cleland, C. (1996). Robustness of taxometric analysis with skewed indicators: II. A Monte Carlo study of the MAXCOV procedure. *Psychological Reports*, *79*, 1035-1039.

Korfine, L., & Lenzenweger, M. F. (1995). The taxonicity of schizotypy: A replication. *Journal of Abnormal Psychology*, *104*, 26-31.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lenzenweger, M. F. (1999). Deeper into the schizotypy taxon: On the robust nature of maximum covariance analysis. *Journal of Abnormal Psychology*, *108*, 182-187.

Lenzenweger, M. F., & Korfine, L. (1992). Confirming the latent structure and base rate of schizotypy: A taxometric analysis. *Journal of Abnormal Psychology*, *101*, 567-571.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

MacLean, C. J., Morton, N. E., Elston, R. C., & Yee, S. (1976). Skewness in commingled distributions. *Biometrics*, *32*, 695-699.

Maser, J. D., & Cloninger, C. R. (Eds.). (1990). *Comorbidity of mood and anxiety disorders*. Washington, DC: American Psychiatric Press.

Meehl, P. E. (1962). Schizotaxia, schizotypy, schizophrenia. *American Psychologist*, *17*, 827-838.

Meehl, P. E. (1973). MAXCOV-HITMAX: A taxometric search method for loose genetic syndromes. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 200-224). Minneapolis: University of Minnesota Press.

Meehl, P. E. (1979). A funny thing happened to us on the way to the latent entities. *Journal of Personality Assessment*, *43*, 564-581.

Meehl, P. E. (1990). Toward an integrated theory of schizotaxia, schizotypy, and schizophrenia. *Journal of Personality Disorders*, *4*, 1-99.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, *60*, 117-174.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, *50*, 266-274.

Meehl, P. E. (1999). Clarifications about taxometric method. *Applied and Preventive Psychology*, *8*, 165-174.

Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127-181). New York: John Wiley.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194-216.

Meehl, P. E., & Yonce, L. J. (1994). Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports*, *74*, 1059-1274.

Meehl, P. E., & Yonce, L. J. (1996). Taxometric analysis: II. Detecting taxonicity using covariance of two quantitative indicators in successive intervals of a third indicator (MAXCOV procedure). *Psychological Reports*, *78*, 1091-1227.

Murphy, E. A. (1964). One cause? Many causes? The argument from the bimodal distribution. *Journal of Chronic Disease*, *17*, 301-324.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcements. *Psychological Monographs*, *80*(Whole No. 609).

Ruscio, A. M., Borkovec, T. D., & Ruscio, J. (2001). A taxometric analysis of the latent structure of worry. *Journal of Abnormal Psychology*, *110*, 413-422.

Ruscio, A. M., & Ruscio, J. (in press). The latent structure of analogue depression: Should the BDI be used to classify groups? *Psychology Assessment*.

Ruscio, A. M., Ruscio, J., & Keane, T. M. (in press). The latent structure of posttraumatic stress disorder: A taxometric investigation of reactions to extreme stress. *Journal of Abnormal Psychology.*

Ruscio, J. (2000). Taxometric analysis with dichotomous indicators: The modified MAXCOV procedure and a case removal consistency test. *Psychological Reports*, *87*, 929-939.

Ruscio, J., & Ruscio, A. M. (2000). Informing the continuity controversy: A taxometric analysis of depression. *Journal of Abnormal Psychology*, *109*, 473-487.

Ruscio, J., & Ruscio, A. M. (2001). *Distinguishing types from continua: A nontechnical introduction to the taxometric method*. Manuscript submitted for publication.

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 300-308.

Sneath, P.H.A., & Sokal, R. R. (1973). *Numerical taxonomy.* San Francisco: Freeman.

Sokal, R. R., & Sneath, P.H.A. (1963). *Principles of numerical taxonomy.* San Francisco: Freeman.

Strube, M. J. (1989). Evidence for the *type* in Type A behavior: A taxometric analysis. *Journal of Personality and Social Psychology*, *56*, 972-987.

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Trull, T. J., Widiger, T. A., & Guthrie, P. (1990). Categorical versus dimensional status of borderline personality disorder. *Journal of Abnormal Psychology*, *99*, 40-48.

Tyrka, A. R., Cannon, T. D., Haslam, N., Mednick, S. A., Schulsinger, F., Schulsinger, H., et al. (1995). The latent structure of schizotypy: I. Premorbid indicators of a taxon of individuals at risk for schizophrenia-spectrum disorders. *Journal of Abnormal Psychology*, *104*, 173-183.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Jr., Mislevy, R. J., Steinberg, L., et al. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum.

Waller, N. G. (1999). Searching for structure in the MMPI. In S. Embretson & S. Herschberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 185-217). Mahwah, NJ: Lawrence Erlbaum.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA: Sage.

Waller, N. G., Putnam, F. W., & Carlson, E. B. (1996). Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences. *Psychological Methods*, *1*, 300-321.

Waller, N. G., & Ross, C. A. (1997). The prevalence and biometric structure of pathological dissociation in the general population: Taxometric and behavior genetic findings. *Journal of Abnormal Psychology*, *106*, 499-510.

John Ruscio, Ph.D., is an assistant professor of psychology at Elizabethtown College. His primary areas of interest include the taxometric method and its applications, clinical judgment, statistical decision making, and conceptual and practical issues in diagnosis and classification.

Ayelet Meron Ruscio, M.S., is a doctoral candidate in clinical psychology at the Department of Psychology at Pennsylvania State University. Her interests include the nature and origins of normal and pathological anxiety, as well as general issues in the assessment, classification, and differential diagnosis of mental disorders.

Validity of the Wechsler Abbreviated Scale of Intelligence and Other Very Short Forms of Estimating Intellectual Functioning

Bradley N. Axelrod
John D. Dingell Department of
Veterans Affairs Medical Center

Performance on the Wechsler Adult Intelligence Scale–III (WAIS-III) was compared to performance on the Wechsler Abbreviated Scale of Intelligence (WASI), as well as short form estimations of intellectual functioning derived from WAIS-III performance, in a mixed clinical sample of 72 participants. The WASI verbal IQ (VIQ) score was significantly higher than the WAIS-III VIQ, whereas performance IQ (PIQ) estimates all differed from actual WAIS-III PIQ and full scale IQ (FSIQ). Correlations of WAIS-III scores with WASI scores were consistently lower than were correlations between the WASI-III and all other short forms. Although maintaining administration times of 15 minutes for a two-subtest FSIQ and 30 minutes for a four-subtest FSIQ, the WASI did not consistently demonstrate desirable accuracy in predicting scores obtained from the WAIS-III. The results suggest that clinicians should use the WASI cautiously, if at all, especially when accurate estimates of individuals' WAIS-III results are needed.

Keywords: WAIS-III, WASI, short forms, intelligence, assessment

In response to requests over the years for a short and reliable measure of intellectual functioning, the Psychological Corporation released the Wechsler Abbreviated Scale of Intelligence (WASI) in 1999 as an independent test of intelligence. As the WASI was developed at the same time as the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III) (Wechsler, 1997), the measures were linked.

As noted in the WASI manual (Psychological Corporation, 1999), the scale was developed in an effort to establish a reliable short form of measuring intellectual functioning that was normed across the life span. The utility of a shortened scale would be for use as a screening instrument, an estimation of general intellectual functioning for research purposes, or as a reassessment for someone who had previously been given a more comprehensive evaluation. Some less convincing reasons noted in the WASI manual related that although there were numerous ways to shorten the Wechsler scales, "it is very time consuming for clinicians to review volumes of literature to decide which short form best suits their needs" (Psychological Corporation, 1999, p. 2). The WASI was seen as a standardized method of obtaining an estimate of WAIS-III summary scores.

The WASI is composed of the four subtests, two verbal and two performance, that had previously been shown to correlate most strongly with general intellectual functioning. Specifically, the subtests of Vocabulary and Similarities are used to estimate verbal IQ (VIQ), whereas Block Design and Matrix Reasoning are used to estimate performance IQ (PIQ). None of the items on any of the subtests are included in the corresponding subtests in the WAIS-III. Performance on each subtest is converted to an age-adjusted standardized score, from which VIQ and PIQ scores can be generated. Full scale IQ (FSIQ) estimates are

Extreme gratitude is extended to Elizabeth Abraham Lopez and Scott Larson for the data collection required for this study. Joseph F. Kulas is thanked for his participation in data collection and his helpful comments on an earlier draft of this manuscript. Correspondence regarding this article should be addressed to Bradley N. Axelrod, Psychology Section (11-MH-PS), 4646 John R. Street, John D. Dingell DVAMC, Detroit, MI 48201-1916; e-mail: bradley.axelrod@med.va.gov.

Assessment, Volume 9, No. 1, March 2002 17-23

generated by using the results from all four subtests (FSIQ-4), or by using only two subtests (Vocabulary and Matrix Reasoning) (FSIQ-2).

Of the 1,145 adults included in the standardization sample, 248 were evaluated twice, once with the WASI and once with the WAIS-III, with an intertest interval of approximately 1 month. The resulting correlations for Vocabulary, Similarities, Block Design, and Matrix Reasoning were .88, .76, .83, and .66, respectively. More important, correlations between IQ scores for verbal and performance were .88 and .84, respectively. Finally, WAIS-III FSIQ correlated .92 with the WASI FSIQ-4 and .87 with the WASI FSIQ-2. No studies to date have evaluated comparability of these two measures in a clinical sample.

An additional benefit of the WASI is that it generates an estimate of intellectual functioning that could produce VIQ and PIQ scores in addition to an overall summary score. Although a number of short forms exist for the WAIS-III (see Table O-7 through O-11 in Sattler & Ryan, 1999), few have been objectively evaluated with clinical samples. Although not the strongest short form in terms of validation analyses, the four-subtest short form proposed by Kaufman, Ishikuma, and Kaufman-Packer (1991) for the WAIS-revised (WAIS-R) generates separate estimates of VIQ and PIQ and is also one of the shortest to administer. Composed of Arithmetic, Similarities, Picture Completion, and Digit Symbol, this tetrad correlated .97 with FSIQ in the standardization sample of the WAIS-R (Kaufman et al., 1991) and only .90 in the standardization sample of the WAIS-III. This short form was computed via the formula from Sattler (1992, p. 1069) using data found in the WAIS-III Wechsler Memory Scale–III (WMS-III) technical manual (Psychological Corporation, 1997).

With regard to the length of time required for administration, the WASI manual (Psychological Corporation, 1999) states that the time required to generate the FSIQ-4 is approximately 30 minutes. The time required for the even briefer FSIQ-2 is approximately 15 minutes. The administration times offered in the WASI manual appear to be estimates derived from the standardization sample. For the WAIS-III, studies of administration time have reported slight differences (cf. Axelrod, 2001; Ryan, Lopez, & Werth, 1998), with average administration time for all of the 11 subtests required to obtain the three summary scores ranging from 60 to 90 minutes. Administration of the four WAIS-III subtests included for WASI FSIQ averaged 25 to 41 minutes, and the time required to administer the two subtests to compute FSIQ-2 averaged 11 to 22 minutes. Administration time for the tetrad proposed by Kaufman and colleagues (1991) is significantly faster than all other four-subtest short forms and approximately 25% of the time required to administer the full WAIS-III (Axelrod, 2001; Ryan et al., 1998). The actual time for administering the four subtests averaged from 16 to 23 minutes.

The present study sought to assess the validity of the WASI in estimating VIQ, PIQ, and FSIQ scores of the WAIS-III. The subtests included in the WASI were evaluated as a new short form, as was the existing tetrad (Kaufman et al., 1991) previously shown to be the fastest of the short forms. In addition to examining the validity of the WASI and two short forms, administration time for the WASI and its comparable WAIS-III subtests was measured.

METHOD

Participants

The sample was a heterogeneous group of 72 male patients seen for neuropsychological evaluation at a large urban tertiary veterans medical center. Included were 37 patients with neurological disorders (Alzheimer's disease, *n* = 13; seizure disorder, *n* = 9; history of concussion, *n* = 8; stroke, *n* = 4; vascular dementia, *n* = 2; and Huntington's chorea, *n* = 1) and 29 with psychiatric diagnoses (depressive disorder, *n* = 8; schizophrenia, *n* = 6; alcohol abuse, *n* = 5; bipolar, *n* = 4; personality disorders, *n* = 4; and adjustment disorder, *n* = 2). Six of the patients seen were found to have no neurological or psychological disorder. The patient sample averaged 53.7 (*SD* = 15.0) years of age and had obtained 12.2 (*SD* = 2.6) years of education. Premorbid intellectual functioning was estimated to be 92.9 (*SD* = 13.9) by performance on the North American Adult Reading Test (Blair & Spreen, 1989). Eighty-eight percent of the sample was right handed and 60% of White racial composition.

Procedure

The WASI (Psychological Corporation, 1999) and WAIS-III (Psychological Corporation, 1997) were administered and scored according to the standardized procedures specified by their respective manuals. The measures were administered by either psychology interns or a psychology technician, all of whom were quite familiar with both measures before beginning the study. Deviation quotients were computed as estimates of IQ scores as recommended by Tellegen and Briggs (1967). These IQ estimates were computed for the subtests from the WAIS-III that are used in the WASI to estimate VIQ (Vocabulary and Similarities), PIQ (Matrix Reasoning and Block Design), and FSIQ with both two (Vocabulary and Matrix Reasoning) and four subtests. For clarity sake, these versions will be referred to as the WAIS-III dyads and WAIS-III tetrad, respectively.

Similarly, VIQ, PIQ, and FSIQ deviation quotients were computed for the subtests from the tetrad (Kaufman et al., 1991) composed of Arithmetic, Similarities, Picture Completion, and Digit Symbol.

Administration order of the WAIS-III and WASI was counterbalanced, based on the last digit (odd or even) of the participants' medical record number, thereby dividing the sample into two groups. Participants performed one task in the morning and the other in the afternoon of the same day. Additional measures included in the neuropsychological battery were administered between the WASI and WAIS-III. The two groups (WASI first, WAIS-III second: $n = 35$; WAIS-III first, WASI second: $n = 37$) did not differ from each other with regard to age, education, or estimations of premorbid intellectual functioning.

RESULTS

Administration Order

The performance data for the subtests and summary scores of the WASI and WAIS-III appear in Tables 1 and 2. However, prior to addressing those data, the potential impact of administration order was evaluated via 2 × (2) (Order × [Task]) mixed factorial analyses of variance. Mixed design ANOVAs were performed for each of the four subtests as well as the four summary scores. For all four analyses on the subtests, significant main effects for task were observed (all Fs > 522, all ps < .001). However, no significant findings were observed for the main effect of order or for the order-by-test interaction. In other words, there was essentially no effect of practice, regardless of which test was administered first. The only difference on the subtests was the difference in performance on the two sets of tests, differences that will be more clearly explicated in the next section. WASI subtest scores were transformed to scaled scores ($M = 10$, $SD = 3$) to confirm the findings on comparable scales. Again, no interactions or main effects for order were observed. Only main effects for each of the measures were noted.

When the 2 × (2) mixed-design ANOVA was applied to the VIQ, PIQ, and FSIQ estimates, similar findings were observed. Specifically, no main effect for order or order-by-task interaction was observed for any of the measures. In contrast, main effects for task were seen for PIQ, $F(1, 70) = 65.5$, $p < .001$, across the tasks and between WAIS-III FSIQ and WASI FSIQ-4, $F(1, 70) = 9.4$, $p < .01$. No main effects were observed between WAIS-III and WASI VIQ, $F(1, 70) = 3.6$, $p = .06$, and FSIQ-2, $F(1, 70) = 1.2$, *ns*, scores. WASI PIQ and FSIQ scores were higher than comparable WAIS-III scores, whereas WASI VIQ was lower than WAIS-III VIQ. Because no main effect for order or interaction between order and task was observed, the two groups were combined for all subsequent analyses.

TABLE 1
Means and Standard Deviations of Relevant Subtest Scores for WASI and WAIS-III

Variable	*Mean*	*Standard Deviation*
WAIS-III (age scaled scores)		
Vocabulary	7.2	2.8
Similarities	7.4	2.7
Arithmetic	7.6	2.7
Picture Completion	6.4	2.2
Digit Symbol	6.2	2.4
Block Design	8.0	2.4
Matrix Reasoning	8.2	2.4
WASI (age *T*-scores)		
Vocabulary	36.5	12.1
Similarities	41.8	10.9
Block Design	44.8	9.0
Matrix Reasoning	43.2	10.2
WASI (age scaled scores)[a]		
Vocabulary	5.9	3.6[b]
Similarities	7.5	3.3[c]
Block Design	8.4	2.7[d]
Matrix Reasoning	8.0	3.1[c]

NOTE: WASI = Wechsler Abbreviated Scale of Intelligence; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition.
a. Calculated arithmetically from *T*-scores, scores have a mean of 10 and standard deviation of 3.
b. WASI score is significantly lower than WAIS-III score ($p < .001$).
c. WAIS-III and WASI scores are not significantly different.
d. WASI score is significantly higher than WAIS-III score ($p < .05$).

Comparison of the WASI and WAIS-III Short Forms to the WAIS-III

The performance statistics for each of the WASI and WAIS-III subtests appear in Table 1. The subtest age scaled scores used in computation of short form estimates appear first. For the WASI, scores are presented both in terms of the age-corrected *T*-scores generated by the WASI manual as well as the scaled scores for the same scores computed arithmetically.

Table 2 presents a comparison of WAIS-III VIQ, PIQ, and FSIQ summary scores to the WASI as well as to the short forms of the WAIS-III. For VIQ, the WASI score was significantly lower than the WAIS-III, although the difference was clinically trivial. Neither the WAIS-III VIQ dyad nor the Kaufman VIQ dyad differed significantly from WAIS-III VIQ. In contrast, PIQ for the WASI and the WAIS-III dyad was significantly higher than the WAIS-III PIQ. The Kaufman PIQ dyad resulted in scores that were significantly lower than the WAIS-III. FSIQ estimates derived from the WASI (FSIQ-4) and the WAIS-III tetrad were significantly higher than actual FSIQ from the WAIS-III. In contrast, the Kaufman FSIQ tetrad was sig-

TABLE 2
Comparison of WAIS-III Summary Scores to WASI and WAIS-III Short Forms

Variable	M *(*SD*)*	F *Value*[a]	η^2	*Internal Consistency*[b]	r[c]	r[*d*]
Verbal IQ (VIQ)						
WAIS-III VIQ	86.3 (12.3)			.969		
WASI VIQ	84.1 (14.4)	3.8*	.05	.961	.75	
WAIS-III dyad (V + S)	85.5 (13.0)	1.5	.02	.940	.90	.84
Kaufman VIQ (A + S)	86.0 (13.2)	0.3	.00	.917	.89	.80
Performance IQ (PIQ)						
WAIS-III PIQ	82.2 (10.0)			.941		
WASI PIQ	90.4 (12.6)	66.7***	.49	.956	.74	
WAIS-III dyad (MR + BD)	89.2 (11.6)	119.5***	.63	.925	.88	.80
Kaufman PIQ (PC + DSym)	77.8 (10.6)	39.7***	.36	.881	.84	.72
Full scale IQ (FSIQ)						
WAIS-III FSIQ	83.3 (11.1)			.975		
WASI FSIQ (four subtest)	86.0 (12.8)	9.4*	.12	.976	.82	
WAIS-III tetrad (V + S + MR + BD)	85.9 (12.1)	29.4***	.30	.959	.94	.90
Kaufman tetrad (A + S + PC + DSym)	79.9 (11.1)	38.5***	.36	.937	.91	.85
WASI FSIQ (two subtest)	84.5 (13.0)	1.2	.19	.963	.71	
WAIS-III FSIQ dyad (V + MR)	86.7 (12.1)	20.6****	.99	.945	.85	.79

NOTE: WAIS-III = Wechsler Adult Intelligence Scale–Third Edition; WASI = Wechsler Abbreviated Scale of Intelligence; V = Vocabulary; S = Similarities; A = Arithmetic; MR = Matrix Reasoning; BD = Block Design; PC = Picture Completion; DSym = Digit Symbol Coding.
a. Comparison of summary scores to WAIS-III scores.
b. Computed according to Moiser (1943).
c. Correlation with WAIS-III score.
d. Based on the correction for redundancy by Levy (1967).
$*p \leq .05$. $***p < .005$. $****p < .001$.

nificantly lower than obtained from WAIS-III FSIQ. In examining two-subtest estimations of WAIS-III FSIQ, the WASI FSIQ-2 did not differ significantly, whereas the WAIS-III dyad was significantly higher than obtained from WAIS-III FSIQ.

The correlations of each of the WASI and WAIS-III short forms with the summary scores derived for the full versions of the WAIS-III appear in the sixth column of Table 2. The comparison, within each summary score, of the correlations was accomplished via test comparisons of Fisher's *z* transformed scores. Specifically, for estimates of WAIS-III VIQ, the correlations for the WASI scores were significantly lower than the WAIS-III two-subtest short forms ($p < .001$), which did not differ from each other. These same results were obtained for correlations with WAIS-III PIQ. WAIS-III FSIQ correlated significantly higher for the four- and two-subtest short forms than it did with either the FSIQ-4 or FSIQ-2 from the WASI.

The fifth column of Table 2 presents the internal consistency of each of the tests and variations of the WAIS-III evaluated in this study. Using the formula presented by Moiser (1943), reliability data of the WASI were constructed from the data in Tables 5.1 and 5.8 in the WASI manual (Psychological Corporation, 1999). The data for the WAIS-III were derived from Tables 3.1 and 4.12 in the WAIS-III WMS-III technical manual (Psychological Corporation, 1997) using the same equation. The inclusion of reliability information is useful first to determine the acceptability of a short form based on its meeting the minimal criteria of internal consistency. Reliability scores for all but the Kaufman PIQ dyad fall above the "tolerated" composite reliability cutoff ($r > .90$) and many even fall above the "preferred" cutoff ($r > .95$) (Nunnally, 1978, p. 246). Sattler (1992) offered a more liberal level of acceptable internal consistency ($r > .80$), above which all of the estimation equations fall.

The second reason for calculating the reliabilities of the shortened forms of the WAIS-III lies in the application of a correction factor that takes into account the redundancy of correlating a short form with the longer version of a test that includes the same items. This correction (Levy, 1967) was applied to the correlations of the WAIS-III short forms to the WAIS-III summary scores that appear in column 6. The resulting corrected correlations are presented in the final column of the table (see Table 2). Once corrected for redundancy, the correlations of the WASI scores with WAIS-III scores do not appear to be as weak relative to correlations of the various two- and four-subtest short forms. In fact, the WASI correlations did not differ significantly from the short forms for any of the VIQ, PIQ, or FSIQ-2 estimates. The correlation of the WAIS-III tetrad with WAIS-III FSIQ was significantly higher than the correlation with the WASI FSIQ-4. The correlation between the WAIS-III FSIQ and Kaufman tetrad was no different from the other correlations.

The percentage of cases in which WAIS-III estimates fell within 3, 6, and 10 points of full WAIS-III scores appear in Table 3. VIQ, PIQ, and FSIQ have standard errors of measurement of 2.5, 3.7, and 2.3, respectively. Whereas one would expect 68% of the cases to fall within one standard error of measurement, none of the short forms appear to come close to this level of accuracy. PIQ estimates are clearly weaker than estimates for either VIQ or FSIQ. Even when allowing a range of error of ± 10, the WASI PIQ only captures 54% of the WAIS-III PIQ. The percentage of cases that fell within a range across the measures was evaluated relative to the WAIS-III summary scores (see Bruing & Kintz, 1977, pp. 222-224). The WASI summary scores of VIQ, PIQ, FSIQ-2, and FSIQ-4 consistently captured fewer scores within the range of the WAIS-III scores than did any of the other short forms (all *p*s < .01). The WAIS-III and Kaufman short forms did not differ from each other in prediction accuracy.

TABLE 3
Accuracy of WASI and WAIS-III Short Forms in Estimating WAIS-III Summary Scores

Variable	*Percentage of Cases ± 3 Points*	*Percentage of Cases ± 6 Points*	*Percentage of Cases ±10 Points*
Verbal IQ (VIQ)			
WASI VIQ	24	41	71
WAIS-III dyad (V + S)	49	74	97
Kaufman VIQ (A + S)	40	79	93
Performance IQ (PIQ)			
WASI PIQ	14	33	54
WAIS-III dyad (MR + BD)	20	51	76
Kaufman PIQ (PC + DSym)	40	67	83
Full scale IQ (FSIQ)			
WASI FSIQ (four subtest)	30	66	84
WAIS-III tetrad (V + S + MR + BD)	54	80	97
Kaufman tetrad (A + S + PC + DSym)	43	83	91
WASI FSIQ (two subtest)	29	50	73
WAIS-III FSIQ dyad (V + MR)	34	61	89

NOTE: WASI = Wechsler Abbreviated Scale of Intelligence; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition; V = Vocabulary; S = Similarities; A = Arithmetic; MR = Matrix Reasoning; BD = Block Design; PC = Picture Completion; DSym = Digit Symbol Coding.

Administration Duration of the WASI

The time required to administer each of the four subtests from the WASI as well as the comparable subtests from the WAIS-III appears in Table 4. Both Vocabulary and Similarities required significantly more time to administer with the WASI than was observed for the WAIS-III. Not surprisingly, the time required to obtain a VIQ summary score was likewise longer for the WASI than for the same two subtests for the WAIS-III VIQ dyad. Although the difference was statistically significant, the reader should be aware that the difference was slightly more than 2 minutes. This difference in administration time was also seen in the time required to administer all four subtests. The administration times for the present sample for FSIQ-2 and WASI FSIQ-4 of 17 and 34 minutes, respectively, are generally consistent with the times of 15 and 30 minutes, respectively, presented in the WASI manual (Psychological Corporation, 1999).

TABLE 4
Administration Time (in minutes) for WASI and WAIS-III Subtests and Estimation Forms

	WASI		*WAIS-III*		
Variable	M	SD	M	SD	t-*Value*
Subtests					
Vocabulary	10.9	5.9	9.3	3.0	2.1*
Similarities	5.3	2.4	4.4	1.8	3.0***
Block Design	11.3	3.4	10.9	2.6	0.8
Matrix Reasoning	5.9	4.6	6.0	3.8	0.2
Summary scores					
Verbal IQ (V + S)	16.2	7.6	13.7	4.1	2.7**
Performance IQ (MR + BD)	17.3	6.5	17.0	5.2	0.5
Four subtest full scale IQ[a]	33.6	12.6	30.7	7.1	2.2*
Two subtest full scale IQ[b]	16.8	8.8	15.3	5.1	1.5

NOTE: WASI = Wechsler Abbreviated Scale of Intelligence; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition; V = Vocabulary; S = Similarities; MR = Matrix Reasoning; BD = Block Design.
a. Four subtests include Vocabulary, Similarities, Matrix Reasoning, and Block Design.
b. Two subtests include Vocabulary and Matrix Reasoning.
**p* ≤ .05. ***p* < .01. ****p* < .005.

DISCUSSION

Touted as an alternative measure of assessing general intellectual functioning, the WASI was evaluated in a heterogeneous clinical sample. In contrast to the correlations in the standardization sample comparing WASI to WAIS-III summary scores of .84 to .92, the present clinical sample had correlations that ranged from only .71 to .82. The WASI VIQ summary score underestimated WAIS-III VIQ, whereas WASI PIQ and FSIQ-4 overestimated the comparable WAIS-III scores. Even less convincing was the relatively small number of cases that fell within one standard error of measurement (SEM) (≈ 3.0) for any of the WASI summary scores. Specifically, WASI scores fell within one SEM for at most only 30% of the cases.

The same subtests from the WAIS-III that are included in the WASI were used to derive VIQ, PIQ, FSIQ-2, and FSIQ-4 deviation quotients. The resulting VIQ as well as two- and four-subtest FSIQ summary scores were more highly correlated with WAIS-III scores than were the WASI scores. Furthermore, the number of cases that fell within one SEM of WAIS-III scores was better than the WASI for VIQ (49% of the cases) and FSIQ-4 (54% of the cases). As a comparison, the shortest four-subtest short form (Kaufman et al., 1991) was analyzed. Not only did the resulting VIQ, PIQ, and FSIQ forms have the weakest composite reliabilities when compared to the other short forms, but the PIQ and FSIQ summary scores were significantly lower than WAIS-III scores. Furthermore, corrected correlations were slightly lower than those obtained from the WAIS-III dyads and tetrad.

Required administration time for the WASI is similar to that reported in the WASI manual. For either the WASI or the comparable subtests of the WAIS-III, a little more than 15 minutes is required to obtain a two-subtest (Vocabulary and Matrix Reasoning) FSIQ and slightly more than 30 minutes is required for a four-subtest FSIQ. The administration times for Similarities, Block Design, and Matrix Reasoning were comparable to a previous study from our clinic (Axelrod, 2001). However, the administration of the Vocabulary subtest was 50% slower than in the prior study. As noted in the introduction, the average time needed to administer the subtests included in Kaufman's FSIQ tetrad is about 60% of the time required for the WASI (Axelrod, 2001; Ryan et al., 1998).

Clinicians should be cautioned that the findings discussed in this article may not extend to patient or nonpatient samples that differ significantly from the individuals evaluated here. Another warning regarding the generalizability of the findings pertains to the use of short forms derived from administration of a full test. It is certainly possible that fatigue or procedural learning on subtests administered early in the battery might affect performance on tests given later in the battery. Consequently, extracting scores from subtests imbedded in the complete battery might not generalize to administering those subtests in isolation. Finally, although some readers might argue that the sample size for the present study is small, the use of a repeated measures design resulted in sufficient power to complete the analyses while minimizing both Type I and Type II errors.

Even though the validity findings do not strongly support the regular use of the WASI, the authors should be praised for creating a task that is not affected by practice. In this clinical sample, there was no procedural benefit for the patient in first having been given the WAIS-III. This finding would not be surprising for Vocabulary, which does not appear to have a practice effect (cf. Brines, 1996). However, Similarities and Block Design, the latter of which is thought to benefit from procedural learning, have improved as much as one third of a standard deviation on reassessment. With the WASI and WAIS-III, it appears that being given either first does not affect performance on the other. On the other hand, the lack of practice effect can also be interpreted as demonstrating poor concurrent validity between the WAIS-III and WASI. In that case, one cannot use these measures to evaluate potential changes in performance over time.

The reader is encouraged to use caution when deciding to use any short form, as they are usually less stable than are the full versions of the same measure. With regard to the WASI, the utility of estimating VIQ and PIQ with the WASI is difficult to endorse when less than one half of the clinical cases obtained scores within 6 points of their WAIS-III VIQ and PIQ scores. Finally, if the clinician's goal is to obtain an accurate estimation of general intellectual functioning, the current results suggest that the WASI should not be used in the assessment of individual patients. This conclusion is particularly noteworthy when the margin of error of 6 points captures less than two thirds of the sample for FSIQ.

REFERENCES

Axelrod, B. N. (2001). Administration duration for the Wechsler Adult Intelligence Scale-III and Wechsler Memory Scale-III. *Archives of Clinical Neuropsychology*, *16*, 293-301.

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist*, *3*, 129-136.

Brines, D. B. (1996). Practice effects of the WAIS-R in a normal population (Doctoral dissertation, Wayne State University, 1996). *Dissertation Abstracts International*, *57*, 2935.

Bruing, J. L., & Kintz, B. L. (1977). *Computational handbook of statistics* (2nd ed.). Glenview, IL: Scott, Foresman, and Company.

Kaufman, A. S., Ishikuma, T., & Kaufman-Packer, J. L. (1991). Amazingly short forms of the WAIS-R. *Journal of Psychoeducational Assessment*, *9*, 4-15.

Levy, P. (1967). The correction for spurious correlation in the evaluation of short-form tests. *Journal of Clinical Psychology*, *23*, 84-86.

Moiser, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, *8*, 161-168.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Psychological Corporation. (1997). *WAIS-III WMS-III technical manual*. San Antonio, TX: Author.

Psychological Corporation. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI) manual*. San Antonio, TX: Author.

Ryan, J. J., Lopez, S. J., & Werth, T. R. (1998). Administration time estimates for WAIS-III subtests, scales, and short forms in a clinical sample. *Journal of Psychoeducational Assessment*, *16*, 315-323.

Sattler, J. M. (1992). *Assessment of children* (Rev. and updated 3rd ed.). San Diego, CA: Jerome M. Sattler Publisher, Inc.

Sattler, J. M., & Ryan, J. J. (1999). *Assessment of children: WAIS-III supplement* (Rev. and updated 3rd ed.). La Mesa, CA: Jerome M. Sattler Publisher, Inc.

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology*, *31*, 499-506.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–third edition: Administration and scoring manual*. San Antonio, TX: Psychological Corporation.

Bradley N. Axelrod, Ph.D., is currently the staff clinical neuropsychologist at the John D. Dingell Veterans Affairs Medical Center in Detroit, Michigan. His primary research interests fall in evaluating the reliability and validity of neuropsychological assessment measures. Other publications have evaluated alternative and shortened versions of tests, as well as the construction of new measures.

The Minnesota Multiphasic Personality Inventory–2 Across the Human Immunodeficiency Virus Spectrum

Tina Hanlon Inman
Julie K. Esther
Wendy T. Robertson
Colin D. Hall
Kevin R. Robertson
University of North Carolina at Chapel Hill

The Minnesota Multiphasic Personality Inventory–2 (MMPI-2) was used to assess individuals' patterns of psychological symptoms across the spectrum of HIV illness. Two hundred and twenty-five participants in the present sample were administered the MMPI-2, 61 were HIV-seronegative controls, 61 were asymptomatic, 36 were symptomatic, and 67 met criteria for AIDS. Symptomatic HIV-seropositive patients scored higher on the Hypochondriasis, Conversion-Hysteria, and Depression Scales. These differences appeared to be largely due to an increase in somatic complaints rather than an increase in other depressive symptoms. Group differences did not appear to be due to HIV-associated neuropsychological dysfunction. Interpretive strategies for the MMPI-2 and treatment considerations are discussed.

Keywords: HIV, AIDS, MMPI-2, distress, antiretroviral therapy

Initial studies of psychological symptoms in individuals infected with human immunodeficiency virus Type 1 (HIV-1) have found increased symptoms of distress (Atkinson et al., 1988; Catalan et al., 1992). Symptoms of depression and anxiety are the most frequently reported and may result from multiple stressors including loss of physical health and well-being, social isolation due to death of friends or fear of contagion, and loss of income due to physical illness. The central nervous system effects of HIV can also produce apathy and depression-like symptoms and must be considered when evaluating symptoms of depression.

Reports of anxiety and depression in HIV-seropositive (HIV+) individuals are common, with several studies having quantified these symptoms (Chuang, Jason, Pajurkova, & Gill, 1992; McDaniel, Fowlie, Summerville, Farber, & Cohen-Cole, 1995; Miller & Riccio, 1990; Pakesch et al., 1992). These studies illustrated that HIV+ patients have increased levels of self-reported anxiety and depression compared to HIV-seronegative (HIV–) controls. As might be expected in those with psychological distress, an increase in suicidal ideation has been found in HIV+ individuals (Marzuk et al., 1988; Perry, Jacobsberg, & Fishman, 1990).

This work was supported by grants from the National Institute of Neurological Diseases and Stroke (1 PO1 NS26680-01, NS34243, NS36518), National Institutes of Allergy and Infectious Diseases (AI-25868), and the National Center for Research Resources (GCRC RR00046). Parts of this work were presented at the Neuroscience of HIV Infection: Basic and Clinical Frontiers, Vienna, June 2-5, 1993. The abstract is published in *Clinical Neuropathology* (Vol. 12). Correspondence concerning this article should be sent to Kevin R. Robertson, Ph.D., AIDS Neurological Center, Neurology, 751 Clinical Sciences Building, Campus Box #7025, University of North Carolina at Chapel Hill, School of Medicine, Chapel Hill, NC 27599-7025; phone: 919-966-5549; fax: 919-966-2922; e-mail: kevinr@glial.med.unc.edu.

Assessment, Volume 9, No. 1, March 2002 24-30

Several studies have used the Minnesota Multiphasic Personality Inventory (MMPI) or a shortened variation to assess patterns of psychological distress in HIV+ individuals, primarily in homosexual males or intravenous drug users (IVDUs). Consistent with previous research findings indicating HIV+ individuals experience increased symptoms of psychological distress (e.g., Atkinson et al., 1988; Catalan et al., 1992), studies comparing the MMPI scores of HIV+ and HIV– homosexual males have shown HIV+ males to have elevated MMPI scores, primarily on the Depression (D) Scale (Drebing et al., 1994; Kovner et al., 1989; Moore, van Gorp, Hinkin, Holston, & Weisman, 1994).

Researchers also have compared the MMPI scores of asymptomatic (ASX) and symptomatic (SX) HIV+ homosexual males but have not consistently shown higher levels of psychological distress associated with more advanced stages of HIV (Moore et al., 1994). Drebing et al. (1994) found SX HIV+ males in their study scored significantly higher on the D Scale than ASX HIV+ males; however, it was concluded that the differences on the D Scale were due to an increased endorsement of somatic symptoms as the HIV-1 disease progressed.

The findings of research on patterns of psychological distress in HIV+ IVDUs have revealed elevated scores on specific scales of the MMPI. Hestad, Aukrust, Ellertsen, and Klove (1994) found that the HIV+ IVDU had significantly higher scores than the HIV– IVDU. In contrast, Pakesch et al. (1992) found that both the HIV+ and HIV– IVDU groups displayed significantly greater psychopathology than the non-IVDU controls and concluded that the symptoms were more likely due to lifestyle factors rather than HIV status. These findings suggest that demographic factors need to be taken into consideration in research on the psychosocial functioning of HIV+ IVDUs.

Only one study has been published that has used the MMPI-2 to examine psychological distress in HIV+ individuals. Svikis, Gorenstein, Paluzzi, and Fingerhood (1998) examined MMPI-2 scores in a small sample (N = 21) of HIV+ pregnant, inner-city, drug-dependent women and found that the most frequent two-point code type in their sample of HIV+ women was the Paranoia (Pa) Scale and Schizophrenia (Sc) Scale (n = 4). However, they did not include a comparison group or compare across disease stage.

To date, no studies in the literature have utilized the MMPI-2 to assess psychological symptoms across the spectrum of HIV-1 illness. Examination of the prior MMPI research with HIV+ samples reveals several key issues that need to be addressed. First, initial research efforts focused on homosexual males, resulting in a paucity of information about the symptoms of distress experienced by HIV+ women. There is a great need for women to be included in research studies because women make up approximately 19% of the HIV+ population and are one of the segments of the population that has the fastest growing rate of HIV infection (Centers for Disease Control and Prevention [CDC], 1999). Second, appropriate control groups should be included in the design of the research. Third, the influence of demographic factors should routinely be examined. In addition, the role of somatic symptoms, including neurocognitive complaints, in scale elevations should be examined, particularly when elevations in the D Scale are found.

The purpose of the present study is to assess psychological distress across HIV disease state using the MMPI-2, with the inclusion of female participants and an HIV– control group. Analyses were planned to control for the influence of both demographic characteristics and neurocognitive impairment on test scores. Examination of the Harris-Lingoes subscales (Harris & Lingoes, 1955, 1968) and the Content Scales (Butcher, Graham, Williams, & Ben-Porath, 1990) of the MMPI-2 was also planned to attempt to identify elevations on the D and Conversion-Hysteria (Hy) Scales that were due to somatic complaints versus other factors. It was hypothesized that elevations in depression and somatic complaints would exist independently from the neurocognitive impairments that can affect those with HIV-1. It was thought that SX HIV+ patients would exhibit elevations in depression and anxiety, as well as concerns related to health issues, compared to HIV– controls and ASX HIV+ participants.

METHOD

Participants

The sample consisted of 225 patients who were voluntary participants in a longitudinal study at the AIDS Neurological Center. Sixty-one participants were HIV–controls (CTRL). Participants in the CTRL group were recruited from family and friends of HIV+ patients, hospital employees, and volunteers who responded to newspaper ads and flyers. Sixty-one participants were HIV+ but ASX (CDC Categories A1-A2) (CDC, 1992), 36 were SX and met criteria for AIDS-related complex (CDC Categories B1-B2), and 67 met criteria for AIDS (CDC Categories A3, B3, C1-C3). Participants' demographic characteristics are presented in Table 1.

Among patients in the ASX group, the mean time since HIV+ diagnosis was 29.68 months (SD = 25.41, range = 2-109 months, n = 60). Thirty-three percent of participants in the ASX group were on antiretroviral medications, and 3% were on anti-infective medications. Patients in the SX group were diagnosed with HIV an average of 37.88 months prior to the assessment (SD = 30.28, range = 3-109

TABLE 1
Demographic Characteristics by HIV Serostatus and Disease Stage

Group (n)	*Mean Age* (SD)	*Mean Education* (SD)	*Percentage Female*	*Percentage White*
HIV– control (61)	35.98 (9.93)	15.41 (2.39)$_a$	66$_a$	79$_a$
HIV + asymptomatic (61)	36.23 (7.97)	13.98 (2.98)$_b$	21$_b$	67$_{a,b}$
HIV + symptomatic (36)	37.22 (6.99)	12.72 (2.29)$_c$	28$_b$	50$_b$
HIV + AIDS (67)	37.69 (8.45)	13.19 (2.30)$_{b,c}$	24$_b$	52$_b$

NOTE: Means with the same letter subscripts are not significantly different at $p < .05$. Age, $F(3, 221) = .55$, ns; Education $F(3, 221) = 11.65$, $p < .05$; Percentage Female, $\chi^2(3, N = 255) = 34.50$, $p < .05$; Percentage White, $\chi^2(3, N = 225) = 12.86$, $p < .05$.

months, $n = 34$). Participants in the SX group self-reported HIV-related symptoms starting an average of 23.56 months prior to the assessment ($SD = 23.68$, range = 1-74 months, $n = 16$). Fifty percent reported taking antiretroviral medications, and 29% reported taking anti-infectives. Participants in the AIDS group were diagnosed with HIV an average of 55.07 months prior to the assessment ($SD = 41.53$, range = 3-162 months, $n = 65$). Participants in the AIDS group reported HIV-related symptoms starting an average of 28.56 months prior to the assessment ($SD = 30.18$, range = 1-141 months, $n = 43$). Seventy-three percent reported taking antiretroviral medications, and 46% reported taking anti-infectives.

Materials

Participants were administered the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) as part of a larger protocol. Participants were also administered a neuropsychological battery that has been described elsewhere (Wilkins et al., 1990). Neuropsychological tests were scored using age- and education-based norms and were converted to *z* scores.

Procedure

Participants were recruited and interviewed from September 1988 to December 1999. Participants were enrolled from infectious disease clinics, hemophilia clinics, and statewide support groups, as well as by peer nomination. No individual was entered into the study specifically due to identified neurological or psychological difficulties. Participants were admitted to the General Clinical Research Center at the University of North Carolina at Chapel Hill. All participants completed a protocol that included psychological, neurological, neuropsychological, clinical neurophysiological (electroencephalogram and evoked potentials), and laboratory evaluations. Participants were administered the MMPI-2 by a clinical psychologist as part of the comprehensive psychological protocol. The comprehensive neuropsychological battery was also administered, and the results were used as a covariate in some analyses.

Statistical Analyses

All profiles were examined for validity. Based on elevations on either the Lie Scale (raw score > 7), Infrequency Scale (raw score > 21 for men and > 18 for women), Correction Scale (K) (raw score > 22), Variable Response Inconsistency Scale (raw score > 13), and/or True Response Inconsistency Scale (raw score > 13 or < 5), 33 profiles were identified as invalid. A chi-square analysis was performed to determine if differences existed between groups in the number of profiles that were found to be invalid. This analysis was not significant, $\chi^2(3, N = 225) = 2.26$, $p = .52$. These profiles were subsequently excluded from all further analyses.

Because there were significant demographic differences between groups on the variables of gender, race, and education, initial analyses were conducted to determine whether any demographic variables were significantly correlated with MMPI-2 basic clinical scale scores. A priori hypotheses focused on group differences based on progression of the disease and specific scales; therefore, a series of ANCOVA tests was conducted with disease state as the independent variable, MMPI-2 scores as the dependent variable, and significantly correlated demographic variables as covariates. The Harris-Lingoes subscales were examined for all significant basic clinical scales.

RESULTS

Preliminary Analyses

Validity checks were completed first, with all invalid profiles subsequently eliminated from the analyses. A correlational analysis was conducted to determine if each of the K-corrected MMPI-2 scales was significantly correlated with the following variables: disease state (CTRL, ASX, SX, AIDS), race, gender, and education. The

Hypochondriasis (Hs), D, and Psychasthenia (Pt) Scales were significantly correlated with disease state and education. The Hy Scale was significantly correlated with disease state, whereas the Psychopathic Deviate (Pd) and Sc Scales were significantly correlated with disease state, race, and education. The Masculinity-Femininity (Mf) Scale was significantly correlated with disease state and gender. The Pa, Mania, and Social Introversion Scales were not significantly correlated with disease state. Pearson correlation coefficients are presented in Table 2.

TABLE 2
Correlation of K-Corrected MMPI-2 Scale Scores, Disease State, and Demographic Variables

MMPI-2 Scale	*Disease State*	*Race*	*Gender*	*Education*
Hypochondriasis	.44*	.05	–.05	–.23*
Depresssion	.33*	.04	–.10	–.16*
Conversion-Hysteria	.33*	–.07	–.05	–.12
Psychopathic Deviate	.25*	.35*	.06	–.25*
Masculinity-Femininity	.18*	–.10	–.44*	–.07
Paranoia	.14	.12	–.04	–.15*
Psychasthenia	.22*	.10	–.07	–.21*
Schizophrenia	.27*	.27*	–.03	–.30*
Mania	.09	.41*	.13	–.34*
Social Introversion	.03	.04	–.03	–.20*

NOTE: K = Correction Scale; MMPI-2 = Minnesota Multiphasic Personality Inventory–2.
*$p < .05$.

Basic Clinical Scales

On those scales where a significant correlation was found with disease state, an ANCOVA was performed using significantly correlated demographic variables as covariates. Follow-up pairwise comparisons were performed to determine significant differences between individual groups on the K-corrected MMPI-2 basic scale *T*-scores. Significance levels were set at $p < .01$ to control for an inflated Type I error secondary to multiple comparisons. Significant differences were found according to disease state on the Hs, Hy, and D Scales with no significant covariates. On each of these scales, the CTRL group scored significantly lower than the SX and AIDS groups, which did not differ. On the D and Hy Scales, the ASX group scored significantly lower than the AIDS group but did not differ from the SX or CTRL groups. On the Hs Scale, the ASX group scored significantly lower than the SX and AIDS groups and did not differ from the CTRL group. Including the mean *z* score on neuropsychological tests as a covariate did not change the outcome of these analyses. The ANCOVA revealed that differences between groups on the Pd, Mf, Pt, and Sc Scales were due primarily to demographic differences (level of education, race, and gender). Table 3 contains mean *T*-scores and *F* values for the comparisons described above.

Harris-Lingoes Scales

The Harris-Lingoes Scales were examined only for the basic clinical scales that were found to differ significantly across disease state as determined by univariate ANCOVA. An ANCOVA was performed for the D and Hy Harris-Lingoes subscales with the demographic covariates for the parent scale included in the analysis. A significance level of $p < .01$ was again set to control for multiple comparisons. One of the five D subscales was found to differ between groups. Physical Malfunctioning (D3) was significantly higher in the SX and AIDS groups than in the CTRL nd ASX groups, $F(3, 191) = 10.07$, $p < .01$. Two Hy subscales were significantly different when compared between groups. On the Lassitude-Malaise (Hy3) subscale, the SX and AIDS groups reported more symptoms than the CTRL and ASX groups, $F(3, 191) = 13.31$, $p < .01$. Participants in the SX and AIDS groups reported significantly more symptoms on the Somatic Complaints (Hy4) subscale than did those in the CTRL group, $F(3, 191) = 5.13$, $p < .01$.

Content Scales

A correlational analysis was run to determine which of the 15 content scales was significantly correlated with the following variables: disease state, race, gender, and education. The Health Concerns (HEA) and Work Interference Scales were significantly correlated with disease state and education. The Depression Scale was significantly correlated with disease state, race, and education. The Fears Scale was significantly correlated with disease state, race, gender, and education. Follow-up ANCOVAs were run on each of these scales using the significant demographic correlates as covariates. Only the HEA Scale remained significantly different between disease state groups, with the SX and AIDS groups scoring significantly higher than the CTRL group and the SX group scoring significantly higher than the ASX group, $F(3, 191) = 9.53$, $p < .01$. Including neuropsychological test results as a covariate did not change the results of these analyses.

Code Types and Profiles

MMPI-2 profiles were examined by disease state group for common two-point code types. Code types were defined as the two highest basic clinical scale scores that exceeded a *T*-score of 65, excluding the Mf Scale. Interestingly, there was a high degree of variability in all groups,

TABLE 3
Comparisons Between Disease State Groups on K-Corrected Basic Scale *T*-Scores With Demographic Covariates

	CTRL		*ASX*		*SX*		*AIDS*			
Scale	*LSM*	SE	*LSM*	SE	*LSM*	SE	*LSM*	SE	F*(3, 191)*	*Significance*
Hypochondriasis	51.36$_a$	1.89	53.67$_a$	1.82	65.84$_b$	2.54	65.07$_b$	1.79	13.34**	CTRL = ASX < SX = AIDS
Conversion-Hysteria	52.00$_a$	1.94	54.77$_{a,b}$	1.96	62.25$_{b,c}$	2.69	63.97$_c$	1.89	8.25**	CTRL < SX = AIDS, ASX < AIDS
Depression	52.10$_a$	1.90	55.35$_{a,b}$	1.82	62.39$_{b,c}$	2.54	62.03$_c$	1.80	5.92**	CTRL < SX = AIDS, ASX < AIDS
Psychopathic Deviate	53.83	1.71	57.70	1.64	60.42	2.29	59.08	1.62	2.21	
Masculinity-Femininity	56.55	1.51	59.84	1.44	59.43	1.96	58.95	1.37	0.87	
Psychasthenia	52.81	1.92	56.25	1.84	60.69	2.57	58.75	1.81	2.42	
Schizophrenia	54.04$_a$	2.02	58.72$_{a,b}$	1.94	63.51$_b$	2.70	60.51$_{a,b}$	1.91	2.89	

NOTE: K = Correction Scale; CTRL = control group; ASX = asymptomatic; SX = symptomatic; LSM = least squares means. Means with the same letter subscripts are not significantly different at $p < .01$.
**$p < .01$.

with few code types represented repeatedly. No single code type was representative of any group. However, scale patterns became more homogeneous as the disease state progressed, with scales more highly correlated with somatic symptoms (Hs and Hy) becoming common high points. Table 4 contains the mean *T*-scores for the basic clinical scales by disease state group.

DISCUSSION

No published studies, to date, have examined differences on the MMPI-2 based on HIV-1 disease progression. The present findings support previous research indicating that HIV+ participants tend to score higher on measures of depression and somatic complaints than HIV– controls (Catalan et al., 1992; Hestad et al., 1994; Kovner et al., 1989; Miller & Riccio, 1990). Specifically, differences on the Hs, D, Pd, Mf, Pt, and Sc Scales between groups were found in this study. These findings are consistent with the results of prior studies of distress and clinical symptoms on the MMPI that have found that HIV+ patients reported greater amounts of depression, somatic complaints, and anxiety, as well as elevations on the Pd, Mf, and Sc Scales (Hestad et al., 1994; Moore et al., 1994; Pakesch et al., 1992). However, in the present study, when the analyses were controlled for demographic differences, the differences on these scales appeared to be due to disparities in race and gender distribution in the groups rather than the presence of HIV-1.

This study also established that patients meeting the criteria for AIDS and SX HIV-1 tended to report more symptoms of depression and somatic complaints than did ASX HIV patients. Closer examination of the Harris-Lingoes subscales that make up the D and Hy Scales revealed that the differences between groups may be largely due to physical symptoms related to the progression of HIV-1 as suggested by Drebing et al. (1994). SX and AIDS participants reported more symptoms on the Physical Malfunctioning subscale (D3) that assesses somatic symptoms and physical health. SX and AIDS participants also reported higher levels of weakness, fatigue, and dysphoria as measured by the Lassitude-Malaise subscale (Hy3). These results suggest that among SX HIV+ and AIDS patients, increased scores on scales Hs, D, and Hy may be largely due to physical symptoms associated with the progression of HIV-1.

From a clinical standpoint, practitioners using the MMPI-2 should note that although differences between HIV– and HIV+ participants in this study seemed to be related to demographic differences, this may not be true on an individual basis. An examination of two-point code types revealed a large amount of heterogeneity in all groups, with a substantial proportion of HIV+ participants (40%-69%) obtaining significant elevations on at least two scales. This suggests that a large number of HIV+ patients experience psychological distress that may or may not be directly related to the physical effects of HIV-1. Overall, one must be sensitive to the particular individual and the various factors (race, gender, education, socioeconomic status, psy-

TABLE 4
Mean K-Corrected *T*-Scores on Basic MMPI-2 Clinical Scales by Disease State

Scale	*CTRL*	*ASX*	*SX*	*AIDS*	*Difference Source*
Lie	49.06 (8.23)	50.62 (7.69)	50.25 (6.32)	51.14 (7.48)	
Infrequency	54.44 (13.25)	56.79 (16.28)	59.96 (14.42)	57.51 (14.74)	
Correction	49.69 (9.18)	45.51 (10.30)	46.61 (8.75)	48.23 (10.66)	
Hypochondriasis	50.65 (10.33)$_a$	53.66 (12.55)$_a$	66.36 (14.95)$_b$	65.49 (15.37)$_b$	Disease state
Depression	51.48 (11.66)$_a$	55.23 (13.75)$_b$	62.82 (14.43)$_b$	62.53 (13.76)$_b$	Disease state
Conversion-Hysteria	52.00 (12.20)$_a$	54.77 (13.00)$_{a,b}$	62.25 (15.25)$_{b,c}$	63.96 (16.47)$_c$	Disease state
Psychopathic Deviate	52.11 (12.82)$_a$	57.45 (10.20)$_{a,b}$	61.29 (14.27)$_b$	60.51 (13.48)$_b$	Race
Masculinity-Femininity	53.52 (11.61)$_a$	61.28 (10.43)$_b$	60.79 (9.82)$_b$	59.83 (11.72)$_b$	Gender
Paranoia	52.72 (11.86)	58.04 (14.48)	60.43 (12.00)	57.70 (15.32)	
Psychasthenia	51.80 (10.89)$_a$	56.25 (13.10)$_{a,b}$	61.43 (15.15)$_b$	59.35 (15.14)$_b$	Education/disease state trends
Schizophrenia	51.85 (10.84)$_a$	58.53 (14.83)$_{a,b}$	64.82 (17.25)$_b$	62.12 (16.05)$_b$	Race
Mania	52.50 (11.38)	55.08 (10.81)	54.64 (12.46)	55.51 (11.75)	
Social Introversion	49.41 (10.78)	49.43 (10.98)	53.29 (13.42)	49.68 (11.43)	

NOTE: K = Correction Scale; MMPI-2 = Minnesota Multiphasic Personality Inventory–2; CTRL = control group; ASX = asymptomatic; SX = symptomatic. Means with the same letter subscript are not significantly different at $p < .01$ with no covariates. Numbers in parentheses are standard deviations.

chiatric history, and coping style) that may influence MMPI-2 scores beyond the presence of HIV-1.

REFERENCES

Atkinson, J., Grant, I., Kennedy, C., Richman, D., Spector, S., & McCutchan, J. (1988). Prevalence of psychiatric disorders among patients infected with HIV. *Archives of General Psychiatry*, *45*, 859-864.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory–2; Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 content scales*. Minneapolis: University of Minnesota Press.

Catalan, J., Klimes, I., Day, A., Garrod, A., Bond, A., & Gallwey, J. (1992). The psychosocial impact of HIV infection in gay men: A controlled investigation and factors associated with psychiatric morbidity. *British Journal of Psychiatry*, *161*, 774-778.

Centers for Disease Control and Prevention (CDC). (1992). Revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Morbidity and Mortality Weekly Report*, *41*(RR-17), 1-19.

Centers for Disease Control and Prevention (CDC). (1999). *HIV/AIDS among U.S. women: Minority and young women at continuing risk*. Washington, DC: Author. Retrieved July 3, 2000, from www.cdc.gov/hiv/pubs/facts/women.htm

Chuang, H. T., Jason, G. W., Pajurkova, E. M., & Gill, M. J. (1992). Psychiatric morbidity in patients with HIV infection. *Canadian Journal of Psychiatry*, *37*, 109-115.

Drebing, C. E., Van Gorp, W. G., Hinkin, C., Miller, E. N., Satz, P., Kim, D. S., et al. (1994). Confounding factors in the measurement of depression in HIV. *Journal of Personality Assessment*, *62*, 68-83.

Harris, R., & Lingoes, J. (1955). *Subscales for the Minnesota Multiphasic Personality Inventory* [Mimeographed materials]. San Francisco: Langley Porter Clinic.

Harris, R., & Lingoes, J. (1968). *Subscales for the Minnesota Multiphasic Personality Inventory* [Mimeographed materials]. San Francisco: Langley Porter Clinic.

Hestad, K., Aukrust, P., Ellertsen, B., & Klove, H. (1994). Psychological difficulties related to human immunodeficiency virus-1 infection in intravenous drug users. *Acta Psychiatrica Scandinavica*, *90*, 25-31.

Kovner, R., Perecman, E., Lazar, W., Hainline, B., Kaplan, M. H., Lesser, M., et al. (1989). Relation of personality and attentional factors to cognitive deficits in human immunodeficiency virus-infected participants. *Archives of Neurology*, *46*, 274-277.

Marzuk, P., Tierney, H., Tardiff, K., Gross, E., Morgan, E., Hsu, M., et al. (1988). Increased risk of suicide in persons with AIDS. *Journal of the American Medical Association*, *259*, 1333-1337.

McDaniel, J. S., Fowlie, E., Summerville, M. B., Farber, E. W., & Cohen-Cole, S. A. (1995). An assessment of rates of psychiatric morbidity and functioning in HIV disease. *General Hospital Psychiatry*, *17*, 346-352.

Miller, D., & Riccio, M. (1990). Non-organic psychiatric and psychosocial syndromes associated with HIV-1 infection and disease. *AIDS*, *4*, 318-388.

Moore, L., van Gorp, W., Hinkin, C., Holston, S., & Weisman, J. (1994). Frequencies of MMPI-168 code types among asymptomatic and symptomatic HIV-1 seropositive gay men. *Journal of Personality Assessment*, *63*, 574-578.

Pakesch, G., Loimer, N., Grunberger, J., Pferesmann, D., Linzmayer, L., & Mayerhofer, S. (1992). Neuropsychological findings and psychiatric symptoms in HIV-1 infected and non-infected drug users. *Psychiatry Research*, *41*, 163-177.

Perry, S., Jacobsberg, L., & Fishman, B. (1990). Suicidal ideation and HIV testing. *Journal of the American Medical Association*, *263*, 679-682.

Svikis, D., Gorenstein, S., Paluzzi, P., & Fingerhood, M. (1998). Personality characteristics of treatment-seeking HIV+ pregnant drug dependent women. *Journal of Addictive Diseases*, *17*, 91-111.

Wilkins, J., Robertson, K., van der Horst, C., Robertson, W., Fryer, J., & Hall, C. (1990). The importance of confounding factors in the evaluation of neuropsychological changes in patients infected with human immunodeficiency virus. *Journal of Acquired Immune Deficiency Syndromes*, *3*, 938-942.

Tina Hanlon Inman holds a doctorate in clinical psychology from the University of Kentucky and completed a postdoctoral fellowship at the University of North Carolina at Chapel Hill with an emphasis on neuropsychology.

Julie K. Esther holds a doctorate in clinical psychology from the University of North Carolina at Greensboro and is currently completing a postdoctoral fellowship.

Wendy T. Robertson holds a master's in clinical psychology from Western Carolina University and currently conducts clinical research in the neurological, neuropsychological, and psychological aspects of HIV and AIDS.

Colin D. Hall is a neurologist and vice chair of neurology at the University of North Carolina at Chapel Hill where he directs the AIDS Neurological Center.

Kevin R. Robertson is a clinical psychologist and director of psychological and neuropsychological research for the Department of Neurology and the AIDS Neurological Center at the University of North Carolina at Chapel Hill.

Perinatal Bereavement Grief Scale

Distinguishing Grief From Depression Following Miscarriage

Jennifer Boyd Ritsher

Department of Veterans Affairs and Stanford University

Richard Neugebauer

New York State Psychiatric Institute and Columbia University

The study evaluated the Perinatal Bereavement Grief Scale (PBGS), the first scale designed to measure grief following reproductive loss in terms of yearning for the lost pregnancy and lost baby. Participants included 304 women interviewed by telephone 1 to 3 times within 6 months after miscarriage. The PBGS had high internal consistency and test-retest reliability. It showed convergent validity with measures of attachment and investment in the child and divergent validity against measures of social desirability and depressive symptoms, supporting the conceptual distinction between grief and depression. Cross-cultural validity was acceptable whether tested by language (Spanish vs. English) or ethnicity (Hispanic vs. other). This measure of yearning enables study of the epidemiology and prognostic value of this key feature of mourning.

Keywords: miscarriage, perinatal bereavement, grief, depression, assessment

Attachment occupies a central place in child development and in adult relationships (Bowlby, 1980). Consistent with this view, bereavement—the termination of an attachment—increases risk for a host of illnesses and mortality (Clayton, 1998; Osterweis & Townsend, 1988). Psychological reactions to bereavement include numbness, disbelief, guilt, self-blame, anger, social isolation, and yearning and searching for the deceased (Clayton, 1982, 1998; Elders, 1995; Jacobs et al., 1987). The bereavement literature classifies these responses into three groups: (a) shock and numbness, (b) yearning for and preoccupation with the deceased, and (c) depression, disorganization (e.g., Bowlby, 1980), and more recently, anxiety (Jacobs et al., 1990; Surtees, 1995). To date, most research has focused on the third category, namely, depression-related responses to loss (Bruce, Kim, Leaf, & Jacobs, 1990; Clayton, 1998). However, the key role of attachment in the life cycle and, correspondingly, the serious health consequences of bereavement argue for broadening the scope of inquiry in this field to include the full range of psychological responses to loss so as to elucidate further the mourning process. Such research may also identify early predictors of maladaptive or pathological grief. The present article describes the psychometric properties of a measure of reactions to perinatal bereavement, focusing exclusively on yearning and preoccupation with the loss.

Proposals by Prigerson and colleagues (1999) for recognition of a new diagnostic entity called "traumatic grief" represent a more recent, independent effort to elaborate further our understanding of pathogenic responses to loss. As proposed, traumatic grief constitutes a syndrome distinct both from normal grief and post-traumatic stress dis-

This work was supported in part by Grants MH13043, MH39581, HD15509, and HD12207 from the National Institutes of Health, Bethesda, Maryland. The development of the Perinatal Bereavement Grief Scale in English and Spanish is much indebted to the work of Emma Diaz-Hernandez, Ph.D. We also thank Poorni Otilingam, M.P.H., for editorial assistance. Correspondence concerning this article should be addressed to Jennifer Boyd Ritsher, Center for Health Care Evaluation (MPD 152), VA Health Care System and Stanford University School of Medicine, 795 Willow Road, Menlo Park, CA 94025; e-mail: jennifer.ritsher@med.va.gov.

Assessment, Volume 9, No. 1, March 2002 31-40

order, with the stressor explicitly identified as the traumatic loss of an attachment (Prigerson et al., 1999). The specific components of traumatic grief that differentiate it from depression include yearning for the deceased and feeling stunned (Prigerson, Frank, et al., 1995; Prigerson, Maciejewski, et al., 1995). Studies of traumatic grief suggest that it significantly enhances risk for suicidality and heart disease even after controlling for depressive symptoms (reviewed in Prigerson et al., 1999). These findings underscore the value of examining the potential contribution of yearning—a primary element distinguishing traumatic grief from depression—to psychiatric and physical morbidity and possibly even mortality subsequent to the loss.

Several researchers investigating the psychological reactions specifically to pregnancy loss have attempted to move beyond the traditional focus in the bereavement field on depression. Building on the Expanded Texas Grief Inventory (Zisook, Devaul, & Click, 1982), Toedter and colleagues developed the 84-item Perinatal Grief Scale (PGS) (Toedter, Lasker, & Alhadeff, 1988) and a condensed 33-item version (Potvin, Lasker, & Toedter, 1989). The PGS measures a wide range of reactions to perinatal loss, including depression, anger, social functioning, spirituality, desire for counseling, locus of control, and guilt. It contains three subscales: Active Grief, Difficulty Coping, and Despair. However, each of these subscales contains items seemingly assessing depressive symptoms (such as "I feel depressed" in Active Grief, "I have considered suicide since the loss" in Difficulty Coping, and "I feel worthless since he or she died" in Despair), thereby precluding measurement of grief in the form of preoccupation with the loss separate from depression (Potvin et al., 1989). Correlations between the PGS and measures of depression are therefore not informative about the level of comorbidity of grief and depression or about the possibly unique role of grief in complicated mourning or other psychopathological outcomes.

Beutel, Will, Voelkl, von Rad, and Weiner (1995) developed the Munich Grief Scale, framed as a shortened version of the PGS-33. One of the expressed aims of the shortened version was to disentangle unique aspects of grief from more generic psychiatric symptoms. Whereas the scale as a whole does not do so, one subscale, Traurigkeit (Beutel, Will, et al., 1995), is restricted to feelings of missing the lost child. (Another paper published in English [Beutel, Deckardt, von Rad, & Weiner, 1995] translates *Traurigkeit* as *sadness*. However, in the context of the purposes of the research, and consistent with one sense of the meaning of *trauig*, the word would have been more appropriately translated as *mourning*.) The six items in the subscale pertain to crying for and missing the baby and having painful memories of the loss.

In prior research, both of conjugal and prenatal bereavement, study samples frequently comprised individuals with losses occurring at widely varying time intervals prior to assessment. Furthermore, the fact and circumstances of bereavement were sometimes based on recall rather than official documentation. Understandably, studies of conjugal bereavement often restrict the bereaved to persons in older age groups (Clayton, 1998), and prenatal bereavement studies sometimes examine reactions only to early or only to late loss (Nicol, Tompkins, Campbell, & Syme, 1986; Peppers & Knapp, 1980). Study samples are comparatively small and relatively homogeneous sociodemographically. By contrast, the current investigation interviewed a sample of more than 300 sociodemographically and ethnically diverse women seeking medical care for a clinically documented spontaneous abortion. The sample included women across the entire span of reproductive years with reproductive losses at any time up through 27 completed weeks of gestation, that is, 27 weeks plus 6 days. Furthermore, miscarrying women were assessed at three fixed time intervals after bereavement: 2 weeks, 6 weeks, and 6 months after loss.

The current literature suggests that Hispanics may have qualitatively different attitudes and behaviors regarding motherhood (Rodriguez & Kosloski, 1998; Sanchez-Ayendez, 1988; Zeskind, 1983), death (Shapiro, 1995), and the death of infants in particular (Walsh & McGoldrick, 1991). For this reason, we devote special attention to testing the validity and reliability of this instrument among Hispanics, an opportunity afforded by the substantial representation of Hispanics in our study population.

METHOD

Study Design

The present study is part of a large-scale investigation of the psychosocial and psychiatric sequelae of miscarriage, described in detail elsewhere (Neugebauer, 1987; Neugebauer et al., 1992a, 1992b, 1997). Conducting interviews at three points in the 6 months following loss allowed for both cross-sectional and longitudinal assessments of the reliability and validity of the Perinatal Bereavement Grief Scale (PBGS). The miscarriage cohort was derived from the cases of an antecedent hospital-based case-control study of risk factors for miscarriage, in which 77% of all women attending a New York City medical center for miscarriages from late 1984 to 1986 were interviewed (Kline, Stein, Susser, & Warburton, 1986).

The sample administered the PBGS was diverse in terms of ethnicity, education, income, and marital status (see Table 1). By self-report, 38% of participants were

Hispanic, 34% were White, and 21% were Black. About one third (29%) were interviewed in Spanish. The mean age was 30 years, and 38% were private patients (versus those receiving care in a clinic). Most women (64%) had at least one child; two thirds had no prior reproductive loss. Overall, the sample completing the PBGS is representative of the miscarriage cohort in the larger study, as regards major sociodemographic and reproductive history variables. The entire miscarriage cohort was in turn representative of the population of women attending the medical center for a miscarriage (Kline et al., 1986; Neugebauer et al., 1992a, 1992b).

We aimed to assess all miscarrying women at each of three time points: the 2nd week after miscarriage (2 weeks), Weeks 6 through 8 (6 weeks), and Weeks 26 through 35 (6 months). However, we were unable to interview all women at each time point. Overall, 73% ($n = 382$) of 523 eligible miscarrying women were interviewed at least once: 232 were first interviewed at 2 weeks, 114 first interviewed at 6 weeks, and 36 first interviewed at 6 months after loss. This staggered recruitment arises largely from the logistic difficulties encountered in scheduling women for assessments within narrow, fixed time intervals after loss.

The PBGS was introduced into the assessment battery several months after the start of participant recruitment. As a consequence, the first 78 miscarrying women out of the total 382 were not administered the PBGS. Of the remaining 304, 213 had no missing items on the PBGS scale. Of these 213, 133 were first interviewed at 2 weeks. At 6 weeks, 63 were interviewed for the first time and 119 were reinterviewed. At 6 months, 17 were interviewed for the first time and 163 were reinterviewed. One item on the PBGS, you wondered whether you would have had a boy or a girl, could not be asked of some women (14, 20, and 56 women at 2 weeks, 6 weeks, and 6 months, respectively) because they knew the child's gender based on a cytogenetic report. Our analysis excludes women missing data on this one item (although analyses not reported here showed that including them would not have altered our findings). In addition, 42 women were pregnant again at 6 months after loss and so could not answer the standard version of the PBGS. These pregnant women are dropped from the analyses of the 6-month data. There were no other missing data except for 1 additional woman missing one other item at each time point. In sum, at 2 weeks, a total of 133 women completed all 15 items of the grief scale; at 6 weeks, a total of 182 women; and at 6 months, a total of 178 women. Analyses reported below pertain to these 178 participants.

The staggered entry of study participants means that the women assessed at 6 weeks and 6 months comprised individuals interviewed previously as well as those being interviewed for the first time. Where relevant, all analyses described below were performed for the overall sample at each time point and separately for women being reinterviewed and those being interviewed for the first time. Similarly, analyses were repeated separately for each sociodemographic and reproductive history subgroup, as categorized in Table 1. Results from these analyses did not differ materially from the psychometric findings for the overall sample, except where noted below.

TABLE 1
Selected Sociodemographic and Reproductive Characteristics of Women Administered the Perinatal Bereavement Grief Scale

Characteristic	*Percentage or Mean* (N = *304)*
Sociodemographic	
Age (in years, *M* ± *SD*)	29.53 ± 6.19
Ethnicity (%)	
White	34.2
Black	21.1
Hispanic	37.5
Other	7.2
Interviewed in Spanish[a] (%)	28.9
Education (%)	
Less than high school	27.0
High school graduate	24.3
Some college	21.7
College graduate plus	26.9
Income (%)	
< $10,000	31.6
> $40,000	30.9
Current marital status (%)	
Single	26.6
Married/cohabiting	68.8
Other	7.9
Private payment of hospital bill (%)	38.2
Reproductive	
Nulliparous (%)	37.8
Living children (%)	
0	36.5
1	29.9
2	19.1
3+	14.5
Prior reproductive loss (%)	
0	66.1
1	21.4
2+	12.5
Weeks' gestation (%)	
11	49.3
12-15	23.0
16-27	27.6
28+	0.0

a. Of those interviewed in Spanish, 76% identified themselves as Hispanic, 3% as Black, 7% as White, and 14% as "other."

Measures

Participants were assessed by telephone in English or Spanish (Neugebauer et al., 1992a, 1992b). Each instrument was included in the assessment battery at all three time points except for the Crowne Marlowe and questions regarding the woman's (a) physical changes during pregnancy, (b) medical procedures associated with the pregnancy or miscarriage, and (c) preparations for the baby. These topics were covered at the time of the woman's first interview only, irrespective of whether the first interview occurred at 2 weeks, 6 weeks, or 6 months after loss.

PBGS. The PBGS (see Table 2) is a 15-item scale designed as a measure of grief and yearning for the lost pregnancy and the lost baby. The 15 items were derived from a review of the theoretical, clinical, counseling, and research literature (Belitsky & Jacobs, 1986; Bowlby, 1980; Kirkley-Best & Kellner, 1982; Jacobs et al., 1987; Leon, 1986; Lewis & Page, 1978; Phipps, 1981; Zeanah, 1989). The set of 7 pregnancy items contains statements such as "you dreamed you were still pregnant" and "you patted or held your belly as though you were still pregnant." Examples of the 7 items about the loss of the baby include "you wanted to hold the baby in your arms" and "you imagined what the baby would have looked like." The remaining item asks if "you felt physically ill when you thought about the miscarriage." Respondents indicate how often the statement has been true in the past week, using a 4-point Likert-type scale ranging from *rarely or none of the time, less than 1 day* (scored 1) to *most or all of the time, 5 to 7 days* (scored 4). Responses are summed to yield a total score (possible range, 15-60).

In general, reverse-coded items are useful to minimize the effects of response bias and to detect random responding. However, the potential intensity of grief and possible emotional vulnerability of the study population posed difficulties in creating comprehensible but inoffensive reverse-scored items. For example, the meaning of an item asking how often "you felt the baby was no longer inside you" is unclear, and asking how often "you did not want to hold the baby in your arms" seems insensitive. Thus, only one reverse-scored PBGS item proved feasible, "you found it easy to think about things other than the baby."

The Spanish translation of the PBGS was constructed through a systematic process of translation and backtranslation. Five female bilingual Ph.D. candidates in psychology composed the translation team. One individual translated all items from English into Spanish. Next, a second individual, blind to the English versions of the items, backtranslated the Spanish version into English. When comparing the two versions, only minor discrepancies emerged, the majority involving grammatical differences such as verb tense but not the intended affective, cognitive, or behavioral content of the item. These minor differences were resolved through a joint meeting of all

TABLE 2
The Perinatal Bereavement Grief Scale[a]—Items and Descriptive Statistics

	2 weeks (n = *133*)		*6 weeks* (n = *182*)		*6 months* (n = *178*)	
	M	SD	M	SD	M	SD
Let me read you a list of thoughts or feelings you might have had in the past week. Please tell me if you felt or thought these things rarely, some of the time, a moderate amount of the time, or most of the time in the past week.						
1. You found yourself walking like a pregnant woman.	1.66	1.04	1.14	0.47	1.15	0.48
2. You felt as if the baby were still inside of you.	1.71	1.05	1.21	0.61	1.14	0.45
3. You dreamed you were still pregnant.	1.42	0.85	1.22	0.50	1.21	0.56
4. You felt physically ill when you thought about the miscarriage.	1.96	1.16	1.51	0.88	1.44	0.87
5. You felt as if you were still pregnant.	1.73	1.00	1.30	0.71	1.29	0.71
6. You wanted to hold the baby in your arms.	2.14	1.29	1.62	0.98	1.51	0.93
7. You found yourself planning things for the baby as though you were still pregnant.	1.56	0.96	1.32	0.75	1.21	0.67
8. You found it easy to think about things other than the baby. (reversed)	2.61	1.15	1.95	1.15	2.13	1.22
9. You patted or held your belly as though you were still pregnant.	1.75	0.98	1.28	0.68	1.24	0.61
10. You felt as if there were an empty space inside of you.	2.35	1.23	1.55	0.84	1.52	0.89
11. You longed for the baby.	2.54	1.29	2.04	1.12	1.92	1.13
12. You felt like wearing maternity clothes.	1.71	1.08	1.31	0.75	1.31	0.78
13. You wondered whether you would have had a boy or a girl.	2.78	1.25	2.42	1.18	2.10	1.17
14. You imagined what the baby would have looked like.	2.42	1.27	1.98	1.04	1.90	1.14
15. You dreamt about the baby.	1.70	1.03	1.38	0.77	1.42	0.87

NOTE: Response options were labeled as follows: *rarely or none of the time (less than 1 day)*, *some of the time (1-2 days)*, *a moderate amount of time (3-4 days)*, and *most or all of the time (5-7 days)* (coded 1-4).

a. Developed by Richard Neugebauer, Ph.D., M.P.H., New York State Psychiatric Institute and Columbia University.

five bilingual translators, the project director, and the principal investigator (R.N.).

Center for Epidemiological Studies–Depression Scale (CES-D). The CES-D is a widely used scale for measuring affective, cognitive, and somatic symptomatology of depression. It consists of 20 items asking about the presence and duration of these symptoms in the preceding 7 days (Radloff, 1977). Response options are fixed in a 4-point Likert-type scale identical to that in the PBGS. Responses are summed to yield an overall index. The CES-D has excellent reliability and validity in psychiatric, general, and obstetric populations (e.g., Myers & Weissman, 1980; Radloff, 1977; Weissman, Sholomskas, Pottenger, Prusoff, & Locke, 1977; Zuckerman, Amaro, Bauchner, & Cabral, 1989).

Crowne Marlowe. The Crowne Marlowe Social Desirability Scale (Crowne & Marlowe, 1960) is a widely used 15-item measure of biased responding on the basis of social desirability. The items force a true or false response to exaggerated claims of social propriety, such as "you are always willing to admit it when you make a mistake." High scores are taken to indicate that the respondent is systematically giving socially appropriate answers or conforming to his or her perceptions of the investigators' wishes during other portions of the interview.

Indicators of attachment: Investment in and wantedness of the pregnancy and baby. The interview battery did not contain a specific measure of attachment to the pregnancy or baby. However, it did include investment behaviors reflecting the woman's commitment to the pregnancy and expected baby (investment items), as well as items aimed at measuring the extent to which the woman wanted to be pregnant (the Wantedness Scale) (Neugebauer et al., 1992a, 1992b).

Investment is a key feature of attachment. Attachment represents an affective bond and a symbolic engagement with the loved object. It is also usually associated with external behaviors aimed at maintaining the relationship with the loved object (Bowlby, 1980). The interview contained three items, separate from the PBGS, conveying the woman's symbolic engagement with, and investment in, the baby: (a) thinking about a name for the baby (yes/no), (b) making changes in the home in preparation for the baby's arrival (yes/no), and (c) purchasing items for the baby (yes/no). A fourth item aimed to assess specifically the woman's view as to the personhood of the unborn child: (d) thinking of what she had lost as the loss of a baby or child (versus any other response—losing a fetus, an embryo, or just blood and tissue). These behaviors and cognitions are predicated on a representation of the pregnancy as involving the creation of a person and new significant other. As such, we judged these actions and thoughts as more direct, clear, and unambiguous indicators of attachment than a woman's report of her feelings about the pregnancy. For brevity, these four items are hereafter referred to as the *investment items*.

The Wantedness Scale (Neugebauer et al., 1992b) measures the degree to which the pregnancy had been wanted. It was constructed from items concerning the woman's wish to conceive, her emotional reaction to learning of her pregnancy, and her consideration of elective abortion. Scores could range from 0 to 12, with the highest scores identifying women who had been trying to conceive, who reported being "very happy" when they learned they were pregnant, and who had given no thought to an elective abortion. At 6 months after loss, the Wantedness Scale correlated strongly with current contraceptive use (beta = –.308, $p < .001$), intent to have children (beta = .288, $p < .0001$), wish to get pregnant as soon as possible (beta = .267, $p < .005$), and pregnancy status (beta = .152, $p < .01$) (Neugebauer et al., 1992b). The internal consistency reliability across various demographic groups was acceptable (.68 to .78) (Neugebauer et al., 1992b).

Quickening. Women were also asked if they had experienced quickening, which is a sensory experience of the physical reality of the child. The response options were yes/no with a follow-up question where applicable regarding frequency. Although not a measure of attachment or attachment behavior per se, we hypothesized that this experience would increase attachment (Muller, 1992) and therefore be associated with higher levels of yearning after miscarriage.

Participant comprehension, response to interview, and interviewer monitoring. After completion of the interview, the interviewers rated participants in terms of their level of comprehension of interview questions, attentiveness and degree of distress during the interview, and any interruptions. For example, the interviewer rated "whether the participant 'cried or was tearful' during the interview." Comfort was measured using a 3-point Likert-type scale evaluating the participant's tension level (nervous, sporadic nervousness, mostly relaxed). If any break-offs occurred in the course of the questioning, the interviewer recorded the reason for the interruptions in an assigned place in the booklet. Moreover, interviewers enumerated items that were "especially hard for the participant to understand." About 10% of interviews were audiotaped and reviewed by office staff or by the principal investigator (R.N.). Interviewers met as a group biweekly so as to maintain uniformity of interviewing style and to discuss any difficulties with implementation of the study protocol or participants' reactions to the interview content.

RESULTS

The internal consistency reliability coefficient was .89 for the PBGS at each wave of data collection (n = 133 at 2 weeks, 182 at 6 weeks, 178 at 6 months), indicating a high, constant level of reliability in the 6 months after loss.

Test-retest reliability was high, statistically significant, and inversely related to the length of time between administrations of the measure. Among women interviewed both at 2 weeks and at 6 weeks after loss (n = 112), test-retest reliability was .69 (p < .001). Among women interviewed both at 6 weeks and at 6 months after loss (n = 105), the test-retest reliability was .67 (p < .001). Among women interviewed both at 2 weeks and at 6 months (n = 68), the coefficient was .48 (p < .001).

Divergent validity—yearning (PBGS) versus depression (CES-D). The CES-D and the PBGS are likely to have shared measurement variance because they both have a 4-point Likert-type format, straightforward content, similar length, and the same method of administration, and they were given during the same interview. In addition, they both measure constructs that involve affective distress. Thus, any difference between the two scales is likely to be largely the product of actual differences between the two constructs. The divergence of the two scales was assessed via correlation of the total scores and with exploratory factor analysis of the entire set of items from both scales combined. The correlation between the PBGS and CES-D scores at each wave ranged from .51 at 2 weeks after loss to .46 at 6 weeks to .34 at 6 months (p < .001 for each). Thus, the PBGS accounts for only 12% to 26% of the variance in the CES-D score.

Next, all items from the CES-D and the PBGS combined were subjected to exploratory factor analysis, using oblique rotation and specifying two factors. At 6 weeks and 6 months after loss, the two factors that emerged consisted of perfectly separated sets of items, one factor containing all 20 CES-D items and the other containing all 15 PBGS items. At 2 weeks after loss, 1 CES-D item (feeling that people dislike you) loaded on the PBGS factor for the full sample. It was the weakest item in the yearning factor and almost equally correlated with the CES-D factor (see Table 3).

Divergent validity—yearning (PBGS) versus social desirability (Crowne Marlowe). Correlations between the PBGS and the Crowne Marlowe would be high if the PBGS primarily measured women's perception of how they ideally "should" respond rather than their actual response to the loss. The divergent validity coefficient for the PBGS and the Crowne Marlowe was .05 (p > .05), indicating that the Crowne Marlowe score explains less than 1% of the variance in the PBGS score.

Convergent validity—yearning (PBGS) versus attachment-related items. We hypothesized that affirmative endorsement of each of the four investment items and elevated scores on the Wantedness Scale would be positively correlated with the PBGS, even after controlling for length of gestation. Because psychological attachment can start as soon as the pregnancy is known (Muller, 1992), the level of yearning following miscarriage is not necessarily strongly influenced by the length of gestation at loss. Consequently, the degree of association between gestational age and PBGS scores was not judged a critical test of convergent validity.

The attachment indicators (the four investment items and the Wantedness Scale) were assessed during the participant's first interview in the study, whether at 2 weeks, 6 weeks, or 6 months after loss. A psychometric examination of the PBGS and attachment indicators at 6 weeks and 6 months might reflect not only the relationship between these sets of measures but also both possible changes in and recollection of attachment indicators with passage of time since the loss. To avoid these psychometric ambiguities, we restrict this particular analysis to women first interviewed at 2 weeks after loss (n = 133).

As predicted, higher PBGS scores were more likely to occur among women who (a) had started thinking of a name for the baby (57% of 118 answering the item, beta = .24, p < .01), (b) thought of their loss as a lost baby or lost child rather than as a fetus or tissue (74% of 132, beta = .32, p < .001), (c) had bought things for the baby (9% of 113, beta = .35, p < .001), and (d) had made changes in their homes (17% of 113, beta = .31, p < .001). The women's report of the extent to which the pregnancy was wanted (wantedness) was only weakly correlated with the PBGS score (r = .16), but the relationship was statistically significant (beta = .18, p < .05). Regarding quickening, which we posited to be related to attachment, PBGS scores were also higher among women who had experienced quickening (30% of 117, beta = .25, p < .05). Each of these six regression models was adjusted for gestational age at time of loss.

Factorial validity—lost baby and lost pregnancy factors. The possible segregation of items pertaining to the lost pregnancy and lost baby into separate factors was tested in two ways. First, items were sorted into the two relevant groupings and subjected to the same reliability and validity analyses as described above for the full scale. It was expected that the correlation of the two sets of items with each other would be lower than the alpha score for the full 15-item scale because the alpha essentially represents the average of all possible split-half correlations. In the second set of analyses, the entire scale was factor analyzed and the resulting statistically derived factors inspected for

TABLE 3
Factor Analyses[a] of the Complete Item Pool From the Perinatal Bereavement Grief Scale (PBGS) and the Centers for Epidemiologic Studies–Depression Scale (CES-D), Replicated at Three Time Points

	Factor Loadings at Each Time Point					
	2 Weeks (n = *133)*		*6 Weeks* (n = *182)*		*6 Months* (n = *178)*	
Item	*Yearning*	*Depression*	*Yearning*	*Depression*	*Yearning*	*Depression*
PBGS items (paraphrased)						
Walking like a pregnant woman	.65	—	.73	—	.58	—
Felt like baby were still inside you	.61	—	.73	—	.59	—
Dreamed you were still pregnant	.63	—	.65	—	.63	—
Physically ill when thought about it	.75	—	.60	—	.67	—
Wanted to hold baby in your arms	.70	—	.66	—	.71	—
Felt you were still pregnant	.67	—	.76	—	.69	—
Planning things for the baby	.68	—	.69	—	.76	—
Think of things other than baby	.32	—	.33	—	.32	—
Patted or held your belly	.76	—	.73	—	.71	—
Empty space inside you	.63	—	.56	—	.69	—
Longed for the baby	.72	—	.62	—	.63	—
Felt like wearing maternity clothes	.61	—	.73	—	.71	—
Wondered if boy or girl	.46	—	.54	—	.64	—
Imagine what baby would have looked like	.58	—	.52	—	.77	—
Dreamt about the baby	.74	—	.73	—	.79	—
CES-D items (paraphrased)						
Felt people dislike you	.23	—	—	.46	—	.43
Keeping mind on doing	—	.58	—	.56	—	.67
Talk less than usual	—	.56	—	.66	—	.65
Felt sadness	—	.69	—	.77	—	.79
Felt fearful	—	.54	—	.59	—	.62
Everything you do is an effort	—	.56	—	.66	—	.65
Appetite poor	—	.48	—	.49	—	.60
Life had been a failure	—	.52	—	.62	—	.62
Felt lonely	—	.66	—	.75	—	.59
Shake and blues	—	.64	—	.80	—	.76
Could not get going	—	.59	—	.67	—	.69
Crying spells	—	.63	—	.68	—	.63
Depressed	—	.75	—	.78	—	.82
As good as other people	—	.37	—	.57	—	.52
Felt happy	—	.69	—	.71	—	.76
People were friendly	—	.24	—	.36	—	.33
Hopeful about the future	—	.34	—	.45	—	.52
Bothered by things	—	.42	—	.63	—	.74
Enjoyed life	—	.63	—	.61	—	.67
Sleep was restless	—	.59	—	.43	—	.56
Correlation between factors	.39		.34		.30	
Correlation between PBGS and CES-D scores	.51		.46		.34	

a. At each time point, a separate factor analysis was conducted, with oblique rotation and two factors specified. Factor loadings are given only for the higher factor.

their resemblance to the two a priori theoretically derived sets of items.

For each time point, the internal consistency reliabilities of each item subset (lost pregnancy and lost baby) were slightly lower than those reported above for the overall scale (all .89). The correlation between the two subscales ranged from .64 to .79. In addition, factor analysis of the PBGS at each time point after loss yielded a single main factor and one or two other weak factors. The weak factors had eigenvalues near 1 and did not represent meaningful subgroups of items. Therefore, the most parsimonious interpretation of the factor analyses is that the PBGS measures one factor, which we hypothesize to be yearning and pining for the deceased.

Cross-cultural validity. The reliability and validity analyses outlined above were repeated separately for interviews conducted in English and in Spanish. These analyses address the adequacy of the Spanish translation and the cross-cultural reliability and validity of the construct of yearning following bereavement. Because language of interview is only one proxy for cultural membership, the analyses were repeated contrasting people identifying themselves as Hispanic versus all non-Hispanic racial/ethnic groups combined (White, Black, other). Of the 304 women administered the PBGS at least once, 88 (29%) were interviewed in Spanish and 216 interviewed in English. Of those 216 interviewed in English, 169 (78%) identified themselves as non-Hispanic.

Analyses reported in this section were conducted for the Spanish versus English versions of the instrument and compared to the sample as a whole. Identical analyses comparing self-identified Hispanics versus non-Hispanics produced the same results unless noted otherwise.

As one measure of the Spanish version's ease of use, it had no more missing data than the English version. Furthermore, bilingual interviewer ratings of participants' level of item comprehension did not vary between the Spanish and English versions of the instrument. Nor was differential comfort with or response to the two versions of the PBGS reported in the postinterview ratings or at the biweekly interviewer meetings.

Across the three time points, the Spanish version's internal consistency reliability (alpha = .85-.91) was similar to the coefficients for the English version (alpha = .81-.89). The test-retest reliability was somewhat higher for the English version (*r* = .52-.72) than for the Spanish (*r* = .30-.66). The *r* = .30 was at the longest time span with an *n* of only 18, whereas there were 29 Spanish speakers completing the PBGS at both 2 weeks and 6 weeks after loss (*r* = .49) and 28 completing it at both 6 weeks and 6 months after loss (*r* = .66).

Tests of divergent validity were similar across cultures. The English and Spanish versions of the PBGS both have statistically significant, moderate correlations with the CES-D that decline over time (Spanish: *r* = .47-.33, English: *r* = .50-.38). There were too few Spanish speakers who had completed both instruments for a factor analysis. However, the factor analysis results for self-identified Hispanics (regardless of interview language) were essentially the same as for non-Hispanics and for the complete sample. Among self-identified Hispanics, factor analysis of CES-D and the PBGS items resulted in perfect separation of the two scales 6 weeks and 6 months after loss (see Table 3). At 2 weeks after loss, the separation between the two scales was less complete, with several CES-D items loading on the PBGS factor. The discrepant CES-D items at 2 weeks after loss were among the lowest correlated of those in the PBGS factor. Comparing social desirability (Crowne Marlowe) to yearning (PBGS), correlations were stronger albeit nonsignificant for Spanish speakers (*r* = .32, *p* = .06, *n* = 35) compared to English speakers (*r* = .10, *p* = .33, *n* = 96). In tests of convergent validity (PBGS vs. the indicators of attachment), results were the same after controlling for language of interview, and the language term was not significant.

DISCUSSION

The PBGS enjoys high reliability in the first 6 months after miscarriage and also evidences satisfactory convergent and divergent validity. The PBGS should be considered to measure a single construct of yearning and preoccupation with the deceased following reproductive loss. We could not confirm the existence of two separate factors, one pertaining to the lost pregnancy and the other to the lost baby.

Among the validity findings, the strong support for divergent validity—as indicated by the only modest correlation between the PBGS and the CES-D and the complete segregation of their respective items into two factors—is especially important. The PBGS was intended to measure yearning for the lost pregnancy and anticipated child. These findings indicate that this construct is distinct from the more general phenomenology of depression and affords psychometric support for further investigation of this key psychological reaction to bereavement.

Evidence for the convergent validity of yearning is also, in general, strong. Yearning tended to be greater among women who were invested in the pregnancy, as evidenced by thinking of a name for the baby, having thought of the loss as that of a baby, having bought things for the baby, or having made changes in their homes. Furthermore, those who had felt the fetus move inside them (quickening) also experienced greater yearning. However, yearning was only weakly correlated with the degree to which the woman had wanted the baby. These findings may reflect the fact that wanting a pregnancy is more distally related to attachment to and investment in the baby.

Finally, the cross-cultural validity and reliability results for the PBGS support the use of this instrument in Hispanic populations whether interviewed in English or Spanish. The cross-cultural validity was acceptable whether tested by language (Spanish vs. English) or ethnicity (Hispanic vs. other).

Despite the general similarity in psychometric findings among Hispanics and non-Hispanics in the sample, we could not assess with these data whether the items in the PBGS covered the relevant cultural domain to the same extent among Spanish and English speakers, only that the scale functioned equivalently. In other words, we did not test if the cross-cultural equivalence of the scale would be further improved by adding new items about other aspects of yearning (if any) that are especially salient in Hispanic populations. Replication would be needed to determine whether the lower test-retest reliability among Spanish speakers was a function of cross-cultural differences in the stability of perinatal grief over time or simply attributable to the smaller sample size.

Assessment of convergent validity was limited by the fact that there exist no other known scales of this construct (yearning) and hence no specialized body of literature with which to compare our results. Therefore, it was necessary to use more distal constructs such as indices of attachment as convergent validity indicators. Not surprisingly, the results, although offering substantial evidence of convergent validity, produced some inconsistencies, for example, the comparatively meager correlation in these analyses between yearning and wantedness. Nonetheless, the overall satisfactory to excellent psychometric performance of the PBGS in the present study supports its use for perinatal bereavement research.

Predictive criterion validity was also not assessed. Evaluation of predictive validity awaits future studies that test the ability of the PBGS to predict relevant outcomes in either a longitudinal observational study or one that monitors changes in PBGS scores in women who do versus do not receive bereavement counseling.

The only other measure in the perinatal field that also assesses yearning is the Sadness subscale from the Munich Perinatal Grief Scale (Beutel, Deckardt, et al., 1995; Beutel, Will, et al., 1995). That scale aims to tap only mourning for the lost child. It does not address yearning for the lost pregnancy. It is also largely confined to affective elements of yearning, for example, "I miss the baby," and does not extend, for example, to cognitive elements of yearning such as those in the PBGS (e.g., "You found yourself planning things for the baby as though were you were still pregnant") or somatic experiences of yearning (e.g., in the PBGS, "You found yourself walking like a pregnant woman"). As anticipated for a brief, narrowly focused subscale, internal consistency reliability was .86; 6-month test-retest reliability was .70. Correlation between this subscale and a scale measuring fearful and irritable depression was .55.

The creation and validation of this reliable measure of yearning, searching, and preoccupation with the deceased and the lost pregnancy permit further study of a construct representing one of the three central symptom groups present subsequent to loss—adding to the existing bodies of work on shock and numbness and on depression and disorganization. The PBGS can facilitate investigations of (a) the descriptive epidemiology of yearning; (b) the nature of the relationship between yearning for the deceased and the emergence, degree, and persistence of other bereavement reactions; and (c) the possible prognostic importance of this symptom cluster in terms of later clinical status and social impairment. Research on adult bereavement, specifically conjugal bereavement, has implicated preoccupation with the deceased and searching and yearning for the deceased as predictive of long-term functional impairment (Prigerson, Frank, et al., 1995). The PBGS permits a similar systematic investigation of this question in the context of perinatal bereavement.

REFERENCES

Belitsky, R., & Jacobs, S. (1986). Bereavement, attachment theory, and mental disorders. *Psychiatric Annals*, *16*, 276-280.

Beutel, M., Deckardt, R., von Rad, M., & Weiner, H. (1995). Grief and depression after miscarriage: Their separation, antecedents, and course. *Psychosomatic Medicine*, *57*, 517-526.

Beutel, M., Will, H., Voelkl, K., von Rad, M., & Weiner, H. (1995). Entwicklung und Erfassung von Trauer am Beispiel des Verlustes einer Schwangerschaft: Entwicklung und erste Ergebnisse zur Validitaet der Muenchner Trauerskala (NITS) [Assessment of grief after pregnancy loss: Development and initial validation of the Munich Grief Scale]. *Psychotherapie Psychosomatik Medizinische Psychologie*, *45*, 295-302.

Bowlby, J. (1980). *Attachment and loss: Vol. 3. Loss: Sadness and depression*. New York: Basic Books.

Bruce, M. L., Kim, K., Leaf, P. J., & Jacobs, S. (1990). Depressive episodes and dysphoria resulting from conjugal bereavement in a prospective community sample. *American Journal of Psychiatry*, *147*, 608-611.

Clayton, P. (1982). Bereavement. In E. S. Paykel (Ed.), *Handbook of affective disorders*. New York: Churchill Livingstone.

Clayton, P. J. (1998). The model of stress: The bereavement reaction. In B. P. Dohrenwend (Ed.), *Adversity, stress, and psychopathology* (pp. 96-110). New York: Oxford University Press.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting and Clinical Psychology*, *24*, 349-354.

Elders, M. A. (1995). Theory and present thinking in bereavement. *Issues in Psychoanalytic Psychology*, *17*, 67-83.

Jacobs, S., Hansen, F., Kasl, S., Ostfeld, A., Berkman, L., & Kim, K. (1990). Anxiety disorders during acute bereavement: Risk and risk factors. *Journal of Clinical Psychiatry*, *51*, 269-274.

Jacobs, S. C., Kosten, T. R., Kasl, S. V., Ostfeld, A. M., Berkman, L., & Charpentier, P. (1987). Attachment theory and multiple dimensions of grief. *Omega—Journal of Death and Dying*, *18*, 41-52.

Kirkley-Best, E., & Kellner, K. (1982). The forgotten grief: A review of the psychology of stillbirth. *American Journal of Orthopsychiatry*, *52*, 420-429.

Kline, J., Stein, Z., Susser, M., & Warburton, D. (1986). Induced abortion and the chromosomal characteristics of subsequent miscarriages (spontaneous abortions). *American Journal of Epidemiology*, *123*, 1066-1079.

Leon, I. G. (1986). Psychodynamics of perinatal loss. *Psychiatry*, *49*, 312-324.

Lewis, E., & Page, A. (1978). Failure to mourn a stillbirth: An overlooked catastrophe. *British Journal of Medical Psychology*, *51*, 237-241.

Muller, M. E. (1992). A critical review of prenatal attachment research. *Scholarly Inquiry for Nursing Practice*, *6*, 5-22.

Myers, J. K., & Weissman, M. M. (1980). Use of a self-report symptom scale to detect depression in a community sample. *American Journal of Psychiatry*, *137*, 1081-1084.

Neugebauer, R. (1987). The psychiatric effects of miscarriage: Research design and preliminary findings. In B. Cooper (Ed.), *The epidemiology of psychiatric disorders* (pp. 136-149). Baltimore: Johns Hopkins University Press.

Neugebauer, R., Kline, J., O'Connor, P., Shrout, P., Johnson, J., Skodol, A., et al. (1992a). Depressive symptoms in women in the six months after miscarriage. *American Journal of Obstetrics and Gynecology*, *166*, 104-109.

Neugebauer, R., Kline, J., O'Connor, P., Shrout, P., Johnson, J., Skodol, A., et al. (1992b). Determinants of depressive symptoms in the early weeks after miscarriage. *American Journal of Public Health*, *82*, 1332-1339.

Neugebauer, R., Kline, J., Shrout, P., Skodol, A., O'Connor, P., Geller, P. A., et al. (1997). Major depressive disorder in the six months after miscarriage. *Journal of the American Medical Association*, *277*, 383-388.

Nicol, M. T., Tompkins, J. R., Campbell, N. A., & Syme, G. J. (1986). Maternal grieving response after perinatal death. *Medical Journal of Australia*, *144*, 287-289.

Osterweis, M., & Townsend, J. (1988). *Mental health professionals and the bereaved* (DHHS Publication No. ADM 88-1554). Rockville, MD: U.S. Department of Health and Human Services.

Peppers, L. G., & Knapp, R. J. (1980). Maternal reactions to involuntary fetal/infant death. *Psychiatry: Journal for the Study of Interpersonal Processes*, *43*, 155-159.

Phipps, S. (1981). Mourning response and intervention in stillbirth: An alternative genetic counseling approach. *Social Biology*, *28*, 1-13.

Potvin, L., Lasker, J., & Toedter, L. (1989). Measuring grief: A short version of the Perinatal Grief Scale. *Journal of Psychopathology and Behavioral Assessment*, *11*, 29-45.

Prigerson, H. G., Frank, E., Kasl, S. V., Reynolds, C. F., Anderson, B., Zubenko, G. S., et al. (1995). Complicated grief and bereavement-related depression as distinct disorders: Preliminary empirical validation in elderly bereaved spouses. *American Journal of Psychiatry*, *152*, 22-30.

Prigerson, H. G., Maciejewski, P. K., Reynolds, C. F., Bierhals, A. J., Newsom, J. T., Fasiczka, A., et al. (1995). Inventory of Complicated Grief: A scale to measure maladaptive symptoms of loss. *Psychiatry Research*, *59*, 65-79.

Prigerson, H. G., Shear, M. K., Jacobs, S. C., Reynolds, C. F., Maciejewski, P. K., Davidson, J.R.T., et al. (1999). Consensus criteria for traumatic grief: A preliminary empirical test. *British Journal of Psychiatry*, *174*, 67-73.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Rodriguez, J. M., & Kosloski, K. (1998). The impact of acculturation on attitudinal familism in a community of Puerto Rican Americans. *Hispanic Journal of Behavioral Sciences*, *20*, 375-390.

Sanchez-Ayendez, M. (1988). Puerto Rican elderly women: The cultural dimension of social support networks. *Women & Health*, *14*, 239-252.

Shapiro, E. R. (1995). Grief in family and cultural context: Learning from Latino families. *Cultural Diversity and Ethnic Minority Psychology*, *1*, 159-176.

Surtees, P. G. (1995). In the shadow of adversity: The evolution and resolution of anxiety and depressive disorder. *British Journal of Psychiatry*, *166*, 583-594.

Toedter, L. J., Lasker, J. N., & Alhadeff, J. M. (1988). The Perinatal Grief Scale: Development and initial validation. *American Journal of Orthopsychiatry*, *58*, 435-449.

Walsh, F., & McGoldrick, M. (1991). *Living beyond loss: Death in the family*. New York: Norton.

Weissman, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A., & Locke, B. Z. (1977). Assessing depressive symptoms in five psychiatric populations: A validation study. *American Journal of Epidemiology*, *106*, 203-214.

Zeanah, C. H. (1989). Adaptation following perinatal loss: A critical review. *Journal of the American Academy of Child and Adolescent Psychiatry*, *28*, 467-480.

Zeskind, P. S. (1983). Cross-cultural differences in maternal perceptions of cries of low- and high-risk infants. *Child Development*, *54*, 1119-1128.

Zisook, S., Devaul, R. A., & Click, M. A. (1982). Measuring symptoms of grief and bereavement. *American Journal of Psychiatry*, *139*, 1590-1593.

Zuckerman, B., Amaro, H., Bauchner, H., & Cabral, H. (1989). Depressive symptoms during pregnancy: Relationship to poor health behaviors. *American Journal of Obstetrics and Gynecology*, *160*, 1107-1111.

Jennifer Boyd Ritsher, Ph.D., is currently a senior research associate at the Center for Health Care Evaluation, VA Palo Alto Health Care System, and Stanford University Medical Center. Much of the work for the present study was conducted while she was a fellow in the Psychiatric Epidemiology Training Program at Columbia University. Her research examines the influence of sociocultural factors on psychopathology, such as stress as a causal factor, stigma as a maintaining factor, and culture as a contextual factor influencing clinical presentation and the accuracy of assessment.

Richard Neugebauer, Ph.D., M.P.H., is a research scientist in the Epidemiology of Developmental Brain Disorders Department of the New York State Psychiatric Institute and in the Faculty of Medicine, College of Physicians and Surgeons of Columbia University. His research interests and publications have focused on the psychiatric effects of bereavement and trauma, the role of prenatal factors in risk for psychiatric illness in offspring, and most recently, randomized controlled trials of psychological interventions for depression and grief.

The Daily Inventory of Stressful Events

An Interview-Based Approach for Measuring Daily Stressors

David M. Almeida
University of Arizona

Elaine Wethington
Cornell University

Ronald C. Kessler
Harvard University

*This study introduces the Daily Inventory of Stressful Events (DISE), an interview-based approach to the measurement of multiple aspects of daily stressors through daily telephone interviews. Using a U.S. national sample of adults aged 25 to 74 (*N *= 1031), the prevalence as well as the affective and physical correlates of daily stressors are examined. Respondents had at least one daily stressor on 40 percent of the study days and multiple stressors on 11 percent of the study days. The most common class of stressors was interpersonal tension followed by work-related stressors for men and network stressors (events that occur to close others) for women. Stressors that involved danger of loss were more prevalent than stressors in which loss actually occurred. Regression analyses showed that specific types of daily stressors such as interpersonal tensions and network stressors were unique predictors of both health symptoms and mood.*

Keywords: daily hassles, stress, psychological distress, physical symptoms

The aims of the present study are threefold. First, the study describes a new and innovative technique for assessing exposure to daily stressors. This new technique introduces methods for the assessment and classification of daily stressors by type of stressor and the area of life in which stressors occur. Such content of a stressor affects its impact on symptoms (e.g., Stone, 1987). The new method also introduces investigator-rated measures of stressor severity and stressor threat, recommended by several investigators as a way to reduce some of the bias introduced by self-reported ratings of stressor severity and appraisal (e.g., Monroe & Kelley, 1995). Second, the study aims to assess the prevalence of daily stressors in a U.S. national sample of adults, aged 25 to 74. Most studies of the prevalence of daily stressors have used small, volunteer, or local samples, whose characteristics cannot necessarily be generalized to the American population at large. Exposure to daily stressors varies a great deal across different social groups (e.g., Almeida & Kessler, 1998), and failure to document the variety of daily stressors may result in less effective and valid measurement of stressor exposure (Herbert & Cohen, 1996). Third, the study aims to demonstrate the concurrent validity of its measures by examining the extent to which aspects of daily stressors are significantly associated with negative mood and physical health symptoms, two commonly used outcomes in research on the relationship between daily stressors and health.

The research reported in this article was supported by grants from the the MacArthur Foundation Research Network on Successful Midlife Development, the National Institute for Mental Health (MH19734), and the National Institute on Aging (AG19293). The authors wish to thank Amy Chandler, Dan MacDonald, and Melanie Horn for their research assistance.

Assessment, Volume 9, No. 1, March 2002 41-55

Measurement of Daily Stressors

Many daily event measures now in use are based on checklist measures of life events (Turner & Wheaton, 1995) and have strengths and weaknesses similar to those measures. Checklist measures of daily stressors have come under increasing criticism. The focus of the critique is that individual differences in stressor appraisals may at least partially determine the propensity to report exposure to a stressor (Stone, Kessler, & Haythornthwaite, 1991). People are less likely to report nonstressful incidents, even if researchers consider that type of incident to have stress-creating potential. Such incidents may not be relevant to goals, an important component of the appraisal process (Lazarus, 1999). This type of bias is difficult to control, short of experimental manipulation of exposure to stressors. Confounding may also occur when the question includes an appraisal judgment (e.g., "a lot of work at home"). To be measured validly, both stressor exposure and appraisal of stressors must be differentiated from underlying personality traits that affect stressor reactivity (Lazarus, 1999).

Monroe and Kelley (1995) argued for an interview-based, investigator-rated approach (e.g., Brown & Harris, 1978) as a way to differentiate measures of stressor exposure from appraisal. They recommended personal, semistructured interviews to collect information appropriate for investigators to rate the objective content, severity, and threat of events. The information collected is rated for severity and threat using documented dictionary guidelines for each type of life event or stressor (Wethington, Brown, & Kessler, 1995). The threat and severity ratings estimate the amount of emotional arousal likely to result from life events with different types of objective characteristics. Obviously, the diagnosis of inoperable cancer is rated as being objectively much more severe than a diagnosis of a bad cold. The threat to life represented by the cancer diagnosis is hypothesized to create a qualitatively more severe appraisal of threat for the individual reporting it than the diagnosis of a cold, no matter how uncomfortable the cold symptoms. Such investigator-rated methods can be applied to rating the content and severity of daily events. For example, an argument with a child over household chores would be rated as less severe than an argument over damaging the family car.

Almeida (1998) has developed a personal interview assessing the occurrence of daily stressors, as well as their severity and threat, called the Daily Inventory of Stressful Events (DISE). This interview technique uses narrative descriptions of daily stressors reported in telephone interviews for investigator rating of stressor severity and dimension of threat. Dimension of threat and severity ratings are based not on self-report appraised measures (e.g., "How stressful was this for you?") but on objective characteristics of the stressors as reported by participants. The threat dimensions were adapted from the dimensions of contextual threat developed by Brown and Harris (1978), subsequently modified for use in semistructured personal interview surveys (Wethington, 1997; Wethington et al., 1995). These dimensions reflect the revised model of stress, appraisal, and coping proposed by Lazarus (1999). Each daily stressor is rated by the investigators for the type of threat (loss, danger, disappointment, frustration, and opportunity) and the severity of threat ("none" to "extreme") it would pose to the average individual in that situation.

Prevalence of Daily Stressors

The second aim of the present study is to estimate the prevalence of daily stressors in a U.S. national sample of adults. Although the relationship between exposure to stressors and a specific psychological or health outcome can often be estimated accurately in a small and restricted sample, such findings cannot always be generalized to include all groups in the population. It is useful to know the prevalence of daily stressors in the general population because different social groups have been shown to have different rates of exposure to particular types of daily events. Such group differences in exposure may play an important role in the differences in disorder prevalence, such as the gender difference in psychological distress (Almeida & Kessler, 1998).

Previous studies of daily stressors have relied primarily on self-administered checklists that often yielded relatively coarse information, typically whether a stressor occurred. The present study uses telephone interview methodology. Investigators (e.g., Stone et al., 1991) have recommended telephone interviews as the most feasible way to conduct a nationally representative study of daily stressors. Telephone interviews make it possible to interview larger numbers of people about daily stressors at a reasonable cost. They also make it easier to obtain more detailed and accurate information about daily stressors (e.g., timing and duration) through the use of question probes and complex skip patterns. In addition, the gain in greater control over data recording in telephone interviews will also lead to higher response rates and less missing data. Telephone interviews have been shown to have higher response rates than self-administered questionnaires in general population samples (Dillman, 1989). The researcher also has more control over the quality of the interviews (e.g., whether the respondent is paying full attention to the diary completion task, whether diaries are completed every day). Data are recorded more completely in phone interviews than self-administered diaries because the interviewer can ensure that no questions are skipped. Telephone interviews can also enhance the quality of data through probing incomplete or unclear responses. Finally, phone administration permits rapid feedback about nonresponse such as missed phone appointments, making it possible to implement special efforts to complete the interview. For example, interviewers can contact a participant who missed an appointment or convert refusals (cf. Stone et al., 1991).

Daily Stressors, Mood, and Health Symptoms

The final aim of this study is to demonstrate the concurrent validity of the DISE measures by examining their relationship to health symptoms and mood. Promoters of personal interviews acclaim their potentially greater validity through improved assessment of degree of severity, more precise classification of stressor content, and more valid differentiation between severity and stressor appraisal (Brown, 1989). Interviews improve severity assessment by allowing investigators to analyze whole narratives of stressful events rather than brief responses to abstract phrases and by permitting investigators to rate if experiences meet preestablished thresholds of severity and "seriousness" rather than relying on respondent interpretation of the stimulus question.

A second way in which personal interviews may increase concurrent validity is by better measuring content. A trained interviewer can probe to determine if the situation reported in fact matches the intent of the question and should be counted as a stressor of that type (McQuaid et al., 1992; Raphael, Cloitre, & Dohrenwend, 1991). Previous research has found that the area of life in which a daily stressor occurs, as well as its severity, is differentially related to various outcomes (Stone, 1987). Yet relatively few studies (e.g., Almeida & Kessler, 1998; Bolger, DeLongis, Kessler, & Schilling, 1989) have evaluated the differential impact of stressor content on mood and symptoms. The two studies cited in the previous sentence, moreover, were limited by their inability to assess the differential impact of both stressor content and severity.

A third way in which personal interviews may increase the validity of daily stressor assessment is by providing a way to further differentiate two components of the transactional model of stress: stressor severity and appraisal of stressors. Stressor severity should be measured separately and objectively from stressor appraisal to minimize confounding of stressor severity and components of threat with personal dispositions. Previous studies of daily stressors have often relied on self-rated measures of primary and secondary appraisal (Folkman, Lazarus, Dunkel-Schetter, DeLongis, & Gruen, 1986) to measure severity and threat of an event. Although worded as objectively as possible, the items may still be prone to confounding with underlying mood disturbance, which would allow unmeasured individual differences to affect the judgment whether or how much a stressor threatens physical safety or personal health. Interviewers and investigators can eliminate experiences that do not qualify as serious or that are descriptions of symptoms rather than external events per se. Reporting minor events as more severe than they are, moreover, may be related to the respondent's health status at the time of interview (Bebbington, 1986). One innovation of this study was to introduce investigator-rated measures of stressor threat, rather than relying on self-ratings.

In sum, the present article introduces a novel approach for measuring the content, severity, and threat of daily stressors using a national U.S. sample of adults. The goals are to describe the prevalence of exposure to various aspects of daily stressors and to assess the extent to which these measures of daily stressors are associated with physical symptoms and negative mood.

METHOD

Sample

Data for the analyses are from the National Study of Daily Experiences (NSDE). Respondents were 1,031 adults (562 women, 469 men), all of whom had previously participated in the Midlife in the United States Survey (MIDUS), a nationally representative survey of 3,032 people in the age range 25 to 74 carried out in 1995-1996 under the auspices of the John D. and Catherine T. MacArthur Foundation Network on Successful Midlife (Keyes & Ryff, 1998; Lachman & Weaver, 1998; Mroczek & Kolarz, 1998). The MIDUS survey was designed by an interdisciplinary team of 28 researchers to study patterns and correlates of midlife development in the United States with special emphasis on physical health, psychological well-being, and social responsibility. MIDUS respondents were obtained through random digit dialing of telephone numbers. Data collection involved a telephone interview that lasted an average of 30 minutes, as well as mailed questionnaires that were estimated to take an average of an additional 2 hours to complete.

Respondents in the NSDE were randomly selected from the MIDUS sample and received $20 for their participation in the project. Over the course of 8 consecutive evenings, respondents completed short telephone interviews about their daily experiences. On the final evening of interviewing, respondents also answered several questions about their previous week. The interviews took approximately 10 to 15 minutes to complete. Data collection spanned an entire year (March 1996 to April 1997) and consisted of 40 separate "flights" of interviews, with each flight representing the 8-day sequence of interviews from approximately 38 respondents. The initiation of interview flights was staggered across the day of the week to control for the possible confounding between day of study and day of week. Of the 1,242 MIDUS respondents we attempted to contact, 1,031 agreed to participate, yielding a response rate of 83%. Respondents completed an average of seven of the eight interviews, resulting in a total of 7,221 daily interviews.

Table 1 compares characteristics of the NSDE respondents with the MIDUS respondents who did not participate in the NSDE. The two samples had very similar distributions across these demographic characteristics. The NSDE had slightly more female and fewer minority respondents than the MIDUS sample. Respondents for the present analysis were on average 47 years old. Seventy-seven percent of the women and 85% of the men were married at the time of the study. Thirty-eight percent of the households reported having at least one child in the household. The average family income was between $50,000 and $55,000. Men were slightly older than women and had similar levels of education.

TABLE 1
Demographic Comparison of the MIDUS Sample and the NSDE Subsample (in percentages)

Demographic Variable	*MIDUS*[a]	*NSDE*[b]
Age		
Young adults (25-39)	32.4	33.5
Midlife adults (40-59)	45.6	45.0
Older adults (60-74)	22.0	21.5
Gender		
Males	47.6	45.5
Females	52.4	54.5
Education		
12 years or less	38.4	37.7
13 years or more	61.6	62.3
Marital status		
Married	61.2	65.4
All others	38.8	34.6
Children in household[c]		
Yes	39.3	37.8
No	60.7	62.2
Race		
Caucasian	86.9	90.3
African American	7.2	5.9
All other races	5.9	3.8

NOTE: MIDUS = Midlife in the United States Survey; NSDE = National Study of Daily Experiences.
a. Respondents in the MIDUS survey who did not participate in the NSDE daily study (N = 2,001).
b. Respondents in the NSDE study, all of whom had previously participated in the MIDUS survey (N = 1,031).
c. Whether respondent has at least one child age 18 or younger living in the house.

Measures

Daily negative mood. The telephone diary included an inventory of 10 emotions from the Negative Affect Scale designed specifically for the MIDUS survey (Mroczek & Kolarz, 1998). This scale was developed from the following well-known and valid instruments: the Affect Balance Scale (Bradburn, 1969), the University of Michigan's Composite International Diagnostic Interview (Kessler et al., 1994), the Manifest Anxiety Scale (Taylor, 1953), and the Center for Epidemiological Studies Depression Scale (Radloff, 1977). Examples of items include sad, hopeless, anxious, and restless. Each day, the respondents indicated how much of the time they experienced each emotion over the past 24 hours on a 5-point scale from *none of the time* to *all of the time*. Mean scores across the 10 items were computed. Cronbach's alpha for the scale was .89.

Daily physical symptoms. These were measured using a shortened version of Larsen and Kasimatis's (1991) physical symptom checklist. Items that overlapped with the Psychological Distress Scale (e.g., "urge to cry") were omitted. Our 5-item scale assessed aches (headaches, backaches, and muscle soreness), gastrointestinal symptoms (poor appetite, nausea/upset stomach, constipation/diarrhea), upper respiratory symptoms (sore throat, runny nose), and other physical symptoms or discomforts. Open-ended responses to the other physical symptoms question were subsequently coded and placed into an existing category, deleted if symptom was psychological, or left in a miscellaneous category if no other category existed. Each day, the respondents indicated how frequently they experienced each symptom over the past 24 hours on a 5-point scale from *none of the time* to *all of the time*. Mean scores across the five items were computed. Cronbach's alpha for the scale was .71.

Daily stressors. These were assessed through a semi-structured Daily Inventory of Stress Events (DISE) (Almeida, 1998). The inventory consisted of a series of stem questions asking whether certain types of daily stressors had occurred in the past 24 hours, along with a set of interviewer guidelines for probing affirmative responses to rate stressor content, severity, and threat as well as a series of structured questions that measured respondents' primary appraisal of the stressors. The stem questions, examples of the probe questions, and appraisal questions are provided in the appendix. The aim of this interviewing technique was to acquire a short narrative of each stressor that included descriptive information (e.g., topic or content of the stress, who was involved, how long the stressor lasted) as well as what was at stake for the respondent. Open-ended information for each reported stressor was tape recorded then transcribed and coded for several characteristics. Coders were graduate and advanced undergraduate students who received approximately 10 hours of initial training. Subsequent 2-hour weekly meetings were held to check accuracy and discuss discrepant ratings. As new coders joined the project, they were required to demonstrate interrater reliability similar to other coders. This interview-based approach allowed us to distinguish between a stressful event (e.g., conflict with spouse) and the affective response to the stressor (e.g., crying or feeling sad). Another benefit of this approach was its ability to identify overlapping reports of stressors. In the present study, approximately 5% of the reported stressors were discarded because they were either solely affective responses or they were identical to stressors that were previously described on that day.

Table 2 presents the description and interrater reliability of the DISE measures. For each stressor, expert coders rated (a) content classification of the stressor (e.g., work overload, argument over housework, traffic problem), (b) focus of who was involved in event, (c) dimensions of environmental threat (loss, danger, disappointment, frustration, opportunity), and (d) severity of stress. In addition, respondents provided reports of (e) degree of severity and (f) primary appraisal (i.e., areas of life that were at risk because of the stressor).

The first two measures in Table 2 pertain to the objective nature of the stressor. Each stressor was initially placed into a *content classification*. This taxonomy of daily stressors combined the broad type of stressors (e.g., an argument) with specific content or topic (e.g., housework). A pilot study of a national sample of 1,006 adults was initially conducted to generate the content classification list of daily stressors common to adults in the United States. The initial list included eight broad classifications and 39 specific classifications. This list was then lengthened to incorporate 10 additional specific classifications of arguments and tensions and five other miscellaneous classifications. *Focus of involvement* assessed whether other individuals were involved in the stressor and, if so, what their relations were to the respondent (Brown & Harris, 1978). Codes assigned to content and focus were derived directly from respondent self-reported descriptions. They closely parallel individual items in life event and daily hassles checklists (e.g., Zautra, Guarnaccia, & Dohrenwend, 1986).

The remaining measures in Table 2 assessed the meaning of the stressor for the respondent. *Environmental threat dimensions* were the investigator-rated stressful implications for the respondent. These dimensions were similar to Lazarus's (1999) dimensions of environmental threat, with the addition of disappointment, an expected positive experience that did not occur (Wheaton, 1999), and frustration (i.e., stressors in which the respondent has little or no control). *Objective severity* ratings were similar to Brown and Harris's (1978) ratings of short-term contextual threat and were based on the degree of disruptiveness and unpleasantness associated with the stressor. The final two DISE measures were obtained from the respondents' own ratings (see the appendix for the items). These included the perceived or *subjective severity* of the stressor and self-reports on seven *domains of primary appraisal*, (i.e., the degree of risk the stressor posed to goals and values at stake in the stressful encounter, including self-esteem, life goals, and other persons' well-being) (Lazarus, 1999). Approximately 20% (800 events) of the stressors were rated by two coders. The interrater reliability for investigator ratings (Kappa) ranged from .66 to .95 across all of the codes. Specific classification codes had the lowest reliability partially because of the high number of possible codes (n = 54).

The documentation and guidelines for all of these ratings are provided in an interview and coding manual (Almeida, 1998). In addition, all of the transcribed descriptions of daily stressors and their corresponding ratings are contained in an "electronic dictionary" stored on a computer spreadsheet. This dictionary consists of more than 4,000 rated daily stressors and can be searched and cross-referenced by any of the DISE measures.

RESULTS

The initial set of analyses examined how often respondents experienced daily stressors as well as their average level of daily physical health symptoms and daily negative mood. Across the 8 study days, we calculated the percentage of days that respondents reported any daily stressors (i.e., an affirmative response to any of the stressor stem questions) and multiple daily stressors (i.e., an affirmative response to two or more of the stressor stem questions). The daily symptoms and daily distress scores were also averaged across the study days. Table 3 summarizes these results. On average, respondents reported experiencing at least one stressful event on 39.4% of the study days and multiple stressful events on 10.4% of the study days. Women had more frequent days in which they reported one stressful event, $t(1030) = 5.1, p < .01$, but both men and women had similar numbers of days involving multiple stressors. The scores for daily physical symptoms and daily negative mood represent the daily level of these variables averaged across the study days. Women reported higher levels of physical symptoms, $t(1030) = 3.9, p < .01$.

TABLE 2
Description of the Daily Inventory of Stressful Events Measures

Coding Category	*Description*	*Codes*
Content classification	Stressful events are categorized into one of seven broad classifications organized by interpersonal tensions, life domains, network events, and miscellaneous events. Next, they are placed in 1 of 54 specific classifications. Broad classifications are listed in the cell to the right, followed by the number of specific classifications associated with each heading.	Interpersonal tensions (21) Work/education (9) Home (9) Finances (3) Health/accident (5) Network (7) Miscellaneous (9)
Focus of involvement	Focus of involvement refers to who was involved in the event.	Respondent Other Joint
Threat dimensions	The threat dimension describes the implications of the event for the respondent. Loss is the occurrence of a deficit. Danger is the risk of a future negative occurrence. Disappointment occurs when something does not turn out as the respondent had expected. Frustration occurs when the respondent has little or no control over the events. Opportunity is a chance for positive outcome.	Loss Danger Disappointment Frustration Opportunity
Objective severity	The objective assessment of the severity of an event refers to the degree and duration of disruption and/or unpleasantness created for the respondent. Ratings range from 1, *a minor or trivial annoyance*, to 4, *a severely disruptive event*.	Low severity events Medium severity events High severity events Extreme severity events
Subjective severity	The subjective assessment of severity is the respondent's assessment of the degree of stressfulness involved in the event.	Not at all stressful Not very stressful Somewhat stressful Very stressful
Primary appraisal domains	Primary appraisal domains refer to the respondent's report of how much the following areas were at risk or at stake in the situation: (a) daily routine disruption, (b) finances, (c) how respondent feels about self, (d) how others feel about respondent, (e) health or safety, (f) well-being of one close to respondent, and (g) future plans.	Not at all A little Some A lot

TABLE 3
Description of Daily Stressors, Physical Symptoms, and Negative Mood

	Total		*Men*		*Women*	
	M	SD	M	SD	M	SD
Any stressors (% days)	39.4	26.8	37.5	27.3	40.9*	26.4
Multiple stressors (% days)	10.4	16.7	9.4	16.6	11.2	16.6
Physical symptoms (level)[a]	1.38	0.27	1.34	0.27	1.41*	0.39
Negative mood (level)[a]	1.19	0.30	1.18	0.29	1.20	0.30

NOTE: *N* = 1,031.
a. Range is 0 (*none of the time*) to 5 (*all of the time*).
*$p < .05$, gender difference.

Content and Focus of Daily Stressors

Using measures from the DISE, the next set of analyses provided a much more detailed taxonomy of daily stressors. Table 4 shows the percentage of study days that respondents reported each type of stressor as well as the relative prevalence (i.e., the proportion of reported stressors that fell into each of the categories). Broad content classifications are shown in bold followed by their corresponding specific classifications. Interpersonal arguments and tensions composed the most common broad class of daily stressors occurring on 22% of the study days (i.e., 1.5 days per week). This class of stressor accounted for approximately one half of all reported daily stressors. The most prevalent types of tensions were unspecified dis-

TABLE 4
Distribution of Stressor Content Classification and Focus of Involvement (in percentages)

	Total		*Men*		*Women*	
	Days	*Events*	*Days*	*Events*	*Days*	*Events*
Arguments/tensions	**22.6**	**50.0**	**21.4**	**49.1**	**23.6**	**50.3**
General disagreement[a]	3.8	8.6	3.6	8.7	4.0	8.5
Job procedures	3.7	7.8	4.5	9.6	3.1	6.3*
Financial issues	2.4	4.6	2.2	4.7	2.5	4.6
Discipline/correct child	1.9	3.5	1.4	2.8	2.3	4.1
Timing/schedules	1.6	3.1	1.4	2.7	1.7	3.4
Value differences	1.4	2.7	1.8	3.2	1.2	2.3
Household chores[b]	1.4	2.6	0.9	1.8	1.9*	3.3*
Family issues	1.5	3.1	1.2	2.6	1.7	3.4
Respect/disrespect[b]	1.1	1.9	0.7	1.2	1.5*	2.5*
Personal tastes[b]	1.0	1.8	0.7	1.4	1.2	2.1
Miscommunication	0.7	1.8	1.0	2.6	0.5	1.2
Transportation	0.6	1.4	0.5	1.6	0.6	1.1
Disciplining employee[b]	0.7	1.1	1.1	1.6	0.4	0.7*
Safety/health[b]	0.6	1.1	0.6	1.2	0.7	0.9
Interaction with boss	0.7	1.0	0.6	0.8	0.8	1.2
Substance use[b]	0.3	0.8	0.3	0.5	0.4	0.9
Recreational activities[b]	0.3	0.7	0.2	0.4	0.4	1.0
Possessions[b]	0.3	0.7	0.2	0.4	0.4	0.9
Schoolwork[b]	0.3	0.6	0.1	0.3	0.5	0.8
Receiving bad news[b]	0.3	0.5	0.3	0.4	0.3	0.6
Sex[b]	0.2	0.3	0.3	0.5	0.1	0.1
Work/school	**8.5**	**13.2**	**9.1**	**15.7**	**8.0**	**11.2**
Work overload/demand	3.8	6.7	3.6	6.9	3.9	6.6
Technical breakdown	0.6	1.4	1.1	2.3	0.3*	0.6*
Mistakes	0.6	1.2	1.0	2.0	0.4*	0.6*
Job security	0.5	0.8	0.7	1.2	0.3	0.5
Time/schedules	0.5	0.7	0.4	0.5	0.6	0.9
Job structure[b]	0.4	0.7	0.5	1.2	0.3	0.3
Other work events	0.5	0.7	0.4	0.6	0.3	0.7
School overload/demand	0.3	0.6	0.3	0.7	0.3	0.6
Starting job	0.2	0.4	0.2	0.3	0.2	0.4
Home	**5.6**	**8.2**	**5.3**	**8.0**	**5.9**	**8.3**
Overload/demand	1.7	2.7	1.1	2.2	2.2*	3.0
Household/car repairs	1.4	2.6	1.5	2.9	1.4	2.3
Financial problems	0.9	1.3	1.3	1.7	0.6	0.9*
Pet event	0.4	0.9	0.2	0.6	0.6	1.1
Mistakes	0.4	0.8	0.3	0.7	0.4	0.8
Moving	0.2	0.4	0.1	0.4	0.2	0.4
Other home events	0.2	0.4	0.2	0.5	0.1	0.2
Neighborhood concerns	0.2	0.3	0.3	0.3	0.2	0.3
Time/schedules[b]	0.2	0.2	0.2	0.2	0.2	0.2
Purchase/sale	0.2	0.6	0.4	1.0	0.1	0.2
Health care	**1.0**	**2.2**	**0.9**	**1.6**	**1.2**	**2.7**
Accident/illness	0.7	1.3	0.6	1.1	0.7	1.6
Visit/contact	0.4	0.8	0.2	0.5	0.6	1.1
Network (events that happen to others)	**8.0**	**15.4**	**6.1**	**12.5**	**9.6***	**17.8**
Health	3.8	7.0	3.0	5.9	4.5*	8.0
Social concerns	3.5	6.4	2.7	5.0	4.2*	7.5*
Death/bereavement	0.7	1.7	0.5	1.3	0.9	2.0
Finances[b]	0.2	0.3	0.1	0.2	0.2	0.3
Miscellaneous	**1.7**	**3.5**	**1.8**	**4.4**	**1.5**	**2.7**
Traffic/transportation	0.8	1.7	1.7	2.1	0.8	1.3
Weather	0.4	0.9	0.5	1.1	0.4	0.7
Mistakes	0.2	0.4	0.2	0.4	0.1	0.4
News	0.1	0.3	0.1	0.4	0.1	0.2
Public speaking[b]	0.2	0.2	0.2	0.4	0.1	0.1
Focus categories						
Respondent	12.7	25.7	13.1	29.4	12.3	22.8*
Other	5.3	10.8	4.3	9.2	6.1	12.2*
Joint	26.5	63.3	24.4	61.5	28.2	64.8

NOTE: $N = 1{,}031$. Distribution of broad classifications are shown in bold.
a. Tense moments with others, topic not specified.
b. Event not identified in pilot study.
*$p < .01$, significant gender difference.

agreements, job procedures, financial issues, and disciplining children.

A series of independent *t* tests assessed gender differences in the stressor content classifications. Because of the large number of comparisons, a more stringent alpha level was used ($p < .01$) to reduce the risk of chance findings. Men reported a higher proportion of tensions involving job procedures and disciplining employees, whereas women reported more tensions regarding household work and being respected. For men, the second most common broad class of stressors was that associated with paid work such as technical breakdowns and mistakes. These work stressors did not involve interpersonal tensions. The second most common class of stressors for women was network stressors—stressors that happened to a network of relatives and close friends. Women reported these network stressors on 50% more of the study days than did men. Home-related stressors composed the third most prevalent class of stressors for both men and women. However, within this broad classification, women were more likely to report overloads and men were more likely to report financial stressors.

The final category in Table 4 presents the daily frequency and proportion for the focus of involvement of daily stressors. More than 60% of the daily stressors were joint focused, involving the respondent and another person. The next most common focus of stressors was self-focused, involving the respondent only, followed by other focused, involving only someone else. When stressors included other individuals (i.e., joint- or other- focused stressors), they most likely involved a spouse or partner. The gender differences in the content classification are mirrored in the measure of focus of involvement. Men were more likely to experience self-focused stressors, whereas women's stressors were more focused on other people.

Environmental Threat, Severity, and Primary Appraisal of Stressors

As part of the DISE interview, respondents answered a series of structured and semistructured questions that pertained to the dimension and degree of environmental threat as well as their own appraisals of the stressors. These included investigator-rated dimensions of threat, the investigator-rated (objective) and self-reported (subjective) degree of stressor severity, and primary appraisal domains (i.e., goals and values that were at risk because of the stressor). Table 5 provides a summary of these ratings of daily stressors broken down by gender. The figures for the threat dimensions reflect the percentage of stressors that fell into each of five threat categories. Of the stressors that the total sample experienced, roughly 30% involved some sort of loss, nearly 37% posed danger, and 27% were frustrating or out of the control of the respondent. Table 5 also presents the level of severity ratings averaged across all of the daily stressors. On average, the respondents subjectively rated stressors as having medium severity, whereas objective coders rated the stressors as posing low severity. Figures for the domains of primary appraisal represent the amount of risk stressors imposed on seven goals and values. Daily stressors posed the most risk to disrupting the respondent's daily routine.

Gender differences were observed in subjective ratings of stressor severity in two of the domains of primary appraisal. On average, women subjectively rated stressors as more severe than did men, $t(1030) = 5.4$, $p < .01$. Interestingly, there were no significant gender differences in the objective investigator ratings of stressor severity. This suggests that one's gender does not necessarily expose one to stressors that are inherently more severe, at least according to our trained coders. Compared to women, men reported that stressors posed more risk to their financial situations, $t(1030) = 3.4$, $p < .01$, and less risk to how other individuals felt about them, $t(1030) = 3.9$, $p < .01$.

TABLE 5
Description of Daily Inventory of Stressful Events Measures of Stressor Threat, Severity, and Primary Appraisal

	Total	*Men*	*Women*
Threat dimensions (% events)			
Loss	29.7	29.9	29.5
Danger	36.2	35.7	36.6
Disappointment	4.2	4.0	4.4
Frustration	27.4	28.3	26.6
Opportunity	2.3	2.1	2.4
Stressor severity (*M*)[a]			
Objective assessment	1.8	1.7	1.9
Subjective assessment	2.7	2.5	2.9*
Primary appraisal domains (*M*)[b]			
Disrupting daily routine	2.3	2.3	2.3
Financial situation	1.3	1.4	1.2*
Way feel about self	1.5	1.4	1.5
Way others feel about you	1.4	1.3	1.4*
Physical health or safety	1.3	1.3	1.3
Health/well-being of someone you care about	1.5	1.5	1.5
Plans for the future	1.4	1.4	1.3

NOTE: $N = 1,031$.
a. Range is from 1 (*not at all stressful*) to 4 (*very stressful*).
b. Range is from 1 (*no risk*) to 4 (*a lot of risk*).
*$p < .01$, significant gender difference.

Table 6 shows the relationships among the DISE threat, severity, and primary appraisal measures. The first five columns list the average levels of the severity and primary appraisal broken down by the dimensions of threat. Results of a series of one-way ANOVAs with Tukey multiple comparison tests revealed that stressors that involved danger and loss were subjectively and objectively rated as more severe than stressors associated with other threat dimensions, $F(5, 1025) = 4.2$, $p < .01$, and $F(5, 1025) = 3.8$, $p < .01$, respectively. Stressors associated with danger and loss were also more likely to pose the greatest risk to respondent's physical health and safety, $F(5, 1025) = 6.8$, $p < .01$.

Table 6 also shows the intercorrelations between the severity and the primary appraisal measures. If these measures tap different aspects of daily stressors, we would expect a low degree of correlation between them. The overall pattern of correlations did indicate a modest degree of independence between the severity and appraisal domains as well as within the appraisal domains. Only 3 of the 36 correlations were above .30. Not surprisingly, coders' objective ratings of severity were associated with the respondents' subjective ratings of severity. Within the primary appraisal domains, stressors that risked the way respondents felt about themselves tended also to involve risk for the way others felt about them. Stressors that posed risk to respondents' financial plans were related to respondents' plans for the future.

Daily Stressors, Physical Symptoms, and Negative Mood

In the final set of analyses, zero-order correlations and hierarchical multiple regressions assessed the associations of the daily stressor variables with measures of physical symptoms and negative mood. Although physical symptoms and mood were moderately correlated, $r = .34$, $p <$

TABLE 6
Interrelations Among the Daily Inventory of Stressful Events Measures of Stressor Threat, Severity, and Primary Appraisal

	Threat Dimension					*Intercorrelation*							
	Loss	*Danger*	*Disappoint-ment*	*Frustration*	*Challenge*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Stressor severity													
1. Objective severity	2.04	2.03	1.79	1.36	1.84*								
2. Subjective severity	2.96	2.95	2.59	2.53	2.37*	.36*							
Primary appraisal domains													
3. Disrupting daily routine	2.58	2.31	2.27	2.13	2.31	.16*	.27*						
4. Financial situation	1.39	1.42	1.32	1.19	1.31	.21*	.17*	.20*					
5. Way feel about self	1.52	1.52	1.54	1.51	1.50	–.02	.14*	.21*	.12*				
6. Way others feel about you	1.46	1.44	1.53	1.48	1.47	–.04	.22*	.22*	.15*	.36*			
7. Physical health or safety	1.37	1.39	1.21	1.26	1.29*	.13*	.23*	.27*	.24*	.12*	.24*		
8. Health/well-being of someone you care about	1.38	1.74	1.36	1.35	1.34	.21*	.07	.05	.01	.01	.06	.15*	
9. Plans for the future	1.42	1.49	1.36	1.29	1.40	.23*	.22*	.21*	.47*	.14*	.24*	.15*	.14*

NOTE: N = 1,031.
*p < .01.

.01, we wanted to examine if the DISE measures demonstrated differential prediction to symptoms versus mood. The order of entry of the hierarchical regression was as follows: On the first two steps, the objective measures of daily stressors were entered. On Steps 3 to 5, the severity, threat, and appraisal measures of daily stressors were entered. This strategy allowed us to examine the unique prediction of each daily stressor variable holding the other variables within each class of measure constant. In addition, we were able to test whether the severity, threat, and appraisal measures accounted for variance in physical symptoms and negative mood beyond the content and focus measures of stressors.

Table 7 shows the results from these analyses. Seventeen of the daily stressor variables were significantly correlated with physical health symptoms, and 16 daily stressor variables were significantly correlated with negative mood. Results from the regression analyses indicate that 10 of these variables made unique predictions to both daily symptoms and negative mood. The largest coefficients suggested that higher levels of daily physical symptoms and negative mood were associated with more frequent interpersonal, network, and respondent-focused stressors as well as with stressors that were rated as severe and appraised as posing risk to physical health and safety. The R^2 change coefficients indicated that each type of threat and appraisal measure accounted for variance above the content and focus measures. The entire set of stressor variables accounted for 17% of the variance in daily physical symptoms and 31% of the variance in daily negative mood.

Significant gender differences in the associations of the daily stressor measures with daily symptoms and negative mood were assessed by computing correlations for men and women separately and calculating the Fisher's R to z transformation. The association between daily physical health symptoms and frequency of network stressors was greater for women ($r = .20$) than for men ($r = .05$, $z = 2.50$). The association of daily negative mood with interpersonal tensions was also greater for women ($r = .35$) than for men ($r = .22$, $z = 2.00$). However, the association between daily negative mood and work stressors was greater for men ($r = .19$) than for women ($r = .04$, $z = 2.00$). There were no significant gender differences in the association of daily symptoms and negative mood with the measures of stressor threat and appraisal.

DISCUSSION

The Benefits of the Investigator-Based Approach

This study introduces an investigator-based approach to the measurement of multiple aspects of daily stressors through daily telephone interviews that utilize a series of structured and semistructured questions. A relatively small number of stem questions were designed to cover a wide range of stressor content domains such as family, work, and interpersonal tension. Affirmative answers to the stem questions were then probed by trained interviewers, who sought to establish if an objective stressful occurrence had actually happened and to collect specific details that could be used to classify the stressor by content, severity, and dimensions of threat. Trained coders then classified the stressors, producing detailed documentation for their ratings. These rating "dictionaries" allowed this study to pro-

TABLE 7
Correlations and Hierarchical Multiple Regressions of Daily Inventory of Stressful Events Measures With Daily Physical Symptoms and Daily Negative Mood

	Physical Symptoms				*Negative Mood*			
	r	β	R^2 *Change*	R^2	r	β	R^2 *Change*	R^2
Step 1: Broad content classification								
Interpersonal tensions	.23*	.20*			.31*	.29*		
Work stressors	.07*	–.01			.06	–.02		
Home stressors	.11*	.06			.12*	.06*		
Network stressors	.16*	.11*			.14*	.08*		
Miscellaneous stressors	.06	.05	.07*		.02	.01	.10*	
Step 2: Focus of involvement								
Respondent	.15*	.23*			.14*	.24*		
Other	.14*	.04			.06	–.13		
Joint	.22*	.03	.01	.08*	.30*	.03	.02*	.12*
Step 3: Threat dimensions								
Loss	.21*	.08*			.24*	.11*		
Danger	.23*	.08*			.27*	.13*		
Disappointment	.04	–.01			.03	–.03		
Frustration	.08*	.04			.07*	–.07		
Opportunity	.03	–.03	.02*	.10*	–.01	–.08*	.03*	.15*
Step 4: Stressor severity								
Objective assessment	.18*	.09*			.20*	.09*		
Subjective assessment	.23*	.15*	.03*	.13*	.32*	.22*	.08*	.23*
Step 5: Primary appraisal domains								
Disrupting daily routine	.16*	.02			.25*	.03		
Financial situation	.10*	.02			.16*	.02		
Way feel about self	.20*	.09*			.29*	.08*		
Way others feel about you	.05	–.06*			.22*	.07		
Physical health or safety	.28*	.20*			.32*	.15*		
Health/well-being of someone you care about	.00	–.06			.02	–.06		
Plans for the future	.24*	.04	.04*	.17*	.22*	.09*	.08*	.31*

N = 1,031 respondents.
*p < .01.

duce ratings of sufficient detail and consistency, enabling future researchers to replicate our findings.

The interview-based approach provided us several advantages in the present study. The first was that the instrument produced a comprehensive accounting of daily stressors with a minimal number of questions. Such a strategy reduces the burden on study participants, who for the most part need to devote only about 15 minutes a day to the interview. In contrast, comprehensive daily stressor checklists (e.g., Brantley & Jones, 1993; Zautra et al., 1986) involve checking off a much larger number of questions, which requires more time.

The second advantage of the investigator-based approach was that the open-ended narratives allowed us to classify stressors with more precision than a conventional checklist. The narratives made it possible to classify stressors by the area of life affected (i.e., content) and by who was involved (i.e., participant focused, other focused, or both). Both content and focus are known to be related to the health impact of the stressor (Kessler & McLeod, 1984; Stone, 1987). Intercoder reliability estimates suggest that coders were able to make such distinctions.

The third advantage of the investigator-based approach was the reduction of some types of response bias to particular stressor questions. By having content classification of the stressor performed by the coding team rather than participants, it was possible to classify responses more reliably and validly. The coders were able to detect and eliminate incidents reported by participants that did not meet criteria set for objective stressors. This strategy prevents investigators from inadvertently classifying an episode of mood disturbance or ill health as a putative "cause" of bad mood or ill health that day.

A fourth advantage was the investigator-based system of rating severity. Self-report methods of assessing stressor severity are likely confounded with mood at the time of recall, usually the end of the day. The investigator system of probing for objective details makes it possible to rate severity by objective criteria. It is important to note that investigator-rated severity tended to be lower than severity estimated by participants, suggesting that self-rated severity does indeed tend to misspecify the relationship between stressor exposure and health or mood.

Finally, the investigator-based approach allowed us to clarify and disentangle the classification of stressors and dimensions of threat from the self-reported appraisal of the stressor. Investigator ratings of severity and self-reported domains of primary appraisal were only modestly correlated. The differentiation of severity, investigator-rated dimensions of threat, and self-reported primary appraisal of the stressful encounter could promote useful advances in testing the components of the revised transactional model of stress (Lazarus, 1999).

Daily Stressors in the United States

This study was the first of its kind to conduct daily telephone interviews on a national sample of adults. Response and retention rates were above 80%, both of which are substantially higher than other diary studies that rely on self-report questionnaires (cf. Almeida & Kessler, 1998). In addition, the demographic distribution of the current sample is almost identical to the national sample from which it was drawn. The telephone interviews also allowed much more control over when interviews were completed as well as clarification of potentially ambiguous questions. For these reasons, the current study limits sources of sampling error that could bias estimates of the prevalence of daily stressors. Respondents had at least one daily stressor on 40% of the study days and multiple stressors on 11% of the study days. These findings indicate that adults experience some form of daily stressor on 12 days within an average month.

It is important to note that these estimates are somewhat lower than checklist studies (e.g., Bolger, DeLongis, Kessler, & Schilling, 1989; Stone & Neale, 1984). Three potential reasons help explain this discrepancy. First, the DISE measure has objective criteria for what constitutes a stressor. Affective states such as feeling sad or crying do not meet these criteria. In addition, each stressor must be independent from other stressors. In checklist measures, the researcher has little control over how respondents interpret questions, which may contribute to overestimates of stressor prevalence. Second, checklist measures often include more than 20 and can have more than 100 items (for a review, see Eckenrode & Bolger, 1995). Although the large number of stressor items expands the comprehensiveness of the list of stressors, more items increase opportunities to double report the same stressor (Schwarz, 1999). On the other hand, the DISE measure has only seven stem questions that may limit the opportunity for a respondent to report a stressor. The DISE addresses this issue of comprehensiveness by including a final stem question that allows respondents to report any stressors not previously indicated. Respondents rarely reported these miscellaneous stressors when they reported other stressors (1.7% of days). Third, a possible explanation of lower prevalence is falloff in reporting over the week. Because the DISE relies on follow-up probes, it is possible that respondents "learned" to under- report stressors on subsequent daily interviews to avoid answering open-ended questions. We examined this explanation by testing if respondents reported fewer stressors on later study days using linear and quadratic forms of a variable defining the number of days that had elapsed since the respondent first began filling out the diary. Although there was some evidence for falloff of reporting, it was not as great as that found in self-report checklist studies.[1]

A major benefit of the investigator-based measurement of daily stressors is its ability to obtain detailed information regarding both objective characteristics of daily stressors as well as the multiple aspects of appraised meaning of stressors. These characteristics include level of severity and dimensions of environmental threat, as well as self-reported appraisals of the goals and values at stake. Such data provide a more complete picture of the types of stressors individuals experience as well as the implications of the stressors for individuals than are typically measured by more standard designs. In our U.S. sample of adults, the most common class of stressors was interpersonal tension followed by work-related stressors for men and network stressors for women. Such gender differences are consistent with prior research on life events and daily stressors (Almeida & Kessler, 1998; Bolger, DeLongis, Kessler, & Wethington, 1989). In terms of the investigator ratings of environmental threat, it appears that stressors that involve danger of loss are actually more prevalent than stressors in which loss actually occurs. In addition, stressors were most frequently appraised as disrupting daily routines, according to respondent self-report. This may indicate that adults anticipate future stressors arising out of daily life and actively strive to prevent them.

Daily Stressors, Mood, and Physical Symptoms

The final aim of the study was to demonstrate the concurrent validity of the DISE measures by examining their relationship to health symptoms and mood. Some life event researchers (e.g., Brown, 1989) have argued that interview measures are more valid than checklist measures of life events. Their assertions suggest that the use of personal interview techniques should introduce major improvements to the measurement of daily stressors through more valid assessment of severity and content and more valid differentiation among severity, content, and stressor appraisal. Whereas most other studies have found consistent links between the frequency of daily stressors with negative mood and physical symptoms (Almeida & Kessler, 1998; Bolger, DeLongis, Kessler, & Schilling, 1989; Larson & Kasimatis, 1991; Stone, Reed, & Neale, 1987), the current multidimensional measure helps address under what conditions and how daily stressors are associated with these health outcomes.

Of the 22 stressor variables we examined, 17 were associated with health symptoms and 16 were associated with negative mood. The stressor content and focus variables alone accounted for 8% of the variance in physical symptoms and 12% of the variance in negative mood. These effect sizes are consistent with previous self-report checklist studies on the frequency of daily stressors (Almeida & Kessler, 1998; Rehm, 1978; Stone, 1987). The regression analyses also showed that specific types of daily stressors such as interpersonal and network stressors (events that occur to close others) were unique predictors of both health symptoms and mood. This pattern of findings is similar to previous research assessing the content of daily stressors (Bolger, DeLonger, Kessler, & Schilling, 1989; Stone, 1987).

Our analyses highlighted the importance of both investigator-rated threat and self-rated primary appraisal measures in predicting health symptoms and mood. Each type of measure (i.e., investigator-rated dimension of threat, objective and subjective level of severity, and self-rated domain of primary appraisal) contributed to the explained variance in these outcomes above and beyond content classification and focus. Individuals who had a greater proportion of stressors that pose high severity, loss, or danger reported more symptoms and higher negative mood. In addition, stressors appraised as disrupting daily routines or posing risk to physical health and safety were also shown to be unique predictors of symptoms and mood.

Gender Differences

The present study also observed gender differences in several aspects of daily stressors. Consistent with other studies, women reported more frequent daily stressors and higher levels of negative mood than did men (Almeida & Kessler, 1998; Bolger, DeLongis, Kessler, & Wethington, 1990). In terms of the objective characteristics of daily stressors, men had more arguments with coworkers about job procedures, miscommunication, and discipline, whereas women reported more arguments over household chores and getting respect. Similarly, men reported proportionally more self-focused stressors involving financial and work-related problems, whereas women's stressors were more likely to be focused on concerns for their network of close friends and relatives. These findings are consistent with previous studies of stress spillover and crossover. Women's roles may expose them to more stressors experienced by significant others (Kessler & McLeod, 1984). Traditional family roles, moreover, obligate women to provide more services to family members, which may increase demands from others (Wethington, McLeod, & Kessler, 1987). A previous daily diary study found wives more likely than husbands to react to the stressors of their partners (Bolger, DeLongis, Kessler, & Wethington, 1989).

Gender differences were also observed in subjective ratings of stressor severity. Women subjectively rated stressors as more severe than did men. Interestingly, there were no significant gender differences in the objective investigator ratings of stressor severity. This suggests that one's gender does not necessarily increase exposure to stressors that are inherently more severe, at least according to our trained coders. However, the *perception* of stressor severity may be gender graded. This may be due to the possibility either that men downplay the significance of stressors or that women perceive events as relatively more dramatic than do men. Recall that women reported higher levels of negative mood than did men. It is possible that women are more likely to report stressors as more severe because of their average higher negative affect at the time of recall (Fiedler & Stroehm, 1986).

The final set of analyses addressed whether there were differences in reactivity to the content of the stressors, their threat, and their appraised meanings. Although there was some evidence for differential reactivity to the content of daily stressors, there was no evidence for differential reactivity to threat or stressor appraisal. Some research suggests that women may be more reactive to stressors that occur to their network of friends and relatives (Kessler & McLeod, 1984). We found that women were more reactive to daily stressors that included others (interpersonal tensions and network events). They also reported more "joint-focused" and "other-focused" events, implying that they may also be more behaviorally reactive to others' problems (Bolger, DeLongis, Kessler, & Wethington, 1989).

Limitations and Future Directions

Of course, findings from this study should be considered in light of its limitations. First, although the sample represents a broad cross-section of adults in the United States, the respondents are predominately Caucasian. Our findings regarding the prevalence of stressors await future replication in more ethnically diverse samples. Second, because all of the respondents previously participated in a larger, complex study, involving a lengthy telephone interview followed by two self-administered questionnaires, the diary study participants are likely to be a highly compliant sample of people. A diary study attempted in a new sample might not yield such a high rate of response and might result in a higher proportion of missing days. Third, compared to many other diary studies, the present study sampled a relatively modest number of days per participant. The large sample size and the use of telephone interviews required a reduction in the number of study days of interviews. Although missing days were not a major problem, the limited number of days per respondent places more weight on the remaining ones. Fourth, there was modest falloff in reporting stressful events as the study week progressed. Although the falloff was not as steep as that reported in self-administered diary studies, it should be noted for future studies that using the more personal method of telephone interviewing does not by itself eliminate falloff. Fifth, this study represents the first use of the DISE measure, and all findings should be considered preliminary, pending replication. Sixth, a complication for replication is that use of the DISE is more expensive and burdensome for the investigator than the use of conventional self-administered diary questionnaires. The DISE requires interviewers who need more intense training and supervision than interviewers in studies that do not demand quick but accurate judgments about probing and question sequencing. Coders must also classify the narrative descriptions of events immediately after the interviews, so that this information can be used to improve interview quality.

Another limitation is that the DISE may reduce, but does not eliminate, all potential confounding between individual appraisal of stressors and the propensity to report them. The DISE technique is dependent on self-report and assumes that most people will respond to the questions honestly and report the content of daily stressors with accuracy. Although the DISE attempts to provide objective ratings of stressors, this method still relies on subjective accounts of stressful events. In the final analysis, all naturalistic stress research must assume that the people who experience stressful experiences are the best judges of what constitutes them.

The DISE contains several innovations that provide a new way to examine the role of daily stressors in promoting health and mood. One potential new area of research is greater specification of the health consequences of different types of daily stressors based on the type of threat they pose. Interview-based measures of major life events have provided evidence that the dimensions of threat involved in a major life event produce different types of emotional and physical disturbance. Life events posing the threat of loss or disappointment may be more likely to produce depression (Brown, 1989; Brown, Harris, & Hepworth, 1995). Life events characterized by danger may be more likely to produce physical and psychological symptoms of anxiety (Finlay-Jones, 1989; Harris, 1991). Repetitive, persistent interpersonal difficulties are related to chronic depression (Brown, Harris, Hepworth, & Robinson, 1994). Events resolving stressful conditions may in some cases end episodes of anxiety and depression (e.g., Brown & Moran, 1994). Thus, it is likely that daily stressors posing loss and danger as threats may be more likely to result in sad or anxious mood, respectively (Lazarus, 1999). Daily interpersonal stressors that persist over several days without resolution may lead to increasing mood disturbance over time. More generally, daily stressors involving frustration (stressors out of the control of respondents) may be more likely to produce physical or psychological fatigue. Finally, daily stressors that pose the possibility for positive outcomes (opportunity) may be more likely to produce hopeful or optimistic mood or physical energy at the end of the day.

APPENDIX
Daily Inventory of Stressful Events (DISE)

The DISE is a semistructured instrument consisting of four components: (a) a list of seven "stem" questions that pertain to occurrences of stressful events in various life domains, (b) a series of open-ended "probe" questions that ascertain a description of the stressful event, (c) a question regarding the perceived severity of the stressor, and (d) a list of structured primary appraisal questions, inquiring about goals and values that were "at risk" because of the event. An affirmative response to the stem questions prompts the interviewer to probe for a detailed description of the event, which is followed by questions pertaining to "what was at risk" for the respondent as a result of the event.

Stem Questions

1. Did you have an *argument or disagreement* with anyone since this time yesterday? No Yes
2. Since (this time/we spoke) yesterday, did anything happen that you *could have argued* about but you decided to let pass in order to avoid a disagreement? No Yes
3. Since (this time/we spoke) yesterday, did anything happen at *work or school* (other than what you have already mentioned) that most people would consider stressful? No Yes
4. Since (this time/we spoke) yesterday, did anything happen at *home* (other than what you have already mentioned) that most people would consider stressful? No Yes
5. Many people experience *discrimination* on the basis of such things as race, sex, or age. Did anything like this happen to you since (this time/we spoke) yesterday? No Yes
6. Since (this time/we spoke) yesterday, did anything happen to a *close friend or relative*

(other than what you have already mentioned) that turned out to be stressful for you?	No	Yes
7. Did *anything else* happen to you since (this time/we spoke) yesterday that most people would consider stressful?	No	Yes

Examples of Probes for Description

Ask only if "yes" for following stem questions:

1. Think of the most stressful disagreement or argument you had since (this time/we spoke) yesterday. Who was that with?	1
2. Think of the most stressful incident of this sort. Who was the person you decided not to argue with?	2
3. What happened and why did you decide not to get into an argument about it?	2
4. Think of the most stressful incident of this sort. What was the basis for the discrimination you experienced—your race, sex, age, or something else?	5
5. Think of the most stressful incident of this sort. Who did this happen to?	6
6. How does this affect your job?	3
7. What kinds of things were said?	1, 2
8. When did that happen? Was that some time yesterday or today?	All
9. What happened and what about it would most people consider stressful?	All
10. Have you had any problems with this in the past?	All
11. How long has this been going on?	All
12. Does this happen often?	All
13. Was there anything out of the ordinary in this?	All

Subjective Severity Question

1. How stressful was this for you—very, somewhat, not very, or not at all?

   1. Not at all→go to next stem question
   2. Not very→go to primary appraisal questions
   3. Somewhat→go to primary appraisal questions
   4. Very→go to primary appraisal questions

Primary Appraisal Questions	*not at all*	*a little*	*some*	*a lot*
1. How much were the following things at risk in this situation: First, how much did it risk disrupting your daily routine—a lot, some, a little, or not at all?	1	2	3	4
2. How much did it risk your financial situation?	1	2	3	4
3. How much did it risk the way you feel about yourself?	1	2	3	4
4. How much did it risk the way other people feel about you?	1	2	3	4
5. How much did it risk your physical health or safety?	1	2	3	4
6. How much did it risk the health or well-being of someone you care about?	1	2	3	4
7. How much did it risk your plans for the future?	1	2	3	4

NOTE

1. We conducted analyses to determine whether respondents were less likely to report daily stressors on later study days. To this end, correlations between the number of daily stressors and the study day (ranging from 1 to 8) were computed. A similar analysis was performed on the initial 8 days of a 42-day diary study using a self-report checklist of 22 stressors. The correlation between study day and number of reported stressors was –.17 for the present Daily Inventory of Stressful Events study and –.22 for the checklist study.

REFERENCES

Almeida, D. M. (1998). *Daily Inventory of Stressful Events (DISE) expert coding manual*. Tucson: Division of Family Studies and Human Development, University of Arizona.

Almeida, D. M., & Kessler, R. C. (1998). Everyday stressors and gender differences in daily distress. *Journal of Personality and Social Psychology*, *75*, 670-680.

Bebbington, P. E. (1986). Depression: Distress or disease? *British Journal of Psychiatry*, *148*, 479.

Bolger, N., DeLongis, A., Kessler, R. C., & Schilling, E. (1989). Effects of daily stress on negative mood. *Journal of Personality and Social Psychology*, *57*, 808-818.

Bolger, N., DeLongis, A., Kessler, R. C., & Wethington, E. (1989). The contagion of stress across multiple roles. *Journal of Marriage and the Family*, *51*, 175-183.

Bolger, N., DeLongis, A., Kessler, R. C., & Wethington, E. (1990). The micro-structure of role-related stress. In J. Eckenrode & S. Gore (Eds.), *Stress between work and family* (pp. 95-115). New York: Plenum.

Bradburn, N. M. (1969). *The structure of psychological well-being*. Chicago: Aldine.

Brantley P. J., & Jones, G. N. (1993). Daily stress and stress-related disorders. *Annals of behavioral medicine*, *15*, 17-25.

Brown, G. W. (1989). Life events and measurement. In G. W. Brown & T. O. Harris (Eds.), *Life events and illness* (pp. 3-45). New York: Guilford.

Brown, G. W., & Harris, T. O. (1978). *Social origins of depression: A study of depressive disorder in women*. New York: Free Press.

Brown, G. W., Harris, T. O., & Hepworth, C. (1995). Loss, humiliation and entrapment among women developing depression: A patient and non-patient comparison. *Psychological Medicine*, *25*, 7-21.

Brown, G. W., Harris, T. O., Hepworth, C., & Robinson, R. (1994). Clinical and psychosocial origins of chronic depressive episodes: II. A patient inquiry. *British Journal of Psychiatry*, *165*, 457-465.

Brown, G. W., & Moran, P. (1994). Clinical and psychosocial origins of chronic depressive episodes: I. A community survey. *British Journal of Psychiatry*, *165*, 447-456.

Dillman, D. A. (1991). The design and administration of mail surveys. *Annual Review of Sociology*, *17*, 225-249.

Eckenrode, J., & Bolger, N. (1995). Daily and within-day event measurement. In S. Cohen, R. C. Kessler, & L. U. Gordon (Eds.), *Measuring stress: A guide for health and social scientists* (pp. 80-101). New York: Oxford University Press.

Fieldler, K., & Stroehm, W. (1986). What kind of mood influences what kind of memory: The role of arousal and information structure. *Memory and Cognition*, *141*, 181-188.

Finlay-Jones, R. (1989). Anxiety. In G. W. Brown & T. O. Harris (Eds.), *Life events and illness* (pp. 95-112). New York: Guilford.

Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A., & Gruen, R. J. (1986). Dynamics of a stressful encounter: Stressor appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology*, *59*, 992-1003.

Harris, T. O. (1991). Life stress and illness: The question of specificity. *Annals of Behavioral Medicine*, *13*, 211-219.

Herbert, T. B., & Cohen, S. (1996). Measurement issues in research on psychosocial stress. In H. B. Kaplan (Ed.), *Psychosocial stress: Perspectives on structure, theory, life-course, and methods* (pp. 295-332). New York: Academic Press.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-II-R psychiatric disorders in the United States. *Archives of General Psychiatry*, *51*, 8-19.

Kessler, R. C., & McLeod, J. M. (1984). Sex differences in vulnerability to undesirable life events. *American Sociological Review*, *49*, 620-631.

Keyes, C.L.M., & Ryff, C. D. (1998). Generativity in adult lives: Social structural contours and quality of life consequences. In D. P. McAdams, & E. de St. Aubin (Eds.), *Generativity and adult development: How and why we care for the next generation* (pp. 227-263). Washington, DC: American Psychological Association.

Lachman, M., & Weaver, S. L., (1998). Sociodemographic variations in the sense of control by domain: Findings from the MacArthur studies of midlife. *Psychology & Aging*, *13*, 553-562.

Larsen, R. J., & Kasimatis, M. (1991). Day-to-day physical symptoms: Individual differences in the occurrence, duration, and emotional concomitants of minor daily illnesses. *Journal of Personality*, *59*, 387-423.

Lazarus, R. S. (1999). *Stress and emotion*. New York: Springer.

McQuaid, J., Monroe, S. M., Roberts, J. R., Johnson, S. L., Garamoni, G. L., Kupfer, D. J., et al. (1992). Toward the standardization of life stress assessment: Definitional discrepancies and inconsistencies in methods. *Stress Medicine*, *8*, 47-56.

Monroe, S. M., & Kelley, J. M. (1995). Measurement of stress appraisal. In S. Cohen, R. C. Kessler, & L. Gordon (Eds.), *Measuring stress: A guide for health and social scientists* (pp. 122-147). New York: Oxford University Press.

Mroczek, D. K., & Kolarz, C. M. (1998). The effect of age on positive and negative affect: A developmental perspective on happiness. *Journal of Personality and Social Psychology*, *75*, 1333-1349.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Raphael, K. G., Cloitre, M., & Dohrenwend, B. P. (1991). Problems with recall and misclassification with checklist methods of measuring stressful life events. *Health Psychology*, *10*, 62-74.

Rehm, L. P. (1978). Mood, pleasant events, and unpleasant events: Two pilot studies. *Journal of Consulting and Clinical Psychology*, *46*, 337-356.

Stone, A. A. (1987). Event content in a daily survey is differentially associated with concurrent mood. *Journal of Personal and Social Psychology*, *52*, 56-58.

Stone, A. A., Kessler, R. C., & Haythornthwaite, J. A. (1991). Measuring daily events and experiences: Methodological considerations. *Journal of Personality*, *59*, 575-607.

Stone, A. A., & Neale, J. M. (1984). New measure of daily coping: Development and preliminary results. *Journal of Personality and Social Psychology*, *46*, 892-906.

Stone, A. A., Reed, B. R., & Neale, J. M. (1987). Changes in daily event frequency precede episodes of physical symptoms. *Journal of Human Stress*, *13*, 70-74.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal & Social Psychology, 48*, 285-290.

Turner, R. J., & Wheaton, B. (1995). Checklist measurement of stressful life events. In S. Cohen, R. C. Kessler, & L. Gordon (Eds.), *Measuring stress: A guide for health and social scientists* (pp. 29-58). New York: Oxford University Press.

Wethington, E. (1997). The structured life event inventory. In C. Zalaquett & R. Wood (Eds.), *Evaluating stress: A book of resources* (pp. 391-403). Lanham, MD: Scarecrow/University Press.

Wethington, E., Brown, G. W., & Kessler, R. C. (1995). Interview measurement of stressful life events. In S. Cohen, R. C. Kessler, & L. Gordon (Eds.), *Measuring stress: A guide for health and social scientists* (pp. 59-79). New York: Oxford University Press.

Wethington, E., McLeod, J. D., & Kessler, R. C. (1987). The importance of life events for explaining the gender difference in psychological distress. In G. Baruch & R. Barnett (Eds.), *Gender and stress* (pp. 144-156). New York: Free Press.

Wheaton, B. (1999). Social stress. In C. S. Aneshensel & J. C. Phelan (Eds.), *Handbook of the sociology of mental health* (pp. 277-300). New York: Kluwer Academic/Plenum.

Zautra, A. J., Guarnaccia, C. A., & Dohrenwend, B. P. (1986). Measuring small events. *American Journal of Community Psychology*, *14*, 629-655.

David M. Almeida received his Ph.D. in developmental psychology from the University of Victoria and is currently an associate professor in the Division of Family Studies and Human Development at the University of Arizona in Tucson. The topics of his recent publications include sources of gender differences in psychological distress and emotional transmission in the daily lives of families. His current research interests focus on environmental and genetic components of daily stress processes during adulthood.

Elaine Wethington received her Ph.D. in medical sociology from the University of Michigan. She currently holds a joint appointment in human development and sociology at Cornell University. Her research interests are in the areas of situational determinants of exposure to stress, access to social support, and successful coping.

Ronald C. Kessler received his Ph.D. in sociology from New York University and is currently a professor of health care policy at the Harvard Medical School. He is the recipient of a Research Scientist Award and a Merit Award from the National Insitute of Mental Health. His research deals with the psychosocial determinants and consequences of mental illness.

The Use of Reliable Digits to Detect Malingering in a Criminal Forensic Pretrial Population

Scott A. Duncan
Denella L. Ausborn
United States Penitentiary, Atlanta

*The present research is a cross-validation of previous investigation by Greiffenstein, Baker, and Gola; Greiffenstein, Gola, and Baker; and Meyers and Volbrecht on the reliable digits (RELD) method of detecting suspected malingering on the Wechsler Adult Intelligence Scale–Revised (WAIS-R). The results support the use of the RELD method on a criminal forensic pretrial population (*N *= 187). Sensitivities, specificities, and incremental hit rates for two cut levels of the RELD method, Minnesota Multiphasic Personality Inventory–2 (MMPI-2) Infrequency and the Personality Assessment Inventory Negative Impression Scales, as well as multiple combined cut scores, were comparable to those observed in previous studies that used neuropsychologically evaluated participants. The selection of which cut score or combination of cut scores is appropriate on the RELD method is also discussed.*

Keywords: reliable digits, WAIS–R, criminal population

REVIEW OF LITERATURE

According to the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1994), malingering is the "intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs" (p. 683). Although the prevalence of malingering varies widely across forensic settings (Resnick, 1988), the importance of the identification of feigning is paramount to society and/or the accused (Suhr, Tranel, Wefel, & Barrash, 1997).

Many tests and procedures have been employed over the years to detect possible malingering. Although most personality tests traditionally have validity scales built into the questionnaire (Greene, 1991; Morey, 1996; Pope, Butcher, & Seelen, 1993), many popular and frequently used intellectual and neuropsychological tests do not have the ability for such measurements (Wechsler, 1981). In response to this problem, many evaluators utilize single measures of malingering and extrapolate to the test results (Binder, 1993; Hiscock & Hiscock, 1989). However, it is thought that by having validity indicators on specific clinical tests, along with single or even multiple measures of malingering from tests designed to specifically detect dissimulation, one could obtain a more valid and reliable method for detecting feigning (Trueblood & Schmidt, 1993). Recent research (Greiffenstein, Gola, & Baker, 1995) has helped to establish a built-in validity indicator for use with certain populations on the Wechsler Adult Intelligence Scale–Revised (WAIS-R) (Wechsler, 1981).

Most cognitive and neuropsychological tests do not have built-in detectors of malingering. Indeed, it has been shown that many neuropsychological tests can be faked without detection. Heaton, Smith, Lehman, and Vogt (1978) used volunteer dissimilators who produced credible neuropsychological deficits on the Halstead-Reitan neuropsychological battery. Furthermore, the Wechsler Memory Scales were unable to distinguish volunteer

This research was done under the auspices of the Federal Bureau of Prisons. However, the views expressed are those of the authors only and do not represent the official policy or opinions of the Federal Bureau of Prisons. The authors would like to thank Drs. Michael Plasay and Nancy Blywise for their kind review of this research. Correspondence concerning this article and requests for reprints should be addressed to Scott Duncan, Forensic Studies Coordinator, Psychology Services, United States Penitentiary, 601 McDonough Boulevard, S.E., Atlanta, GA 30315; e-mail: sxduncan@bop.gov.

Assessment, Volume 9, No. 1, March 2002 56-61

dissimilators, real-world dissimilators, and TBI (Traumatic Brain Injury) patients (Greiffenstein, Baker, & Gola, 1994; Mittenberg, Azrin, Milsaps, & Heilbronner, 1993). Furthermore, the use of personality test results to establish malingering in a brain-injured population has also been brought into question. Greiffenstein et al. (1995) cautioned against the use of the Minnesota Multiphasic Personality Inventory–2 (MMPI-2) validity scales in a neuropsychological context. Their results showed that domain-specific measures of malingering were generally more sensitive to noncompliance than were indicators from the MMPI-2. Only scale Sc (Schizophrenia) of the MMPI-2 appeared to have even a modest ability to detect noncompliance in a neuro- psychological forensic context.

Prior to Greiffenstein et al.'s (1994, 1995) research, the WAIS-R was without any established indicators of malingering. Greiffenstein et al. utilized the reliable digits (RELD) method from the Digit Span subtest on the WAIS-R to detect dissimilation. This method is simply the sum of the number of digits forward and backward where the examinee last received a 2-point score. As an example, if an examinee recited correctly both trials on digits forward where each trial was four digits long, and then failed one or both trials on all longer sequence digits, then the examinee's RELD forward score would be 4. Following, if the same examinee passed both trials on digits backward where the digits presented were three digits long and failed one or both of all remaining trials, then the examinee's RELD backward score would be 3. The total RELD score in this example would be the sum of RELD forward and backward or 7. The utility of this method in detecting malingering is supported by research that shows the discrepancy between forward and backward spans in brain-damaged populations (Kaplan, Fein, Morris, & Delis, 1991). Malingerers are more likely to malinger both digits forward and digits backward, possibly resulting in a lower total digits score than would be seen in a brain-damaged population.

Greiffenstein et al.'s (1994, 1995) and others' research has primarily been conducted on civil litigants or individuals suspected of head injuries. In addition, Meyers and Volbrecht (1998) researched 47 mild brain-injured litigating and 49 mild brain-injured nonlitigating participants using the RELD method. The results indicate only 4.1% of the nonlitigating participants were classified as malingerers by RELD, whereas 48.9% of the litigating participants were classified as malingerers when the RELD method was employed. When using a forced choice (FC) task as a standard, the RELD achieved 77.8% specificity and 95% sensitivity, when compared to the FC task.

In keeping with studies on the WAIS, which indicate the lack of effect that mental conditions (other than anxiety) have on performance levels on the Digit Span subtest, below that seen in brain-injured participants (Bloom & Goldman, 1962; Ladd, 1964), Suhr et al. (1997) found no difference between the Digit Span performance of depressed and somatic verses head-injured patients. This study compared scores on the Digit Span subtest of the WAIS-R and scores from the Auditory Verbal Learning Test (Rey, 1964) across six groups, including differing levels of head-injured participants, litigating versus nonlitigating participants, probable malingerers, clinically depressed participants, and participants with a somatization disorder. These results showed that several memory tests were useful in distinguishing probable malingerers from the other groups. Furthermore, a complex interaction between malingering status, psychological factors, and medication use in the prediction of memory test performance was investigated. These results emphasize the need to consider nonneurological factors in the interpretation of poor memory performance in the studied population.

Heaton et al. (1978) found that valid and invalid IQ profiles could be distinguished by scores on the Digit Span subtest. Although a history of head trauma may lower Digit Span performance (Evans & Marmorston, 1964; Templer, 1967; Tolor, 1958; Wechsler, 1958), Sterne (1969) found no differences between normals, nonclassified, and organic groups on total digits, number of digits forward, and number of digits recalled backward. Furthermore, Mittenberg, Theroux-Fichera, Zielinski, and Heilbronner (1995) compared 67 nonlitigating head-injured patients with age, IQ, and occupation-matched participants asked to simulate head trauma symptom results on the WAIS-R. It was concluded by these researchers that truly head-injured patients seem to show a pattern on the WAIS-R subtest scores that can be distinguished from dissimulated profiles. However, Rogers and Cruise (1998) warned about the use of simulators by pointing out that simulators as a group do not always perform the same.

To date, no studies reviewed have attempted to research criminal litigants without head injury using the RELD method. In the criminal litigant population, evaluators are warned of the heightened possibility of malingering psychological symptoms (American Psychiatric Association, 1994). However, many criminal litigants will also attempt to malinger low IQ, poor memory, or a combination of all three. Through comparing measures of feigned psychological symptoms with a measure of cognitive malingering, this research sought to contribute to the field of psychological assessment by broadening the literature on the use of the RELD method as a viable tool to detect malingering. These results may have particular implications to the field of criminal forensic psychology considering the impact of assessment on the disposition of criminal cases.

HYPOTHESIS

The purpose of this study is to provide validation of the usefulness of the RELD method in helping to identify malingering in a criminal forensic pretrial/presentence population. It is hypothesized that results similar to Greiffenstein et al. (1994, 1995) and Meyers and Volbrecht (1998) on neuropsychologically assessed populations will be replicated on a criminal forensic pretrial/presentence population. More specifically, we predict the RELD method will detect previously diagnosed malingerers from individuals who were diagnosed as nonmalingerers in a criminal forensic pretrial/presentence population. A comparison between the RELD method and established scales of possible feigned psychological symptoms, MMPI-2 Infrequency Scale (F) (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and the Personality Assessment Inventory (PAI) Negative Impression Scale (NIM) (Morey, 1991), will be performed.

METHOD

All case records ($N = 187$) of adult males with no diagnosed neurological impairment and that contained WAIS-R protocols including the Digit Span subtest, the MMPI-2 F, and the PAI NIM were selected from an archival database maintained at the U.S. Penitentiary in Atlanta, Georgia. This database was obtained from pretrial/presentence detainees who were sent to this facility by federal courts across the United States from 1989 to 1999. Previously made diagnoses for each detainee resulted in the placement of each case into either a malingering group (a priori diagnosis of malingering) or a nonmalingering group (no a priori diagnosis of malingering). A diagnosis of malingering was made based on the clinical judgment of the various evaluators. These clinical judgments were based on the entire evaluation process that included an extensive records review; psychological testing results, which often included validity scores from the MMPI, MMPI-2, PAI, the Structured Interviewed of Reported Symptoms (Rogers, 1986), and Rey's (1964) 15-Item Test; interviews of the detainees; and reported observations of the detainee's behavior throughout the 30- to 45-day evaluation process. The distinction of whether the detainee was malingering emotional, cognitive, memory, or physical symptoms was not recorded as part of the established database. The RELD method was not used to assist in the identification of malingerers as none of the evaluators were aware of this method prior to making their final diagnoses. The cases included 98 Caucasian detainees, 84 Black American detainees, 4 Hispanic detainees, and 1 Native American detainee.

The mean age of the malingering group ($n = 54$) was 34.69 years ($SD = 10.07$ years), with an average of 8.86 years of education ($SD = 2.24$ years). The mean full scale IQ (FSIQ) score on the WAIS-R for the malingering group was 70.69 ($SD = 13.09$). Of the malingering group cases, 23 (43.39%) were Caucasian, 27 (50.94%) were Black American, and 3 (5.66%) were Hispanic.

The nonmalingering group ($n = 134$) displayed a mean age of 38.06 years ($SD = 9.50$ years), with a mean education level of 10.51 years ($SD = 2.7$ years). The mean FSIQ score for the nonmalingering group on the WAIS-R was 87.67 ($SD = 12.55$). Race groupings for the nonmalingering group revealed 75 (55.97%) Caucasian detainees, 57 (42.53%) Black American detainees, 1 (0.74%) Hispanic detainee, and 1 (0.74%) Native American detainee.

The malingering group approached but was not significantly younger ($t = –2.15$, $p = .033$) than the nonmalingering group. There were no significant differences between races in the two groups (all ps < .01). However, the malingering group had a significantly lower level of education ($t = –3.78$, $p = .0002$) and provided significantly lower FSIQ scores ($t = –8.23$, $p < .0001$) than the nonmalingering group.

All cases (malingering group and nonmalingering group) were evaluated at the U.S. Penitentiary in Atlanta, Georgia, as a result of a court-ordered pretrial/presentence evaluation over a period of 30 to 45 days. All detainees were housed in a pretrial unit during their evaluations. Detainees were sent to the Atlanta federal facility from 50 different federal district courts ranging from Montana to Maine in the north and northwest and from Florida and Arizona in the south and southwest. A team approach with a minimum of one licensed psychologist and a predoctoral psychology intern was utilized when conducting the evaluations in this study. Most of the tests were administered by a predoctoral psychology intern under the close supervision of a licensed psychologist. All measures were administered as part of a comprehensive battery of tests that were included in the evaluation to answer the federal court's referral question(s). All tests were administered according to standardized procedures. Of the cases in the malingering and nonmalingering groups, 180 cases (96.25%) were evaluated for the referral question(s) of competency to stand trial and/or responsibility at the time of the offense as per Title 18 U.S.C. § 4241 and 4242 (*Federal Criminal Codes and Rules*, 1993), respectively. The remaining 7 cases (3.25%) were evaluated for varying presentence questions posed by the court in accordance with Title 18 U.S.C. § 3552 (*Federal Criminal Codes and Rules*, 1999). These 7 cases were included as they were also unsentenced detainees who presumably had an equal motivation to malinger deficits to obtain a shorter sentence. They were not viewed as a unique group apart from the 180 cases.

RESULTS

Table 1 displays the means, standard deviations, and ranges for the RELD, MMPI-2 F, and PAI NIM scores. Significant differences between groups were obtained via *t* tests and are noted in Table 1. As can be seen, the malingering and nonmalingering groups differed significantly across all three scores. Furthermore, Pearson's correlation coefficients were obtained between MMPI-2 F, PAI NIM, and RELD. The Pearson's correlation between MMPI-2 F and RELD was –.42 ($p < .05$) with respective correlations of –.28 ($p < .05$) and –.22 ($p < .05$) for the malingering and nonmalingering groups. The PAI's NIM Scale had an overall correlation with RELD of –.34 ($p < .05$). The malingering and nonmalingering groups correlated –.13 (*ns*) and –.18 ($p < .05$), respectively, with NIM and RELD. Finally, the MMPI-2's F Scale correlated overall with the PAI's NIM Scale, .73 ($p < .05$), with respective correlations of .52 ($p < .05$) and .70 ($p < .05$) for the malingering and nonmalingering groups.

Predictive accuracy was examined under two decision rules, following the procedure in Greiffenstein et al. (1994). A conservative cut score for RELD was set at 1.3 *SD* below the mean of the nonmalingering group's RELD score and 1.3 *SD*s above the mean of the nonmalingering group's MMPI-2 F and PAI NIM scores. A second and less conservative decision rule was used, establishing a cut score at –1.0 and 1.0 *SD*s of the nonmalingering group's RELD and MMPI-2 F and PAI NIM scores, respectively. For RELD scores, cases at or below –1.3/–1.0 *SD* for the nonmalingering group were classified as malingering. For MMPI-2 F and PAI NIM scores, participants were classified as malingering if their score was 1.3/1.0 *SD* of the mean for the nonmalingering group. A comparison of actual malingering diagnosis and predicted malingering diagnosis across the three malingering predictor variables is presented in Table 2. Data are expressed in terms of sensitivities, specificitiies, and incremental hit rates. As in Greiffenstein et al. (1995), the incremental hit rate expresses the enhancement of prediction by using a singular cut score, above and beyond that which could have been predicted by the a priori base rate of 28%. This base rate was calculated by establishing the percentage of cases actually diagnosed as malingering in the 187 cases used in this study.

Satisfactory hit rates were achieved by all malingering predictor variables, above that which could have been predicted by a base rate guess. As can be seen in Table 2, all incremental hit rates were within 11 percentage points of each other. The highest incremental hit rate was achieved by the RELD method when using the most conservative cut score of 6, followed by the MMPI-2's F Scale when using the less conservative cut score of 102 T (T scale distribution with a mean of 50 and a standard deviation of 10). The smallest but comparable hit rate was achieved by using the less conservative cut score of 7 on the RELD method. The highest sensitivity achieved (identifying malingerers who had an actual diagnosis of malingering) was from using the less conservative cut score of 102 on the MMPI-2 F Scale. The second highest sensitivity came from using the RELD method at the less conservative cut score of 7. However, the highest specificity level achieved came from using the more conservative cut score of 6 on the RELD method. By using this approach, less than 10% of nonmalingerers were incorrectly classified as malingerers.

Predictive accuracy was further studied by combining the two sets of cut scores on all three measures, thus producing multiple cut-off criteria for each set of cut scores. In other words, in the most conservative case, a detainee was considered malingering only if he had a score of 6 or lower on the RELD method, a score of 100 or above on the PAI's NIM Scale, and a score of 109 on the MMPI-2's F.

TABLE 1
Means, Standard Deviations, Ranges, and *t*-Test Results for the Two Groups on the RELD, MMPI-2's F, and PAI's NIM Scores

	Group					
	Malingerers			*Nonmalingers*		
Measure	M	SD	*Range*	M	SD	*Range*
RELD	5.81*	3.40	0-12	8.87*	2.14	2-15
MMPI-2 F	108.64*	18.55	48-120	76.98*	24.72	42-120
PAI NIM	96.30*	22.62	44-144	70.72*	22.42	44-129

NOTE: RELD = reliable digits; MMPI-2 = Minnesota Multiphasic Personality Inventory–2; F = Infrequency Scale; PAI = Personality Assessment Inventory; NIM = Negative Impression Scale.
*Malingering and nonmalingering groups significantly differed at $p <$ .001.

TABLE 2
Cutting Scores, Sensitivities, Specificities, and Incremental Hit Rates for RELD, NIM, and F (Malingering vs. Nonmalingering)

Measure	*Cutting Score*[a]	*Sensitivity (%)*	*Specificity (%)*	*Incremental Hit Rate*
RELD	6/7	56.6/67.9	90.3/71.6	52.7/42.5
NIM	100/93	41.5/58.5	86.6/82.8	45.7/47.9
F	109/102	66.0/77.3	82.8/79.9	50.0/51.1

NOTE: RELD = reliable digits; NIM = Negative Impression Scale; F = Infrequency Scale.
a. The first cutting score is based on a *z* score of –1.3 for RELD and 1.3 for NIM and F from the mean of the nonmalingering group distribution. The second cutting score is based on a *z* score of –1.0 for RELD and 1.0 for NIM and F from the mean of the nonmalingering group distribution. Actual malingering diagnosis base rate = 28%.

By using this approach with both sets of cut scores, sensitivity was greatly reduced. As can be observed in Table 3, significantly fewer actual malingerers were identified. However, specificity was increased to nearly 100%. Incremental hit rates were not appreciably affected by using the combined approach.

TABLE 3
Cutting Scores, Sensitivities, Specificities, and Incremental Hit Rates for Combined RELD, NIM, and F (Malingering vs. Nonmalingering)

Measure	*Cutting Score*[a]	*Sensitivity (%)*	*Specificity (%)*	*Incremental Hit Rate*
RELD/NIM/F	6/100/109	24.5	98.4	49.0
RELD/NIM/F	7/93/102	39.6	97.7	52.0

NOTE: RELD = reliable digits; NIM = Negative Impression Scale; F = Infrequency Scale.
a. The first cutting scores are based on a *z* score of –1.3 for RELD and 1.3 for NIM and F from the mean of the nonmalingering group distribution. The second cutting scores are based on a *z* score of –1.0 for RELD and 1.0 for NIM and F from the mean of the nonmalingering group distribution. Actual malingering diagnosis base rate = 28%.

DISCUSSION

A priori diagnosed groups separated into malingering and nonmalingering diagnoses were compared across three measures of possible malingering on a criminal forensic pretrial/presentence population. As a result, this study supports the findings by Greiffenstein et al. (1994, 1995) and Meyers and Volbrecht (1998) that report the RELD method as a viable indicator of possible malingering on the WAIS-R. Although Greiffenstein et al. (1995) suggested that MMPI-2 validity scales, including F, may be limited in their ability to detect malingering in a personal injury population, this study would not support this contention when used on a criminal forensic pretrial/presentence population. It appears from these results that the individuals in this criminal forensic pretrial/presentence population who chose to malinger psychotic symptoms on the MMPI-2, as measured by a heightened F Scale score (Dahlstrom, Welsh, & Dahlstrom, 1972; Greene, 1991), were able to be identified as malingers a significant amount of the time. However, it is questionable whether those that attempt to malinger psychopathology, as measured by an elevated F Scale, were also likely to attempt to malinger poor immediate memory recall, as measured by the RELD method, given the similarities between the correlations of these two variables (F and RELD) and the two groups (malingering and nonmalingering). It would appear to be most clinically useful to use varying types of malingering measures to identify malingerers in a criminal pretrial population. The use of these measures in a multiple cut-off fashion does not appear to be clinically useful as some criminal defendants attempted to malinger psychopathology, whereas other defendants malingered cognitive deficits. In addition, Meyers and Volbrecht pointed out that a lower score on the RELD method may simply reflect a general poor performance on the WAIS-R. This contention appears to be supported, in part, in this study when ones observes the Pearson correlation between the FSIQ and RELD of .74 ($p < .05$), with the malingering and nonmalingering groups being .86 and .56 (ps < .05), respectively. If RELD is simply a reflection of poor overall performance on the WAIS-R, one would expect a significantly high correlation between RELD and FSIQ in the malingering group, as noted above. However, to explore this further, one would need to evaluate whether the RELD method had the same correlation to FSIQ as did other WAIS-R subtest scores in the malingering group. Due to the archival nature of the data used in this study, this was beyond the scope of the present research. Further research is needed to replicate Meyers and Volbrecht's findings on a criminal forensic population.

Of most significance in this study are the results of incremental hit-rate comparisons between recognized validity markers on two widely used personality tests and the use of the RELD method derived from the WAIS-R subtest, Digit Span. This study would support the contention that the RELD method can be used as a possible validity marker on the WAIS-R and, presumably, the WAIS-III (Meyers & Volbrecht, 1998). Future research could focus on other validity markers such as the MMPI-2 (Fb) Scale and/or the PAI and Malingering Index Scales. However, due to the archival nature of the database used, these comparisons are not currently obtainable. Research with similar results using these scales would result in more powerful conclusions because these validity indicators have been found to be more specific to malingering per se than F or NIM (Morey, personal communication, February 20, 2001). Furthermore, it should be noted that the hit rates for the MMPI F and PAI NIM involve criterion contamination; that is to say, the decision to place a respondent in the malingering group was based, in part, on these measures. Therefore, for these two measures, the hit rate does not represent an independently validated estimate but rather a part-whole relationship.

The RELD method appears to be more clinically useful when a cut score of 7 is employed in a criminal forensic pretrial/presentence population, if one is concerned with accurately identifying possible malingerers at the expense of identifying someone as nonmalingering when he or she was actually malingering. However, the use of a single indicator to identify malingering would seem to be clinically unacceptable. Although the accuracy of clinical judgment regarding malingering does not improve incrementally with the number of malingering indicators (Sechrest, 1963), it appears prudent to use multiple measures and indicators of malingering before making such a diagnosis in any population. With this general rule, it would seem clinically judicious to use a cut score of 7 when implementing the use of the RELD method, only if other independent indicators of malingering are obtained. However, if one wanted to maximize the correct classification of nonmalingerers and minimize the chance of diagnosing someone as malingering when indeed he or she was not, a cut score of 6 could be used without losing an appreciable amount of sensitivity. In either case, the use of the RELD method as the sole indicator for malingering is not supported by this research. Yet the RELD method appears to be an accurate indicator from the WAIS-R of attempted feigning of memory problems (or perhaps poor general performance), not withstanding significant anxiety (Reitan & Wolfson, 1985) during the administration of the WAIS-R. Replication of these findings on a criminal population is warranted before final conclusions concerning the use of the RELD approach are made.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Binder, L. (1993). An abbreviated form of the Portland Recognition Test. *The Clinical Neuropsychologist*, *7*, 104-107.

Bloom, B., & Goldman, R. (1962). Sensitivity of the WAIS to language handicap in a psychotic population. *Journal of Clinical Psychology*, *18*, 161-162.

Butcher, J., Dahlstrom, W., Graham, J., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory–2: Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Dahlstrom, W., Welsh, G., & Dahlstrom, L. (1972). *An MMPI handbook: Clinical interpretation*. Minneapolis: University of Minnesota Press.

Evans, R., & Marmorston, J. (1964). Perceptual test signs of brain damage in cerebral thrombosis. *Perceptual and Motor Skills*, *18*, 977-988.

Federal criminal codes and rules. (1993). St. Paul, MN: West Publishing Company.

Greene, R. (1991). *The MMPI-2/MMPI: An interpretive manual*. Boston: Allyn & Bacon.

Greiffenstein, M., Baker, W., & Gola, T. (1994). Validity of malingered amnesia measures in a large clinical sample. *Psychological Assessment*, *6*, 218-224.

Greiffenstein, M., Gola, T., & Baker, W. (1995). MMPI-2 validity scales versus domain specific measures in detection of factitious traumatic brain injury. *The Clinical Neuropsychologist*, *9*, 230-240.

Heaton, R., Smith, H., Lehman, R., & Vogt, A. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *46*, 892-900.

Hiscock, M., & Hiscock, C. (1989). Refining the forced choice method for detecting malingering. *Journal of Clinical and Experimental Neuropsychology*, *11*, 967-974.

Kaplan, E., Fein, D., Morris, R., & Delis, D. C. (1991). *WAIS-R as a neuropsychological instrument*. San Antonio, TX: Psychological Cooperation.

Ladd, C. (1964). WAIS performance of brain damaged and neurotic patients. *Journal of Clinical Psychology*, *20*, 114-117.

Meyers, J., & Volbrecht, M. (1998). Validation of reliable digits for detection of malinger. *Assessment*, *5*, 303-307.

Mittenberg, W., Azrin, R., Milsaps, C., & Heilbronner, R. (1993). Identification of malingering head injury on the Wechsler Memory Scale–Revised. *Psychological Assessment*, *5*, 34-40.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R., & Heilbronner, R. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale–Revised. *Professional Psychology: Research and Practice*, *26*, 491-498.

Morey, L. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Morey, L. (1996). *An interpretive guide to the Personality Assessment Inventory (PAI)*. Odessa, FL: Psychological Assessment Resources.

Pope, K., Butcher, J., & Seelen, J. (1993). *MMPI, MMPI-2, and MMPI-A in court*. Washington, DC: American Psychological Association.

Reitan, R., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery*. Tucson, AZ: Neuropsychological Press.

Resnick, P. (1988). Malingering of posttraumatic disorders. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 84-107). New York: Guilford.

Rey, A. (1964). *L'examen clinic en psychologie* [The clinical exam in psychology]. Paris: Presses Universitaires de France.

Rogers, R. (1986). *Structured Interview of Reported Symptoms (SIRS)*. Odessa, FL: Psychological Assessment Resources.

Rogers, R., & Cruise, K. (1998). Assessment of malingering with simulation designs: Threats to external validity. *Law and Human Behavior*, *22*, 273-285.

Sechrest, L. (1963). Incremental validity: A recommendation. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., p. 396). New York: Guilford.

Shur, J., Tranel, D., Wefel, J., & Barrash, J. (1997). Memory performance after head injury: Contributions of malingering, litigation status, psychological factors, and medication use. *Journal of Clinical and Experimental Neuropsychology*, *19*, 500-514.

Sterne, D. M. (1969). The Benton, Porteus, and WAIS Digit Span tests with normal and brain-injured subjects. *Journal of Clinical Psychology*, *25*, 173-175.

Templer, D. (1967). Relation between immediate and short-term memory and clinical implications. *Perceptual and Motor Skills*, *24*, 1011-1012.

Tolor, A. (1958). Further studies on the Bender Gestalt test and the Digit Span test as measures of recall. *Journal of Clinical Psychology*, *14*, 14-18.

Trueblood, W., & Schmidt, M. (1993). Malingering and other validity considerations in the neuropsychological evaluation of head injury. *Journal of Clinical and Experimental Neuropsychology*, *15*, 578-590.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. New York: Psychological Corporation.

Scott A. Duncan is currently the director of the Violence and Psychopathy Assessment Unit at the United States Penitentiary in Atlanta, Georgia. He is also the director of Clinical Training at this facility and has a private practice in the Atlanta area.

Denella L. Ausborn is in full-time private practice in the Atlanta area.

Identifying Faking Bad on the Minnesota Multiphasic Personality Inventory–Adolescent With Mexican Adolescents

Emilia Lucio
Consuelo Durán
National Autonomous University of Mexico

John R. Graham
Yossef S. Ben-Porath
Kent State University

This study examined the extent to which the validity scales of the Minnesota Multiphasic Personality Inventory–Adolescent identified Mexican adolescents who were instructed to fake bad. Validity scales data were used to differentiate between nonclinical adolescents instructed to fake bad and both clinical and nonclinical adolescents who received standard instructions. Participants were 59 male and 87 female Mexican high school students and 59 male and 87 female Mexican adolescents from clinical settings. This is the first study on faking with adolescents in Mexico. The F, F1, and F2 Scales and the F-K index discriminated adequately between the three different groups. Results were similar to those previously reported for adults and adolescents in Mexico and the United States. High positive and negative predictive powers and overall hit rates were obtained in this study. Higher cut scores were needed to discriminate between the groups of girls than between the groups of boys.

Keywords: fake bad, MMPI-A, Mexican adolescents

The Minnesota Multiphasic Personality Inventory–Adolescent (MMPI-A) (Butcher et al., 1992) was developed in the United States to assess issues specific to adolescents, using a current age-appropriate normative sample. The Mexican version of the MMPI-A was developed using a stringent methodology to adapt the test to the Mexican population (Lucio, Ampudia, Duran, Gallegos, & Leon, 1999).

The current study examined the identification of faking bad on the MMPI-A among Mexican adolescents. Previous research has indicated that individuals feign psychopathology for different reasons (Adelman & Howard, 1984; Medoff, 1999; Resnick, 1984). Much research has been conducted with respect to test-taking attitudes of adults on the MMPI and the MMPI-2, but relatively little research has been published with respect to test-taking attitudes on the MMPI-A.

Berry, Baer, and Harris (1991) conducted a meta-analysis of 28 MMPI studies of faking bad in adults. They reported large effect sizes for raw scores and for *T*-scores on the F (Infrequency) Scale and for F-K (Infrequency minus Defensivenss) index, indicating that these measures are effective in detecting fake-bad profiles. Rogers, Sewell, and Salekin (1994) conducted a meta-analysis of 15 studies of malingering on the MMPI-2. The F scale, F-K index, and obvious-subtle index had large effect sizes and were

This article was written while Emilia Lucio was a visiting professor at Kent State University, Kent, Ohio. We appreciate the support given to the project by DGAPA, National Autonomous University of Mexico, Kent State University, and CONACYT Project 27527H. Correspondence regarding this article should be addressed to Emilia Lucio, Facultad de Psicología, División de Estudios de Posgrado, UNAM, Av Universidad 3004, México 04510, D.F.; e-mail: melgm@servidor.unam.mx.

Assessment, Volume 9, No. 1, March 2002 62-69

the best indicators of malingering. The Fb (Back Infrequency Scale) Scale and the revised Social Desirability Scale also yielded large effect sizes. Rogers et al. recommended a raw score greater than 23 on the F Scale or an F-K index greater than 10 as basic screening indicators of malingering. These MMPI-2 results are quite similar to those previously reported for the original MMPI. More recent MMPI-2 faking-bad studies with adults (Bagby, Nicholson, Buis, & Bacchiochi, 2000; Cramer, 1995; Lim & Butcher, 1996) also suggest that the F and Fb Scales and the F-K index are the best ways to discriminate persons faking bad from those completing the MMPI-2 with standard instructions.

Less attention has been given to identification of deviant test-taking attitudes among adolescents, especially to the detection of faking. Consistent with the adult literature, Archer, Gordon, and Kirchner (1987) found that adolescents attempting to fake bad on the original MMPI showed a grossly exaggerated picture of symptomatology. Archer et al., comparing MMPI results of nonclinical adolescents faking bad and psychiatric inpatient adolescents, under standard instructions, found that the best classification rates were obtained with F Scale raw scores ≥ 26. Herkov, Archer, and Gordon (1991) examined the scores of nonclinical adolescents instructed to fake bad on the MMPI and adolescent psychiatric patients who completed the MMPI with standard instructions. They found that the standard validity scales were effective indicators of faking. A *T*-score cutoff greater than 100 on the F Scale achieved an overall hit rate of 93.5% in classifying fake-bad and inpatient participants.

Stein, Graham, and Williams (1995) examined how well the validity scales of the MMPI-A discriminated between nonclinical adolescents instructed to fake bad, adolescents from clinical settings, and nonclinical adolescents taking the test with standard instructions. Adolescents who were instructed to fake bad obtained scores suggestive of overreporting of psychopathology. These adolescents had much higher F Scale and clinical scale scores, and lower K (Defensiveness) Scale scores, than the clinical sample and the nonclinical adolescents under standard instructions. Stein et al. developed cutoff scores to discriminate between the faking-bad and standard instruction conditions. A raw score cutoff of 22 for girls and 26 for boys on the F Scale offered the best discrimination between fakers and adolescents under standard instructions. These scores correctly classified 80% and 72.4% of the faked profiles for girls and boys, respectively, and 97.5% and 98.3%, respectively, of the profiles obtained with standard instructions for girls and boys. Accurate classification was also possible when discriminating between adolescents faking bad and clinical participants completing the MMPI-A with standard instructions. An F Scale raw score cutoff of 23 correctly classified 97.5% of the girls and 100% of the boys in the clinical. The F-K index was as effective as the F Scale in detecting faked profiles.

Rogers, Hinds, and Sewell (1996) compared the clinical usefulness of the MMPI-A and two other measures, the Structured Interview of Reported Symptoms and the Screening Index of Malingered Symptoms, in a within-participants analogue study with 53 dually diagnosed adolescent offenders. There were significant differences on the MMPI-A clinical scales between adolescents instructed to fake bad and those given standard instructions, with feigners tending to overreport symptoms. Although the F Scale was not successful in detecting fakers, the F-K index was clinically useful with positive predictive power (PPP) and negative predictive power (NPP) rates exceeding 80%. The optimal F-K cutoff score was 20, a value that varies from most adult cutting scores.

A study of faking bad among Mexican adolescents is important to establish optimal cutoff scores to detect overreporting in this country. In addition, in the United States there are no studies concerning faking among Hispanic adolescents. However, this is an important group to be considered because the proportion of Hispanics, especially Mexican Americans, in the United States is increasing rapidly. The purpose of the current study was to determine if the Mexican Spanish version of the MMPI-A could differentiate between adolescents who fake bad and those who complete the MMPI-A under standard instructions and if the same scales and cutoff scores used with the standard MMPI-A are appropriate for the Mexican Spanish version.

METHOD

Participants

The participants were 146 nonclinical adolescents (87 girls and 59 boys) in junior and senior high school and 146 clinical adolescents (87 girls and 59 boys) in psychiatric or psychological treatment. The nonclinical adolescents participated voluntarily without payment. The nonclinical sample came from a public school selected because it had some demographics similar to those of the clinical sample. Demographic characteristics of the nonclinical sample are summarized in Table 1.

The clinical sample was selected from a larger sample that was part of the standardization project of the MMPI-A in Mexico (Lucio, Ampudia, & Duran, 1998). Twenty-eight (47%) of the boys were from inpatient settings, and 31 (53%) were from outpatient settings. Seventeen girls (20%) were from inpatient psychiatric settings and 70 (80%) from outpatient psychiatric settings. Demographic

TABLE 1
Description of Participants

	Nonclinical		*Clinical*	
	Boys	*Girls*	*Boys*	*Girls*
Mean age	15.58	15.80	15.56	15.14
SD	0.93	0.95	1.37	1.22
Range	14-18	14-18	14-18	14-18
Living arrangement				
Lived with both parents	52 (88%)	67 (77%)	34 (58%)	43 (49%)
Lived only with the mother	5 (9%)	17 (29%)	11 (19%)	24 (28%)
Lived only with the father	1 (2%)	1 (1%)	3 (5%)	4 (5%)
Lived with other relatives	1 (2%)	0	5 (9%)	8 (9%)
Did not indicate with whom they lived	0	2 (2%)	5 (9%)	2 (2%)
Lived with nonrelatives	0	0	1 (12%)	2 (2%)
Lived alone	0	0	0	4 (5%)

characteristics of the clinical sample are also summarized in Table 1.

The diagnoses for the clinical sample were quite heterogeneous. The most frequent diagnoses for boys were substance abuse (36%), antisocial personality (25%), conduct disorder (14%), and other personality disorders (5%). For girls, the most frequent diagnoses were major depressive disorder (35%), conduct disorder (14%), mood disorder with atypical features (9%), childhood or adolescence disorder not otherwise specified (9%), substance abuse (7%), sexual abuse (5.6%), and mixed anxiety and depressive disorder (5%).

As an incentive to participate in the study, the nonclinical adolescents were told that a prize, consisting of one compact disk, would be given to the adolescent who could be identified as the best faker. Summary results were given to the administration of the school, but results of individual participants were not given. The MMPI-A was administered to the clinical group prior to or during the treatment, and the individual results were given to the person who was in charge of the patient's treatment.

Measures

Biographical Information Form. This form obtained demographic information such as age, grade level, average school grades, school activities, and occupations and educational levels of parents. This form was used in the standardization of the MMPI-A in Mexico and includes some of the items that were used in the standardization of the MMPI-A in the United States.

MMPI-A. The MMPI-A is a self-report measure of personality and psychopathology that consists of 478 true-false items. A sixth-grade reading level is recommended to take the test (Butcher et al., 1992). In this study, the official Mexican version of the MMPI-A was used (Lucio, 1998). The translation procedures followed a strict methodology (Lucio et al., 1999).

Procedures

Test administration. The nonclinical adolescents completed the test twice, first with standard instructions and then under instructions to fake bad. These participants completed the test in groups of 25. The adolescents in the fake-bad condition were given the following instructions, based on those used by Stein et al. (1995) in their study:

> This is the Minnesota Multiphasic Personality Inventory for Adolescents. It is a widely used test for looking at psychological and emotional adjustment. Respond to the items to give the impression that you have very serious emotional problems and that you need treatment where you can talk with a counselor, psychologist or other doctor about your emotional problems. (p. 420)

The clinical adolescents answered the instruments individually or in small groups of 2 or 3 participants with standard instructions.

Scoring and exclusion variables. The MMPI-A raw scores were determined for the three standard validity and 10 clinical scales, as well as for the F1 (Infrequency 1), F2 (Infrequency 2), VRIN (Variable Response Inconsistency), and TRIN (True Response Inconsistency) Scales. Raw scores were transformed to *T*-scores using data provided in the MMPI-A Spanish-version manual (Lucio et al., 1998). Because of the nature of the study, no participants were eliminated on the basis of L (Lie Scale), F, or K scores.

For the nonclinical adolescents, 196 participants were tested in the first session. Of these, 50 were excluded according to the criteria shown in Table 2. Of the 302 clinical adolescents tested, 31 were excluded according to the criteria shown in Table 2. From the remaining 271 clinical adolescents, 89 girls and 57 boys were randomly selected to have the same number of participants as in the nonclinical group. Participants with TRIN or VRIN Scale scores greater than 14 were excluded because they did not respond consistently to the MMPI-A items. TRIN and VRIN Scale cutoff scores greater than 14 were used because the Mexican normative means for these scales are somewhat higher than the U.S. means for these scales.

TABLE 2
Reasons for Exclusion

Frequency	*Criterion*	*Session*
Nonclinical[a]		
1	VRIN greater than 14	First
1	More than 30 CNS	First
36	Did not return	Second
12	VRIN greater than 14	Second
50 (total)		
Clinical[b]		
18	Younger than 14 years old	First
7	More than 30 CNS	First
1	TRIN greater than 14	First
5	Did not specify their age	First
31 (total)		

a. From 196 nonclinical participants tested, 50 were excluded for the reasons shown.
b. From 302 clinical participants tested, 31 were excluded for the reasons shown.

RESULTS

Nonclinical participants who completed the MMPI-A with instructions to appear emotionally disturbed tended to overreport symptoms and difficulties, as indicated by their mean scores in Table 3. This overreporting resulted in much higher F Scale and clinical scale scores than for the clinical sample or for the nonclinical adolescents tested with standard instructions. The validity scale configuration for both boys and girls in the fake-bad condition included very high mean F Scale scores and below average *T*-scores on the K Scale.

Differences between the scores of participants in the fake-bad condition and of the clinical participants, and differences between scores in the fake-bad and standard instruction conditions, were evaluated using *t* tests. As presented in Table 3, girls in the fake-bad condition had significantly higher F Scale scores than girls in the clinical and standard instruction groups and significantly lower K and L Scale scores than girls in the standard instruction group. Girls in the fake-bad condition scored significantly higher than girls in the standard instruction condition for all 10 clinical scales and higher than girls in the clinical group on all clinical scales except Scale 5. It should be noted that differences in mean scores were larger between the fake-bad girls and the standard instruction girls than between the fake-bad girls and the clinical girls.

Very similar results were obtained for boys in the fake-bad condition (see Table 3). However, significant differences were not found between the groups of boys on the L and K Scales or on Scale 5. Boys in the fake-bad group did not score significantly higher on Scale 0 than boys in the standard instruction group.

In addition to comparing mean scores, optimal cutoff scores were also determined to discriminate between participants in the fake-bad and standard instruction conditions. Classification data including hit rates, sensitivity, specificity, PPP, and NPP were calculated for three different base rates. The 50% base rate corresponds to that of the current sample, whereas the 25% and 10% base rates are ones that have been used in other faking studies.

As shown in Table 4, with a 50% base rate, a raw cutoff score of 24 for the girls and of 27 for the boys on the F Scale yielded the highest overall classification accuracy between adolescents in the fake-bad group and adolescents in the standard instructions group. The highest overall classification accuracy between girls and boys in the fake-bad and clinical groups was achieved with F Scale raw score cutoffs of 31 and 23, respectively (see Table 5). Tables 4 and 5 present classification rates for several different cutoff scores in order that clinicians may decide which scores should be used depending on the kinds of errors that are most important to avoid. Classification data for each cutoff score also are presented in Tables 4 and 5 for several other base rates.

Tables 4 and 5 also present classification data for the F-K index. The highest overall accuracy in classifying fake-bad and standard instruction participants was achieved using F-K scores of 13 and 14 for girls and boys, respectively. F-K cutoff scores of 27 for girls and 14 for boys yielded the best overall classification of fake-bad participants and for those who completed the MMPI-A with standard instructions (see Table 5). The new validity scales (F1 and F2) were also successful in detecting fake-bad profiles. However, given that F1 and F2 are subscales of the F Scale and classification accuracy using them was no higher than for the F Scale, classification data are not presented for the F1 and F2 scales.

The PPP and NPP values associated with the F Scale cutoff scores that yielded the best overall classification of fake-bad and standard instruction participants were .927 and .890, respectively, for girls and .980 and .852, respectively, for boys. The optimal F-K index scores yielded PPP and NPP values similar to those for the F Scale for both boys and girls in the fake-bad and standard instruction conditions.

The PPP and NPP values associated with the F Scale cutoff scores that yielded the best overall classification of the fake-bad and clinical conditions were .857 and .783, respectively, for girls and .815 and .886, respectively, for boys. The optimal F-K index scores yielded PPP and NPP values similar to those for the F Scale for both boys and girls in the fake-bad and clinical groups.

TABLE 3
Means, Standard Deviations, and *t*-Test Values for Standard Instructions, Fake-Bad Instructions, and Clinical Participants

	Group/Instructions							
	Nonclinical/Standard Instructions			*Nonclinical/Fake-Bad Instructions*			*Clinical/Standard Instructions*	
Scale	M	SD	t^a	M	SD	t^b	M	SD
Girls (*n* = 87)								
VRIN	52.64	7.81	–4.45*	59.38	11.78	1.80	56.43	9.75
TRIN	53.75	5.26	2.93	56.54	7.15	1.52	58.38	8.64
L	53.38	12.03	3.77*	47.26	9.13	–2.34	50.95	11.49
F	53.84	10.26	–19.82*	102.30	20.27	13.10*	63.00	19.28
K	49.06	10.17	5.89*	40.97	7.78	–0.93	42.16	9.16
Hs	54.75	8.88	–13.02*	75.01	11.48	10.01*	59.87	8.19
D	54.16	10.49	–8.96*	68.36	10.40	6.69*	58.01	10.00
Hy	51.71	9.23	–11.15*	69.87	12.06	7.36*	57.36	10.29
Pd	51.41	9.65	–14.35*	74.47	11.47	6.53*	63.74	10.17
Mf	51.64	9.61	–3.33*	56.72	10.50	2.60	52.57	10.51
Pa	51.56	10.60	–15.96*	82.36	14.54	10.33*	61.48	12.00
Pt	52.48	11.25	–11.77*	72.36	11.03	6.61*	61.29	11.05
Sc	52.99	9.70	–16.86*	83.10	13.55	9.94*	63.91	11.88
Ma	48.40	9.32	–11.96*	66.74	9.85	5.61*	57.33	11.25
Si	54.37	9.27	–7.05*	64.83	10.27	4.91*	57.29	0.08
Boys (*n* = 59)								
VRIN	52.21	7.17	–6.06*	62.02	10.17	4.24*	54.46	9.18
TRIN	53.69	5.51	–3.46*	57.59	6.65	1.39	55.73	8.34
L	52.86	11.01	0.38	52.10	10.31	0.50	51.15	10.34
F	55.03	9.38	–13.90*	93.63	19.14	11.35*	57.66	15.02
K	47.31	10.28	1.04	45.44	9.13	–1.77	48.73	11.01
Hs	55.75	10.68	–10.98*	80.42	13.56	10.79*	55.66	11.26
D	54.42	10.67	–8.24*	71.17	11.40	8.09*	55.20	9.98
Hy	51.29	9.77	–10.60*	72.86	12.20	9.40*	53.41	10.17
Pd	52.68	9.84	–9.51*	71.83	11.94	5.94*	59.66	10.22
Mf	48.86	9.20	–3.20	54.47	9.81	3.18	49.14	8.38
Pa	53.68	10.11	–10.61*	77.68	14.14	9.45*	56.10	10.24
Pt	55.75	10.71	–6.16*	67.34	9.70	5.82*	55.85	11.65
Sc	56.15	9.88	–10.49*	77.22	11.83	9.30*	57.69	10.95
Ma	50.81	9.53	–6.47*	64.68	13.40	4.99*	54.10	9.22
Si	57.31	9.16	–1.63	60.02	8.82	4.73*	52.27	8.97

NOTE: Hs = Hypocondria; D = Depression; Hy = Hysteria; Pd = Psychopathic Deviate; Mf = Masculinity-Femininity; Pa = Paranoia; Pt = Psychasthenia; Sc = Schizophrenia; Ma = Hypomania; Si = Social Introversion.
a. *t* = test between fake-bad and standard conditions.
b. *t* = test between fake-bad and clinical participants.
**p* < .001.

The effects of varying base rates are apparent in Tables 4 and 5. For the fake-bad versus standard instruction condition, higher optimal cutoff scores are needed for both the F Scale and the F-K index as the base rate decreases. However, with these higher cutoff scores, comparable PPP and NPP rates were obtained for the lower base rates. For the fake-bad versus clinical groups, similar cutoff scores for the F Scale and F-K index yielded comparable overall classification rates for the higher and lower base rate conditions, but the PPP values were markedly lower at the lower base rates.

DISCUSSION

The results of this study indicate that the validity scales and indexes of the MMPI-A (F, F1, F2, and F-K) can accurately detect adolescents who complete the test with instructions to fake bad. Scores differed significantly for adolescents who completed the test with instructions to fake bad and clinical and nonclinical adolescents who completed the test with standard instructions. The mean scores for girls and boys in the fake-bad condition were similar to previously reported data for adolescents in the

TABLE 4
Classification Data for Fake-Bad Nonclinical Versus Standard Nonclinical

			Base Rate 10%			*Base Rate 25%*			*Base Rate 50%*		
Cutoff	*Sensitivity*	*Specificity*	*PPP*	*NPP*	*Hit Rate*	*PPP*	*NPP*	*Hit Rate*	*PPP*	*NPP*	*Hit Rate*
Girls											
F Scale											
31	75.8	100.0	1.00	97.3	97.5	100.0	92.2	93.7	100.0	80.5	87.9
30	78.1	98.8	88.3	97.6	96.7	95.9	92.8	93.5	98.5	81.9	88.5
29	80.4	98.8	88.6	97.8	97.0	96.0	93.5	94.0	98.5	83.5	89.6
28	80.4	97.7	79.5	97.8	95.9	92.4	93.4	93.2	97.2	83.3	89.0
27	80.4	97.7	79.5	97.8	95.9	92.4	93.4	93.2	97.2	83.3	89.0
26	81.6	97.7	79.7	97.9	96.0	92.5	93.8	93.5	97.2	84.1	89.6
25	85.0	94.2	62.1	98.2	93.3	92.5	94.7	91.8	93.6	86.3	89.6
24	88.5	93.1	58.7	98.6	93.6	83.7	95.8	91.9	92.7	89.0	90.8
F-K index											
22	75.8	98.8	88.0	97.3	96.5	95.8	92.1	92.9	98.5	80.3	87.3
21	77.0	98.8	88.1	97.4	96.6	95.9	92.4	93.2	98.5	81.1	87.9
20	78.1	98.8	88.3	97.6	96.7	95.9	92.8	93.5	98.5	81.9	88.5
19	78.1	97.7	79.0	97.5	95.7	92.2	92.7	92.6	97.1	81.7	87.9
18	79.3	97.7	79.3	97.7	95.8	92.3	93.1	92.9	97.1	82.5	88.5
17	81.6	96.5	72.4	97.9	95.0	89.2	93.7	92.6	95.9	84.0	89.0
16	82.7	94.2	61.5	98.0	93.1	83.4	94.0	91.2	93.5	84.5	88.5
15	82.7	91.9	53.3	97.9	91.0	78.2	93.8	89.5	91.1	84.2	87.3
14	83.9	91.9	53.6	98.0	91.1	78.4	94.2	89.8	91.2	85.1	87.9
13	88.5	91.9	55.0	98.6	91.6	79.3	95.8	91.0	91.6	88.8	90.2
Boys											
F Scale											
31	71.1	100.0	100.0	96.9	97.1	100.0	90.8	92.5	100.0	77.6	85.5
30	71.1	98.3	82.3	96.8	95.5	93.6	90.7	91.2	97.6	77.3	84.7
29	72.8	98.3	82.6	97.0	95.7	93.7	91.2	91.7	97.7	78.3	85.5
28	77.9	98.3	83.6	97.5	96.2	94.1	92.7	93.0	97.8	81.6	88.1
27	83.0	98.3	84.4	98.1	96.7	94.4	94.3	93.3	98.0	85.2	90.6
26	83.0	96.6	73.1	98.0	95.2	89.5	94.2	93.1	96.0	85.0	89.8
F-K index											
20	66.10	98.3	81.2	96.3	95.0	93.1	89.2	89.9	97.5	74.3	82.2
19	67.80	96.6	68.9	96.4	93.7	87.4	89.5	89.1	95.2	75.0	82.2
18	69.4	96.6	69.4	96.6	93.9	87.7	90.0	89.5	95.3	76.0	83.0
17	72.8	93.2	54.4	96.8	91.1	78.9	90.7	87.1	91.4	77.4	83.0
16	76.2	93.2	55.5	97.2	91.5	79.7	91.8	88.8	91.8	79.7	84.7
15	77.9	93.2	56.1	97.4	91.6	80.0	92.3	89.2	92.0	80.8	85.5
14	79.6	91.5	51.0	97.5	90.3	76.6	92.8	88.4	90.3	81.8	85.5
13	79.6	88.1	42.7	97.5	87.2	70.0	92.5	85.9	87.0	81.2	83.9

NOTE: PPP = positive predictive power; NPP = negative predictive power.

United States (Stein et al., 1995) and also for adults who were instructed to fake bad on the MMPI-2 in Mexico (Lucio & Valencia, 1997). The results of this study are also similar to those reported by other authors using the original form of the MMPI with adolescents in the United States (Archer et al., 1987; Herkov et al., 1991).

The F Scale and the F-K index were about equally effective in identifying fake-bad protocols. Somewhat different optimal cutoff scores for the F Scale and F-K Scale were identified for boys and girls and for the nonclinical and clinical adolescents completing the test with standard instructions. Higher optimal cutoff scores were required for boys than for girls for discriminating the standard and fake-bad conditions for nonclinical adolescents, and the opposite was true for discriminating the fake-bad and clinical groups. In addition, higher cutoff scores were required for optimal discrimination between fake-bad and standard condition groups for base rates lower than that in the present study (50%).

Although Stein et al. (1995) found lower cutoff scores than in the current study, the authors indicated that the scores that they reported seemed to be atypically low. The higher cutoff scores in the current study could also indicate cultural differences between Mexican and U.S. adolescents. This factor should be taken into account when Hispanic adolescents are tested in the United States. It is

TABLE 5
Classification Data for Fake-Bad Versus Clinical Groups

			Base Rate 10%			*Base Rate 25%*			*Base Rate 50%*		
Cutoff	*Sensitivity*	*Specificity*	*PPP*	*NPP*	*Hit Rate*	*PPP*	*NPP*	*Hit Rate*	*PPP*	*NPP*	*Hit Rate*
Girls											
F Scale											
41	60.9	95.4	59.5	95.6	91.9	82.2	87.4	86.4	92.9	70.9	78.1
40	65.5	95.4	61.2	96.1	92.4	83.2	88.8	87.6	93.4	73.4	80.4
39	66.6	94.2	56.3	96.2	91.4	80.1	89.0	87.1	92.0	73.8	80.4
38	68.9	93.1	52.6	96.4	90.6	77.7	89.5	86.8	90.9	75.0	81.0
37	68.9	91.9	48.7	96.3	89.6	74.9	89.4	86.0	89.5	74.7	80.4
36	70.1	91.9	49.1	96.5	89.7	75.2	89.8	86.3	89.7	75.4	81.0
35	70.1	91.9	49.1	96.5	89.7	75.2	89.8	86.3	89.7	75.4	81.0
34	70.1	90.8	45.8	96.4	88.7	72.6	89.7	85.4	88.4	75.2	80.4
33	73.5	89.6	44.1	96.8	88.0	71.2	90.6	85.4	87.6	77.2	81.6
32	74.7	87.3	39.6	96.8	86.0	67.3	90.8	84.0	85.5	77.5	81.0
31	75.8	87.3	40.0	97.0	86.2	67.6	91.2	84.3	85.7	78.3	81.6
F-K index											
31	62.0	95.4	60.0	95.7	92.0	82.4	87.8	86.7	93.1	71.5	78.7
30	64.3	95.4	60.8	96.0	92.3	83.0	88.4	87.3	93.3	72.8	79.8
29	66.6	91.9	47.9	96.1	89.4	74.3	88.7	85.4	89.2	73.3	79.3
28	67.8	90.8	45.0	96.2	88.5	72.0	88.9	84.8	88.0	73.8	79.3
27	68.9	90.8	45.4	96.3	88.6	72.3	89.3	85.1	88.2	74.5	79.8
Boys											
F Scale											
31-30	71.1	91.5	48.2	96.6	81.4	74.5	90.1	86.2	89.3	76.0	81.3
29	72.8	89.8	44.3	96.7	88.1	71.4	90.4	85.4	87.7	76.8	81.3
28	77.9	89.8	46.0	97.3	88.6	72.7	92.1	86.7	88.4	80.3	83.9
27	83.0	88.1	43.7	97.9	87.6	70.9	93.7	86.8	87.5	83.8	85.5
26	83.0	84.7	37.6	97.8	84.5	65.5	93.4	84.3	84.4	83.3	83.9
25	83.0	84.7	37.6	97.8	84.5	65.5	93.4	84.3	84.4	83.3	83.9
24	84.7	79.6	31.6	97.9	80.1	59.2	93.7	80.9	80.6	83.9	82.2
23	89.8	79.6	32.9	98.6	80.6	60.6	95.7	82.2	81.5	88.6	84.7
F-K index											
20	66.1	89.8	41.9	95.9	87.4	69.4	88.3	83.6	86.6	72.6	77.9
19	67.8	89.8	42.5	96.1	87.6	69.9	88.8	84.1	86.9	73.6	78.8
18	69.4	88.1	39.4	96.3	86.2	67.1	89.2	83.3	85.4	74.2	78.8
17	72.8	88.1	40.5	96.6	86.6	68.1	90.3	84.1	86.0	76.4	80.5
16	72.6	88.1	41.6	97.1	86.9	69.1	91.4	85.0	86.5	78.7	82.2
15	77.9	88.1	42.2	97.3	87.1	69.6	91.9	85.5	86.7	80.0	83.0
14	79.6	88.1	42.7	97.5	87.2	70.0	92.5	85.9	87.0	81.2	83.9
13	79.6	86.4	39.5	97.4	85.7	67.2	92.4	84.6	85.4	80.9	83.0
12	83.0	77.9	29.5	97.6	78.4	56.8	92.9	79.2	79.0	82.1	80.5

NOTE: PPP = positive predictive power; NPP = negative predictive power.

possible that clinical Mexican adolescents are more likely to exaggerate their symptoms or admit them more openly than Hispanic adolescents living in the United States. Thus, one should be very cautious in generalizing the findings of the current study to Hispanic adolescents in the United States.

Optimal cutoff scores identified in this study, and the classification data associated with them, should be treated as tentative. Sample sizes in this study, for example, did not permit cross-validation of classification results. Differences in optimal cutoff scores for boys and girls and for clinical and nonclinical comparisons suggest that different cutoff scores may be needed in various settings. In the Mexican fake-bad study with the MMPI-2 (Lucio & Valencia, 1997), higher scores were also needed to discriminate between the fake-bad and clinical groups of men, whereas in this study with adolescents, higher scores were needed to discriminate between the groups of girls. This fact could be related to the kinds of psychopathology presented by participants. Other studies with larger samples and perhaps subgroups of homogeneous psychopathology must be carried out to clarify these results. In conclusion, however, the validity scales and indexes of the MMPI-A (F, F1, F2, and F-K) were successful in discriminating between nonclinical adolescents who were faking bad, nonclinical adolescents who took the test under standard instructions, and clinical patients. Specific recommendations for clinical applications of our findings must wait additional research with larger samples and in various clinical settings.

REFERENCES

Adelman, R. M., & Howard, A. (1984). Expert testimony on malingering: The admissibility of clinical procedures for the detection of deception. *Behavioral Sciences and the Law*, *2*, 5-19.

Archer, R. P., Gordon, R. A., & Kirchner, F. H. (1987). MMPI response set characteristics among adolescents. *Journal of Personality Assessment*, *51*, 506-516.

Bagby, R. M., Nicholson, R. A., Buis, T., & Bacchiochi, J. R. (2000). Can the MMPI-2 validity scales detect depression feigned by experts? *Assessment*, *7*(1), 55-62.

Berry, D.T.R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review*, *11*, 585-598.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R., Tellegen, A., Ben-Porath, Y. S., et al. (1992). *Minnesota Multiphasic Personality Inventory–Adolescent (MMPI-A): Manual for administration, scoring and interpretation*. Minneapolis: University of Minnesota Press.

Cramer, K. M. (1995). The effects of description clarity and disorder type on MMPI-2 fake bad validity indexes. *Journal of Clinical Psychology*, *51*, 831-840.

Herkov, M. J., Archer, R. P., & Gordon, R. A. (1991). MMPI response sets among adolescents: An evaluation of the limitations of the Subtle-Obvious subscales. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*, 424-426.

Lim, J., & Butcher, J. N. (1996). Detection of faking on the MMPI-2: Differentiation among faking bad, denial, and claiming extreme virtue. *Journal of Personality Assessment*, *67*, 1-25.

Lucio, E. (1998). *Spanish version of the Minnesota Multiphasic Personality Inventory: MMPI-A for Mexico*. Mexico City, Mexico: El Manual Moderno.

Lucio, E., Ampudia, A., & Duran, C. (1998). *Manual para la aplicacion y calificacion del MMPI-A. Traduccion y adaptacion al Español* [Spanish translation and adaptation of the MMPI-A manual for administration and scoring]. Mexico City, Mexico: El Manual Moderno.

Lucio, E., Ampudia, A., Duran, C., Gallegos, L., & Leon, I. (1999). La nueva version del Inventario Multifasico de la Personalidad de Minnesota para Adolescents: MMPI-A [The new version of the Minnesota Multiphasic Personality Inventory for Mexican adolescents]. *Revista Mexicana de Psicología*, *16*, 217-226.

Lucio, E., & Valencia, R. (1997). Detección de perfiles de sujetos simuladores y sujetos honestos a traves de las escalas del MMPI-2 [Detection of fakers and honest subjects through MMPI-2 scales]. *Revista Salud Mental*, *20*, 23-33.

Medoff, D. (1999). MMPI-2 validity scales in child custody evaluations: Clinical versus statistical significance. *Behavioral Sciences and the Law*, *17*, 409-411.

Resnick, P. J. (1984). The detection of malingered mental illness. *Behavioral Sciences and the Law*, *2*, 21-28.

Rogers, R., Hinds, J. D., & Sewell, K. W. (1996). Feigning psychopathology among adolescent offenders: Validation of the SIRS, MMPI-A and SIMS. *Journal of Personality Assessment*, *67*, 244-257.

Rogers, R., Sewell, K. W., & Salekin, R. T. (1994). A meta-analysis of malingering on the MMPI-2. *Assessment*, *1*, 227-237.

Stein, L.A.R., Graham, J. R., & Williams, C. L. (1995). Detecting fake-bad MMPI-A profiles. *Journal of Personality Assessment*, *65*, 415-427.

Emilia Lucio is professor of psychology in the Graduate Studies Department in the School of Psychology at the National Autonomous University of Mexico. She conducted the standardization of the MMPI–A in Mexico and works in assessment and intervention with adolescents.

Consuelo Durán is research associate in the Graduate Studies Department of the School of Psychology at the National Autonomous University of Mexico. She collaborated in standardizing the MMPI–A and MMPI–2 in Mexico, has published several articles, and has provided methodological assistance to various undergraduate and graduate theses and dissertations.

John R. Graham is professor of psychology in the Department of Psychology at Kent State University. He has published several books, chapters, and numerous empirical articles concerning the MMPI, MMPI–2, and MMPI–A.

Yossef S. Ben-Porath is professor of psychology in the Department of Psychology at Kent State University. He has published numerous empirical articles concerning the MMPI–2 and the MMPI–A, and also concerning adolescence.

Measures of Self-Efficacy and Optimism in Older Adults With Generalized Anxiety

Melinda A. Stanley
Diane M. Novy
Derek R. Hopko
University of Texas–Houston Medical School

J. Gayle Beck
State University of New York at Buffalo

Patricia M. Averill
Alan C. Swann
University of Texas–Houston Medical School

This study provides initial psychometric data for the Self-Efficacy Scale (SES) and the Life Orientation Test (LOT) in a sample of older adults with generalized anxiety disorder (GAD). Participants included 76 adults, ages 60 to 80, who met Diagnostic and Statistical Manual of Mental Disorders *(4th ed.) criteria for GAD. Self-efficacy and outcome expectancies were lower in older adults with GAD relative to published data from younger and older community samples. Both the SES and LOT demonstrated adequate internal consistency. Confirmatory factor analysis provided evidence for optimism and pessimism factors within the LOT, and exploratory factor analysis of the SES suggested three factors that overlap with previous findings. Overall, the data support the potential utility of these instruments in late-life GAD and set the stage for future investigations of generalized self-efficacy expectancies and outcome expectancies (or optimism) as they relate to the prediction of affect and behavior in this group.*

Keywords: self-efficacy, optimism, anxiety, elderly

A large body of literature has addressed the impact of self-efficacy expectancies and outcome expectancies in the prediction of affective states, health status, and behavior across a range of contexts (Bandura, 1997; Maddux, 1995). Self-efficacy expectancies are characterized as beliefs about one's ability to carry out specific behaviors or behavioral sequences, whereas outcome expectancies are defined as beliefs about what outcomes will occur following certain behaviors. The potentially central role of these cognitive variables was posited initially by Bandura (1977), who maintained that although self-efficacy and outcome expectancies are conceptually distinct, the former generally accounts for the majority of variance in the latter. Indeed, within the self-efficacy literature, high intercorrelations between measures of self-efficacy and outcome expectancies are documented, with generally stronger relationships of self-efficacy to behavior, intentions, and affect.

Although the majority of self-efficacy research has been conducted with young and middle-age adults (see re-

This work was supported in part by a grant from the National Institute of Mental Health (RO1-MH53932) awarded to the first author. The assistance of Dr. James Maddux and two anonymous reviewers, who provided comments on earlier versions of this article, is appreciated. Please address all correspondence to Melinda A. Stanley, Ph.D., 1300 Moursund Street, University of Texas–Houston Medical School, Houston, TX 77030; phone: 713-500-2600; fax: 713-500-2530; e-mail: Melinda.A.Stanley@uth.tmc.edu.

Assessment, Volume 9, No. 1, March 2002 70-81

views in Bandura, 1997, and Maddux, 1995), some data have documented correlations of expectancies with affective and behavioral responses in older individuals. In this age group, perceptions of efficacy and control may be even more important predictors of behavior, health, and adjustment than for younger individuals (e.g., Rodin, 1986; Welch & West, 1995). In fact, research with older adults has indicated significant relations between expectancies and health risk status (Grembowski et al., 1993), fears of falling and associated behaviors (Tinetti, Richman, & Powell, 1990), changes in functional status (Mendes de Leon, Seeman, Baker, Richardson, & Tinetti, 1996), cognitive performance (Seeman, Rodin, & Albert, 1993), and various measures of adjustment, physical and health status, and affect (Waller & Bates, 1992).

Missing from this literature, however, is attention to the roles of expectancies as they pertain to the phenomenology, assessment, and treatment of psychiatric disorders in older individuals. Of particular interest are the predictive utility and theoretical significance of self-efficacy and outcome beliefs for understanding anxiety in late life. The significant relation of self-efficacy beliefs to anxiety and related behaviors among younger adults already has been demonstrated (Williams, 1995). However, no data yet have examined similar relations among older anxious adults, despite relatively high prevalence rates for anxiety disorders and associated disability in this age group (Beekman et al., 1998; de Beurs et al., 1999; Weissman et al., 1985). Of the anxiety disorders diagnosed among older adults, generalized anxiety disorder (GAD) is one of the most common (Beekman et al., 1998; Blazer, George, & Hughes, 1991), but research has only begun to elucidate the nature and treatment of this disorder in later life (J. G. Beck, Stanley, & Zebb, 1996; Stanley, Beck, & Glassco, 1996; Stanley et al., 1999). The ability of self-efficacy theory to enhance understanding of potential cognitive mechanisms and treatment of this disorder has not been examined.

The first step in initiating this line of research is to establish measures of self-efficacy and outcome expectancies with adequate psychometric properties among older adults. However, the measurement of these constructs is a controversial and complicated process. Most theorists and researchers in the field (e.g., Bandura, 1997; Maddux, 1995) argue for problem-specific measures of self-efficacy and outcome expectancies and against the utility of more global, traitlike constructs such as "generalized" self-efficacy or outcome expectancies. However, other authors have suggested the potential predictive utility of a more generalized set of expectancies that develop from a history of successes and failures across a variety of situations (Sherer et al., 1982; Tipton & Worthington, 1984). Even problem-focused efficacy researchers agree that efficacy judgments generalize across similar situations (Maddux & Gosselin, in press), and more global constructs such as generalized self-efficacy or generalized outcome expectancies may be particularly predictive of behavior and affect for people with pervasive anxiety problems such as GAD. People with GAD report a range of worries that often reflect consistent themes (Diefenbach, Stanley, & Beck, 2001) but that are not characterized by a consistent set of circumscribed feared stimuli and associated behaviors or performance, as is the case with other anxiety disorders such as panic disorder and phobias for which problem-focused efficacy judgments may play a stronger role (Barlow, 1988). Cognitive conceptualizations of GAD also emphasize the role of generalized perceptions and beliefs (Dugas, Gagnon, Ladouceur, & Freeston, 1998). Moreover, assessment of problem-specific expectancies in GAD would require idiographic measures that disallow meaningful aggregate data and comparisons across different patient samples. Consequently, more general measures of self-efficacy and outcome expectancies may be useful in the study of GAD, despite the potential limitations of general measures.

The Self-Efficacy Scale (SES) (Sherer et al., 1982) was developed as a measure of generalized self-efficacy. The instrument has demonstrated utility among younger adults in the prediction of academic and occupational behavior (Ferrari, Parker, & Ware, 1992; Sherer et al., 1982), self-esteem (Woodruff & Cashman, 1993), and general adjustment (Martin, Flett, Hewitt, Krames, & Szanto, 1996; Sherer & Adams, 1983). The SES also has been used to measure generalized self-efficacy in older adults (Davis-Berman, 1990; Smits, Deeg, & Bosscher, 1995; Waller & Bates, 1992), with evidence of a moderate relation between generalized efficacy and health-related behavior (Waller & Bates, 1992) and lack of relations between generalized efficacy and potentially related constructs such as mastery and health locus of control (Smits et al., 1995). Evidence of psychometric support for the SES in older adults is limited, however, and no data have assessed the role of self-efficacy expectancies in patients with psychiatric disorders, particularly GAD wherein generalized self-efficacy may play a unique role in the prediction of psychopathology and treatment response.

The Life Orientation Test (LOT) (Scheier & Carver, 1985) is a measure of dispositional optimism that also can be considered a measure of generalized outcome expectancies given the focus on evaluating general inclinations to expect positive life outcomes (Scheier & Carver, 1985, 1993). The LOT has adequate internal consistency, predictive validity, and consistent evidence of a two-factor model (optimism, pessimism) in younger and older community samples (e.g., Marshall & Lang, 1990; Mroczek, Spiro, Aldwin, Ozer, & Bosse, 1993; Robinson-Whelen, Kim, MacCallum, & Kiecolt-Glaser, 1997; Sharpe, Hickey, &

Wolf, 1994). Data from previous research have suggested differential relations of these two factors to mood and personality variables (Marshall, Wortman, Kusulas, Hervig, & Vickers, 1992), as well as predictability of coping and adjustment, even with control of related constructs such as neuroticism and negative affectivity (Mroczek et al., 1993; Robinson-Whelen et al., 1997). However, the psychometric properties of the LOT, including potential validity of two subscales, have not heretofore been investigated among older adults with psychiatric disorders, in particular GAD.

The purpose of this article is to provide initial psychometric data for the SES and the LOT in a sample of older adults with GAD. The study will set the stage for the use of these measures to investigate the role of self-efficacy and outcome expectancies in the psychopathology and treatment of GAD in later life. The data focused on descriptive characteristics, internal consistency, and construct validity. Construct validity was investigated by examination of scale and subscale intercorrelations, factor analyses, and correlations of subscale scores with measures of negative affect and social anxiety.

METHOD

Participants

Participants included 76 adults, ages 60 to 80, who met *Diagnostic and Statistical Manual of Mental Disorders* criteria for GAD (American Psychiatric Association, 1994). Participants were a subsample of 85 individuals evaluated at pretreatment in the context of a clinical trial of cognitive behavior therapy for GAD in older adults. Three of the original 85 participants dropped out of the study after completion of the diagnostic interviews, before all pretreatment assessments were completed. Consequently, self-report data for these patients were not available. Six additional participants also failed to complete the LOT and SES at pretreatment, resulting in a complete data set for 76 individuals.

Participants were recruited through the community, with media announcements and visits to community agencies and church groups that focused on older adults. Interested individuals were screened initially by telephone. Those participants whose symptoms appeared consistent with diagnostic criteria for GAD were evaluated on two separate occasions using the Anxiety Disorders Interview Schedule for *DSM-IV* (ADIS-IV) (Brown, DiNardo, & Barlow, 1994). Postdoctoral fellows, predoctoral interns, and advanced psychology graduate students conducted these interviews. All interviewers had extensive training and experience in the administration of the ADIS-IV. For each participant, the two interviews were separated by at least a 2-week interval (*M* = 20.5 days, *SD* = 9.07). In each case, the second evaluator was unaware of the diagnoses assigned during the initial interview. Following both interviews, patients were discussed at a weekly staff meeting, and a consensus diagnosis was reached. Overall diagnostic agreement for the two interviews was 92%.

As reported in previous studies with subgroups of this sample (Bourland et al., 2000; Stanley, Novy, Bourland, Beck, & Averill, 2001), patients were asked to discontinue any use of psychotropic medications (under the supervision of the prescribing physician) at least 2 weeks prior to the first diagnostic interview. This procedure was established to enhance internal validity of the treatment outcome study. Exceptions were made for participants taking sedative-hypnotic medication for sleep difficulties less than four times per week, low levels of psychotropic medication for pain management, or beta-blockers due to a heart condition. Other exclusion criteria included a history of psychotic symptoms or organic brain syndrome, alcohol or substance abuse within the previous year, dementia as defined by a score of 24 or lower on the Mini–Mental State Examination (Folstein, Folstein, & McHugh, 1975), current involvement in psychotherapy, or the presence of medical conditions that could account for anxiety symptoms or interfere with treatment (e.g., recent stroke, acute cardiac disease or arrhythmias in need of treatment, Parkinson’s disease, hypoglycemia, untreated thyroid or other endocrine disorders, or untreated hypertension). Summary statistics for excluded participants are described in another report (Akkerman et al., in press). It should be noted here that all potential participants received a complete medical evaluation at pretreatment, and none were excluded for medical reasons.

For all included participants, GAD was a principal (*n* = 58, 76%) or coprincipal diagnosis (*n* = 18, 24%) with a global severity rating of at least 4 (*moderate severity*) on a 0 to 8 Likert-type scale. Principal diagnosis was defined as the disorder that was assigned the highest severity rating. When two diagnoses met this criterion, coprincipal diagnoses were assigned. Mean severity of GAD was 5.2 (*SD* = 0.87). Coexistent diagnoses, including those identified as coprincipal as well as those with lesser severity ratings, were as follows: 26% major depressive disorder (*n* = 20), 25% social phobia (*n* = 19), 17% specific phobia (*n* = 13), 9% depressive disorder not otherwise specified (*n* = 7), 5% dysthymic disorder (*n* = 4), 7% panic disorder with or without agoraphobia (*n* = 6), and 1% (*n* = 1) each for obsessive-compulsive disorder, post-traumatic stress disorder, agoraphobia, hypochondriasis, and adjustment disorder with depressed mood. Forty-nine patients (64%) had at least one coexistent disorder.

Mean age of the participants was 66.6 years (*SD* = 5.30), and 76% (*n* = 58) were women. The sample was well educated (M = 14.8 years, SD = 2.66), and ethnic distribution was as follows: 87% White (n = 66), 7% Black (n = 5), 4% Hispanic (n = 3), and 3% Asian (n = 2). Fifty percent of the sample was married, 25% widowed, 24% divorced, and 1% single. Most participants were retired (50%), 35% were employed either part- or full-time, 11% were homemakers, and 1% reported being unemployed.

Measures

All participants completed measures of self-efficacy, optimism, negative affect (worry, anxiety, depression, neuroticism), and social anxiety.

Self-efficacy. The SES is a 23-item instrument developed to assess generalized self-efficacy. The scale contains 7 additional filler items that are not scored. Initial validation studies with young adult samples demonstrated adequate internal consistency and construct validity for two subscales, General Self-Efficacy (17 items) and Social Self-Efficacy (6 items) (Sherer & Adams, 1983; Sherer et al., 1982). A subsequent report suggested an alternative five-factor model for the instrument (Woodruff & Cashman, 1993), with three general self-efficacy and two social self-efficacy factors. Measures of internal consistency for these factors, however, were not consistently high (.57-.75). Among older adults, descriptive data for the SES are difficult to compare across studies given the use of different versions of the scale (e.g., 12 versus 17 items to assess general self-efficacy) and differential scoring (e.g., yes-no rather than Likert-type format [Davis-Berman, 1990] and 14- rather than 5-point Likert-type scale [Waller & Bates, 1992]). In addition, very few formal psychometric analyses have been conducted in samples of older adults. Internal consistency for a 12-item version of the General Self-Efficacy subscale was poor (alpha = .61) (Smits et al., 1995), although factor analysis of data from this sample (Bosscher & Smit, 1998) showed some overlap with the three-factor general self-efficacy model proposed by Woodruff and Cashman (1993). Factors appeared to assess initiative, effort, and persistence, but internal consistency for the factors was low (.63-.64).

Optimism. The LOT is an eight-item instrument developed to assess dispositional optimism. Four items are positively worded and four are negatively worded, and the scale contains four additional filler items that are not scored. Items are rated on a 5-point Likert-type scale. As noted earlier, studies have demonstrated adequate psychometric properties for the LOT in samples of younger and older adults, with support for the potential utility of Optimism and Pessimism subscale scores (Marshall & Lang, 1990; Mroczek et al., 1993; Robinson-Whelen et al., 1997; Scheier & Carver, 1985; Sharpe et al., 1994). As in Robinson-Whelen et al. (1997), the four negatively worded items were reverse scored to create a total LOT score, although these items were summed as rated to create the Pessimism subscale score. This allowed easier interpretation of Pessimism scores such that higher scores indicated greater pessimism.

Negative affect. The Penn State Worry Questionnaire (PSWQ) (Meyer, Miller, Metzger, & Borkovec, 1990) is a 16-item scale that assesses an individual's tendency to worry, with items rated on a 1 to 5 scale. The scale has good internal consistency and adequate convergent validity in older adults with GAD and those without psychiatric diagnoses (J. G. Beck, Stanley, & Zebb, 1995; Stanley et al., 2001). Test-retest reliability in an older GAD sample suggested some temporal instability in PSWQ scores (r = .54), although this finding may be explained partially by the variable range of the test-retest interval and the intervening assessment tasks (Stanley et al., 2001).

The Spielberger State-Trait Anxiety Inventory–Trait Scale (STAI-T) (Spielberger, 1983) is composed of 20 items rated on a 1 to 4 scale of severity. Psychometric data suggest adequate internal consistency and construct validity in heterogeneous samples of older adults (Himmelfarb & Murrell, 1983; Patterson, Sullivan, & Spielberger, 1980), older adult outpatients with a variety of psychiatric disorders (Kabacoff, Segal, Hersen, & Van Hasselt, 1997), and those with well-diagnosed GAD (Stanley, Beck, & Zebb, 1996; Stanley et al., 2001). Test-retest reliability in a normal control sample of older adults also was adequate (Stanley, Beck, & Zebb, 1996), although scores were not stable over time in a GAD sample (r = .58) (Stanley et al., 2001).

The Neuroticism Scale of the revised NEO Personality Inventory (NEO-PI-R) (Costa & McCrae, 1992) is a 48-item self-report measure designed to assess the tendency to experience negative affective states such as fear, sadness, embarrassment, anger, guilt, and disgust. Adequate psychometric properties for the NEO-PI-R have been reported among nonclinical (Costa & McCrae, 1992) and clinical (Fagan et al., 1992) samples of adults. Although psychometric properties have not been examined in older adults, the instrument has been used to investigate personality characteristics in this age group (Allard & Mishara, 1995).

The Beck Depression Inventory (BDI) (A. T. Beck & Steer, 1987) includes 21 items, each of which is rated on a 4-point Likert-type scale. Considerable data have indicated that the BDI has adequate psychometric properties in depressed and normal older adults (Gallagher, Breckenridge, Steinmetz, & Thompson, 1983; Gallagher, Nies, & Thompson, 1982) as well as those with GAD (Snyder, Stanley, Novy, Averill, & Beck, 2000).

Social anxiety. Social anxiety was assessed with two measures, the Extraversion Scale of the NEO-PI-R and the Social Phobia subscale from a revised version of the Fear Questionnaire (FQ) (Marks & Mathews, 1979). The NEO-PI-R Extraversion Scale is a 48-item measure that assesses general sociability and the tendency to be assertive, active, and enthusiastic in social situations. As noted previously, psychometric support for the NEO-PI-R is adequate in younger adults, and the measure also has been used with older adult samples.

In the original FQ, five items each assess degree of avoidance associated with social, agoraphobic, and blood-injury fears. Each item is rated on a 0- to 8-point scale. An initial psychometric study in older adults with and without GAD demonstrated problems with internal consistency and test-retest reliability for the original FQ subscales (Stanley, Beck, & Zebb, 1996). More recent data have demonstrated improved internal consistency and test-retest reliability for the subscales of a modified version of the FQ that focuses on level of subjective fear rather than degree of avoidance (Stanley et al., 2001). Thus, the Social Phobia subscale of the revised FQ was used here to assess social anxiety.

Procedure

All participants provided informed consent and completed the measures described as part of a broader pretreatment assessment within the context of a clinical trial.

RESULTS

Descriptive Data

Means and standard deviations for the LOT and SES total and subscale scores, as well as clinical measures, are presented in Table 1. Scores on clinical scales were comparable to data from other samples of older adults with GAD (Stanley, Beck, & Glassco, 1996; Stanley, Beck, & Zebb, 1996; Wetherell, Gatz, & Craske, 2001). *t* tests were conducted to examine differences in LOT and SES scores based on gender, marital status (married vs. not married), occupational status (employed vs. not employed), and ethnicity (Caucasian vs. non-Caucasian). Bonferroni correction procedures were used to control for experiment-wise error rates within each set of comparisons, with critical alpha set at .017 (.05/3). Results of these analyses revealed no significant relation of gender, marital or occupational status, or ethnicity to any LOT or SES score. Pearson's *r* correlation coefficients also revealed no significant relations of optimism/pessimism and self-efficacy expectancies with age or education.

TABLE 1
Means, Standard Deviations, and Coefficient Alphas for LOT and SES Total and Subscale Scores in a Sample of Older Adults With GAD

	M *(SD)*	*Alpha*	*Item-Remainder Correlations*
LOT			
Total	16.1 (4.59)	.80	.36-.60
Optimism	7.1 (2.72)	.75	.44-.64
Pessimism	7.0 (2.73)	.78	.34-.76
SES			
Total	72.3 (13.61)	.91	.25-.71
General	54.4 (11.09)	.91	.34-.80
Social	18.0 (4.17)	.78	.43-.66
Negative affect			
PSWQ	61.9 (9.60)		
STAI-T	54.7 (9.24)		
NEO-PI-R Neuroticism	63.1 (8.47)		
BDI	17.9 (6.69)		
NEO-PI-R Extraversion	43.3 (9.95)		
FQ Social Phobia	11.1 (6.60)		

NOTE: LOT = Life Orientation Test; SES = Self-Efficacy Scale; GAD = generalized anxiety disorder; PSWQ = Penn State Worry Questionnaire; STAI-T = State-Trait Anxiety Inventory–Trait Scale; NEO-PI-R = revised Personality Inventory of the NEO; BDI = Beck Depression Inventory; FQ = Fear Questionnaire.

z score comparisons of means demonstrated that LOT total scores in the older adult GAD sample were significantly lower, indicating less optimism, than those from a community sample of older men (mean age = 60 years; LOT = 21.2, *SD* = 4.3 [Mroczek et al., 1993]; *z* = –10.32, *p* < .001), a sample of professional women (mean age = 35 years; LOT = 23.8, *SD* = 3.44 [Marshall & Lang, 1990]; *z* = 13.3, *p* < .001), and undergraduate students (Scheier & Carver, 1985) (men: LOT = 21.0, *SD* = 4.56, *z* = –4.71, *p* < .001; women: LOT = 21.4, *SD* = 5.22, *z* = –7.54, *p* < .001). Analysis of LOT subscale scores showed lower levels of optimism and greater pessimism among older patients with GAD compared to adult caregivers (mean age = 56.0 years, *SD* = 13.84; LOT optimism = 10.3, *SD* = 3.29; LOT Pessimism = 5.39, *SD* = 3.27) and noncaregivers (mean age = 61.5 years, *SD* = 13.22; LOT Optimism = 11.5, *SD* = 2.79; LOT pessimism = 4.96, *SD* = 3.28) (Robinson-Whelen et al., 1997) (*z* = –13.66 to 5.80, all *p* < .001). Subscale scores in the GAD group also indicated less optimism but equivalent pessimism relative to a sample of male navy recruits in basic training (mean age = 19 years; LOT scores converted from 1-5 to 0-4 scale: Optimism = 10.2, *SD* = 2.92; Pessimism = 8.4, *SD* = 3.24) (Marshall et al., 1992).

Similar comparisons indicated lower levels of general (*z* = –7.60, *p* < .001) and social (*z* = –10.05, *p* < .001) self-efficacy expectancies in older anxious adults relative to undergraduate students (SES General = 64.3, *SD* = 8.58; SES Social = 21.2, *SD* = 3.63) (Sherer & Adams, 1983). Comparison with other older adult samples who have completed the SES were not possible given the omission of mean scores (Bosscher & Smit, 1998) and the use of alternative scoring in previous work (Davis-Berman, 1990; Waller & Bates, 1992).

Internal Consistency

Internal consistency for the LOT and SES total and subscale scores was estimated with coefficient alpha (see Table 1). Coefficients indicated strong internal consistency for all summary scores (i.e., .75 or higher), with most item-remainder correlations ranging between .34 and .76. Only in the case of the LOT Pessimism subscale did the deletion of one item improve internal consistency. In this case, removal of Item 12 ("I rarely count on good things happening to me") resulted in an alpha of .83, as opposed to an alpha of .78.

Construct Validity

Scale and subscale intercorrelations. The LOT and SES total scores correlated positively ($r = .61$), suggesting a substantial association between optimism and generalized self-efficacy expectancies, although the coefficient indicates that the measures do not overlap completely. The LOT Optimism and Pessimism subscale scores correlated significantly, but only moderately, with each other ($r = –.42$), although both were strongly related to the LOT total score (r = .84 and –.84, respectively). The SES General and Social subscales correlated moderately with each other ($r = .49$), and both were associated strongly with the SES total (r = .96 and .70, respectively).

Factor analysis: LOT. To assess the adequacy of a previously tested model of the LOT (Marshall & Lang, 1990; Robinson-Whelen et al., 1997), a confirmatory factor analysis was conducted on a two-factor model including optimism (positively worded items) and pessimism (negatively worded items) factors. A single-factor model that comprised all the items also was tested as a plausible alternative. Fit indices for these models were derived by the SAS CALIS (Covariance Analysis and Linear Structural Equations) procedure (Hatcher, 1994).

The two-factor model was associated with a $\chi^2(19) = 21.46$, *ns*, and fit indices as follows: chi-square/*df* (fit ratio) = 1.13, root mean square error of approximation (RMSEA) = .04, goodness-of-fit index (GFI) = .93, adjusted GFI (AGFI) = .87, normed fit index (NFI) = .91, and normed noncentrality fit index (NNFI) = .98. Chi-square fit ratios that range between 1.0 and 2.0 are associated with well-fitting models (Novy et al., 1994), although the RMSEA is often viewed as a better descriptive fit index because it is affected less by sample size than chi-square (Quintana & Maxwell, 1999). A RMSEA value of .10 or higher has been suggested as indicating a poor fit (Browne & Cudeck, 1992); GFI and AGFI of .90 and .80, respectively, are generally considered acceptable (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Novy et al., 1994), as are NFIs of .80, with indices closer to 1.0 indicating better fits (Arbuckle, 1997). Factor loadings for the two-factor model ranged from .43 to .92 and are listed in Table 2. The correlation between factors was .47. All fit indices suggested that this two-factor model provided a good fit for the data.

The single-factor model was associated with a $\chi^2(20) = 73.36$, $p < .0001$, and fit indices as follows: chi-square/*df* (fit ratio) = 3.67, RMSEA = .19, GFI = .79, AGFI = .62, NFI = .67, and NNFI = .62. Factor loadings ranged from .32 to .91 and are listed in Table 2. Fit indices suggested that the single-factor model provided a poor fit for the data.

TABLE 2
Factor Loadings for LOT Items in a Sample of Older Adults With GAD

Item	*One-Factor Model*	*Two-Factor Model* *Factor 1*	*Two-Factor Model* *Factor 2*
1. In uncertain times, I usually expect the best.	.39	.55	—
4. I always look at the bright side of things.	.45	.84	—
5. I am always optimistic about my future.	.39	.77	—
11. I am a believer in the idea that "every cloud has a silver lining."	.32	.47	—
3. If something can go wrong for me, it will.	.59	—	.60
8. I hardly ever expect things to go my way.	.85	—	.86
9. Things never work out the way I want them to.	.91	—	.92
12. I rarely count on good things happening to me.	.43	—	.43

NOTE: LOT = Life Orientation Test; GAD = generalized anxiety disorder.

Factor analysis: SES. Because previous factor analytic studies on the SES have yielded inconsistent results, an exploratory factor analysis was conducted for this measure. Principal axis extraction and promax rotation were used, with the number of factors initially unspecified. Five factors had eigenvalues greater than one (8.15, 2.33, 1.86, 1.32, 1.14), accounting for a total of 54.6% variance. However, not all five factors were well defined, and the scree plot justified examination of a three-factor solution. Consequently, another principal axis analysis with promax rotation was conducted, with three factors specified. This model accounted for 47.0% of the variance. The three factors correlated moderately and positively with each other

TABLE 3
Structure and Pattern Coefficients for the SES in a Sample of Older Adults With GAD

		Structure Coefficients			*Pattern Coefficients*		
Item	*Original Specification*	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 1*	*Factor 2*	*Factor 3*
6. It is difficult for me to make new friends.	Social	.76	.45	.30	.73	.16	–.12
10. If I see someone I would like to meet, I go to that person instead of waiting for him or her to come to me.	Social	.70	.42	.34	.65	.11	–.01
14. If I meet someone interesting who is very hard to make friends with, I will soon stop trying to make friends with that person.	Social	.50	.10	.33	.54	–.31	.27
19. When I am trying to become friends with someone who seems uninterested at first, I do not give up very easily.	Social	.54	.28	.24	.54	–.03	–.02
24. I do not handle myself well in social gatherings.	Social	.66	.33	.39	.62	–.04	.14
28. I have acquired my friends through my personal abilities at making friends.	Social	.49	.22	.07	.55	.07	–.22
2. When I make plans, I am certain I can make them work.	General	.32	.50	.32	–.10	.43	–.04
3. One of my problems is that I cannot get down to work when I should.	General	.28	.64	.26	.01	.72	–.14
7. When I set important goals for myself, I rarely achieve them.	General	.43	.78	.62	–.07	.63	.27
8. I give up on things before completing them.	General	.47	.81	.55	–.07	.70	.13
15. When I have something unpleasant to do, I stick to it until I finish it.	General	.39	.60	.23	.17	.63	–.19
16. When I decide to do something, I go right to work on it.	General	.23	.73	.44	–.19	.76	–.10
12. If something looks too complicated, I will not even bother to try it.	General	.21	.29	.74	–.12	–.13	.86
18. When trying to learn something new, I soon give up if I am not initially successful.	General	.39	.48	.77	.03	.06	.73
20. When unexpected problems occur, I do not handle them well.	General	.51	.35	.57	.34	–.07	.46
22. I avoid trying to learn new things when they look too difficult for me.	General	.21	.28	.77	–.12	–.18	.93
23. Failure just makes me try harder.	General	.21	.43	.65	–.15	.15	.64
26. I feel insecure about my ability to do things.	General	.39	.56	.71	.01	.24	.57
27. I am a self-reliant person.	General	.32	.37	.63	.04	.01	.61
29. I give up easily.	General	.51	.58	.83	.13	.13	.70
30. I do not seem capable of dealing with most problems that come up in my life.	General	.43	.46	.76	.10	.01	.71
4. If I cannot do a job the first time, I keep trying until I can.	General	.30	.52	.52	–.02	.34	.34
11. I avoid facing difficulties.	General	.32	.24	.35	.21	–.01	.26

NOTE: SES = Self-Efficacy Scale; GAD = generalized anxiety disorder.

(*r* = .45-.56). Both structure and pattern coefficients were used to define the factors. In both cases, factor loadings with a value of .40 or higher were considered salient. Results of the three-factor model are presented in Table 3.

As expected, pattern coefficients, which consider factors independently of each other, were more helpful in defining the factors than structure coefficients, which take into account factor correlations, given the improved simple structure of the former. However, both sets of loadings suggested that Factor 1 overlapped with the original SES Social subscale, with all six items originally specified for that subscale (6, 10, 14, 19, 24, and 28) loading only or most highly on this factor. Structure coefficients indicated that five additional items had salient loadings on Factor 1 (7, 8, 20, 29, and 30), but these were complex items with higher loadings on other factors. Factor 2 appeared to assess planning/goal setting and demonstrated some overlap with the "strength" factor (i.e., efficacy in spite of obstacles) identified by Woodruff and Cashman (1993) and the "effort" factor identified by Bosscher and Smit (1998). Six of the eight items from the strength factor (2, 3, 7, 8, 15, and 16) and three of five items from the effort factor (2, 15, and 16) loaded only or most highly on Factor 2. Structure coefficients suggested that seven additional items had salient loadings on Factor 2 (6, 10, 18, 23, 26, 29, and 30), but these also were complex items with higher loadings on other factors. Factor 3 appeared to assess initiative/persistence, demonstrating overlap with two factors described similarly by Bosscher and Smit and the "magnitude" (i.e., efficacy over difficult levels of performance) and "competence" (i.e., general sense of competence) factors identified by Woodruff and Cashman. Six of the seven items from Bosscher and Smit's initiative and persistence factors (17, 18, 20, 22, 26, and 30) and eight of the nine items from

Woodruff and Cashman's factors (12, 18, 20, 22, 26, 27, 29, and 30) loaded only or most highly on Factor 3. Only three other items had salient loadings on this factor, but again these were complex items with higher loadings on other factors. Two items failed to load consistently on any single factor (4 and 11).

Calculation of coefficient alpha for each factor was based on items with salient pattern loadings. Coefficients were as follows: .86 (Factor 1), .89 (Factor 2), and .92 (Factor 3).

Subscale validity. To investigate construct validity for the LOT Optimism and Pessimism subscales, scores were correlated with measures of negative affect, with the expectation that pessimism would be associated more strongly with these measures than optimism as in prior research (Marshall et al., 1992; Robinson-Whelen et al., 1997). Correlations are included in Table 4. Statistical comparisons of dependent *r*s (Cohen & Cohen, 1983) indicated that pessimism only correlated more strongly than optimism with depression (BDI) ($t(73) = 1.79, p < .05$). Pessimism and optimism correlated equally with worry (PSWQ), trait anxiety (STAI-T), and neuroticism (NEO-PI-R).

To examine construct validity for the SES subscales, scores were correlated with measures of negative affect and social anxiety. The two general self-efficacy subscales (Initiative/Persistence, Planning/Goal Setting) were expected to correlate more highly with general measures of negative mood, whereas social self-efficacy was expected to correlate more strongly with severity of social anxiety. Correlations are included in Table 5. Statistical comparisons of dependent *r*s indicated that the two measures of general self-efficacy correlated similarly with all measures of negative affect. However, social self-efficacy was more strongly related to NEO-PI-R Extraversion scores than both Initiative/Persistence, $t(73) = 2.55, p < .05$, and Planning/Goal Setting, $t(73) = 3.36, p < .01$. Similarly, FQ Social scores correlated more highly with social self-efficacy than both general self-efficacy subscale scores, Initiative/Persistence: $t(73) = 3.03, p < .01$, and Planning/Goal Setting: $t(73) = 4.28, p < .01$, providing some evidence for construct validity of the SES Social subscale.

DISCUSSION

This study was designed to examine the psychometric properties of the SES and LOT in a sample of older adults with GAD. In general, the results support the reliability and validity of these measures and set the stage for future investigations into the roles of generalized self-efficacy expectancies and outcome expectancies (or optimism) in the prediction of affect and behavior in this group.

TABLE 4
Correlations of Optimism and Pessimism With Measures of Negative Affect

Negative Affect	*Optimism*	*Pessimism*
PSWQ	–.33**	.35**
STAI-T	–.42***	.58***
NEO-PI-R Neuroticism	–.47***	.48***
BDI	$–.17_a$	$.38_b$***

NOTE: PSWQ = Penn State Worry Questionnaire; STAI-T = State-Trait Anxiety Inventory–Trait Scale; NEO-PI-R = revised Personality Inventory of the NEO; BDI = Beck Depression Inventory. Correlations with different subscripts are significantly different at $p < .05$.
**$p < .01$. ***$p < .001$.

TABLE 5
Correlations of Social Self-Efficacy and Measures of Negative Affect and Social Anxiety

	General Self-Efficacy Initiative/ Persistence	*General Self-Efficacy Planning/ Goal Setting*	*Social Self-Efficacy*
Negative affect			
PSWQ	–.30**	–.26*	–.22
STAI-T	–.41***	–.41***	–.55***
NEO-PI-R Neuroticism	–.69***	–.50***	–.57***
BDI	–.27*	–.34**	–.35**
Social anxiety			
NEO-PI-R Extraversion	$.45_a$***	$.37_a$***	$.70_b$***
FQ Social	$–.36_a$**	$–.28_a$*	$–.57_b$***

NOTE: PSWQ = Penn State Worry Questionnaire; STAI-T = State-Trait Anxiety Inventory–Trait Scale; NEO-PI-R = revised Personality Inventory of the NEO; BDI = Beck Depression Inventory; FQ = Fear Questionnaire. Correlations with different subscripts are significantly different at $p < .05$.
*$p < .05$. **$p < .01$. ***$p < .001$.

Overall, descriptive data suggested lower levels of self-efficacy and outcome expectancies in older adults with GAD relative to published data from younger adult control samples and older adult samples of caregivers and noncaregivers. This pattern was expected given prior data suggesting that increased anxiety is associated with reduced efficacy expectancies among younger adults (Bandura, 1997; Williams, 1995). Although the younger adult literature typically relies on more problem-specific efficacy measures, one might expect individuals with generalized anxiety to report lower global expectations of their abilities to behave in certain ways and to obtain desired outcomes as a result of their behaviors. This perspective is consistent with cognitive models of generalized anxiety, which posit that particular sets of general beliefs (e.g., intolerance of uncertainty, erroneous beliefs about the function of worry) (Dugas et al., 1998) are used to interpret a wide range of situations as potentially dangerous or frightening, possibly based on prior experience with success and failure in a variety of contexts. However, the unique role that lowered generalized self-efficacy and outcome expectancies may play among older adults with GAD, relative to those with other psychiatric symptoms or disorders, is not yet known. Of course, other well-known cognitive models have proposed similar global conceptualization schemes for alternative anxiety disorders and depression (A. T. Beck, Emery, & Greenberg, 1985; A. T. Beck, Rush, Shaw, & Emery, 1979). Older adults with symptoms of these syndromes might also be expected to have lower generalized efficacy expectancies, although in many cases these other disorders are characterized more clearly by problem-focused behaviors that support the need for problem-based evaluations of efficacy. The current study provides descriptive data against which to compare the responses of older adults with other anxiety or affective disorders to examine whether efficacy expectancies associated with late-life GAD are unique in any way.

Both the SES and LOT demonstrated adequate internal consistency among older adults with GAD, suggesting that the total and subscale scores of both measures comprise items that are interrelated. Other data have suggested modification of some items from these measures to produce more precise assessments of the underlying constructs (e.g., Bosscher & Smit, 1998; Scheier, Carver, & Bridges, 1994). However, item-remainder correlations here only suggested that removing one item from the LOT Pessimism subscale might improve internal consistency, but internal consistency with this item included was satisfactory.

The study provided some evidence for validity of the SES and LOT subscales. First, the subscales within each of these measures were only moderately related to each other, suggesting that each subscale may measure a slightly different facet of generalized self-efficacy or outcome expectancies. Second, factor analyses supported the potential validity of the subscales. Confirmatory factor analysis provided strong evidence for the existence of two factors within the LOT with items mirroring the Optimism and Pessimism subscales that have been identified consistently in previous literature. Results of the SES exploratory factor analysis were less straightforward given the complex structure demonstrated and the relatively small sample size used for this analysis. Concerns about the overall utility of a generalized self-efficacy construct also may be reflected in the significant overlap among structure coefficients. Nevertheless, interpretation of the analysis indicated three factors that overlapped to some degree with previous findings. The data supported the notion that social and generalized self-efficacy expectancies can be assessed differentially with this instrument. Third, correlational data provided further evidence for validity of the SES Social subscale given that scores on this subscale were more highly related to measures of social anxiety than factors representative of generalized self-efficacy expectancies. Similar correlational data provided minimal evidence of validity for the LOT subscales given that Optimism and Pessimism generally correlated similarly with measures of negative affect. The only exception was depression, which showed a significant relationship only to Pessimism, not Optimism. This finding is consistent with prior reports of the differential predictability of these two subscales (Marshall et al., 1992; Robinson-Whelen et al., 1997), although the data here do not rule out potential confounding effects of other variables (e.g., negative affectivity) on the relations between optimism/pessimism and mood or coping.

SES and LOT total scores correlated moderately with each other, suggesting some overlap between generalized self-efficacy and outcome expectancies, as expected (Bandura, 1997; Maddux, 1995). Correlations between these two types of expectancies should be higher when generalized measures are used relative to more problem-specific assessments. And here, the differentiation of generalized expectancies from related constructs such as personal mastery and self-esteem is important. However, even these more generalized measures of self-efficacy and outcome expectancy do not overlap completely, suggesting the potential value of examining the unique contributions of both constructs in predicting anxiety severity and anxiety-related behaviors for older adults with GAD. Measures of generalized expectancies, in fact, may be particularly important for future research in this domain given the need for standardized tools to evaluate psychiatric syndromes not characterized by fears of a unique set of environmental or internal stimuli. Disorders characterized by fears of clearly identifiable stimuli and associated avoidance behaviors (e.g., specific phobias, agoraphobia, obsessive-compulsive disorder) would lend themselves more easily to assessment of problem-specific expectancies. For patients with GAD, problem-specific assessments would require significant variability from patient to patient.

Conclusions of the study are limited by the relatively small sample size and the omission of alternative measures designed to assess constructs closely related to generalized self-efficacy and outcome expectancies (e.g., mastery, locus of control). Future research will need to replicate the proposed factor structure of the SES and examine the relative roles of generalized self-efficacy, outcome expectancies, and related constructs in the prediction of affective and behavioral responses for older adults with GAD. Prior reports of data from the same sample of patients suggested the potential role of expectancies in self-reported life satisfaction (Bourland et al., 2000) and in the creation of a taxonomy of potential GAD subtypes (Hopko et al., 2001).

However, these data are cross-sectional and do not assess the ability of generalized self-efficacy and outcome expectancies to predict future affective states and behaviors. This issue is of particular relevance in the prediction of treatment response, wherein expectancies are posited to be central mechanisms of change (Bandura, 1997; Williams, 1995).

REFERENCES

Akkerman, R. L., Stanley, M. A., Averill, P. M., Novy, D. M., Snyder, A. G., & Diefenbach, G. J. (in press). Recruiting older adults with generalized anxiety. *Journal of Mental Health and Aging.*

Allard, C., & Mishara, B. L. (1995). Individual differences in stimulus intensity modulation and its relationship to two styles of depression in older adults. *Psychology and Aging*, *10*, 395-403.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arbuckle, J. L. (1997). *AMOS users' guide version 3.6.* Chicago: Smallwaters Corporation.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191-215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Barlow, D. H. (1988). *Anxiety and its disorders*. New York: Guilford.

Beck, A. T., Emery, G., & Greenberg, R. (1985). *Anxiety disorders and phobias*. New York: Basic Books.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.

Beck, A. T., & Steer, R. (1987). *Beck Depression Inventory: Manual*. San Antonio, TX: Psychiatric Corporation.

Beck, J. G., Stanley, M. A., & Zebb, B. J. (1995). Psychometric properties of the Penn State Worry Questionnaire: A descriptive study. *Behaviour Research and Therapy*, *34*, 225-234.

Beck, J. G., Stanley, M. A., & Zebb, B. J. (1996). Characteristics of generalized anxiety disorder in older adults: A descriptive study. *Behaviour Research and Therapy*, *34*, 225-234.

Beekman, A.T.F., Bremmer, M. A., Deeg, D.J.H., van Balkom, A.J.L.M., Smit, J. H., de Beurs, E., et al. (1998). Anxiety disorders in later life: A report from the longitudinal aging study Amsterdam. *International Journal of Geriatric Psychiatry*, *13*, 717-726.

Blazer, D., George, L. K., & Hughes, D. (1991). The epidemiology of anxiety disorders: An age comparison. In C. Salzman & B. D. Lebowitz (Eds.), *Anxiety in the elderly: Treatment and research* (pp. 17-30). New York: Springer.

Bosscher, R. J., & Smit, J. H. (1998). Confirmatory factor analysis of the general Self-Efficacy Scale. *Behaviour Research and Therapy*, *36*, 339-343.

Bourland, S. L., Stanley, M. A., Synder, A. G., Novy, D. M., Beck, J. G., Averill, P. M., et al. (2000). Quality of life in older adults with generalized anxiety. *Aging and Mental Health*, *4*, 315-323.

Brown, T. A., DiNardo, P. A., & Barlow, D. H. (1994). *Anxiety Disorders Interview Schedule for* DSM-IV. Albany, NY: Graywinds.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230-258.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Costa, P. T., Jr., & McCrae, R. R. (1992). *The Revised NEO Personality Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Davis-Berman, J. (1990). Physical self-efficacy, perceived physical status, and depressive symptomatology in older adults. *Journal of Psychology*, *124*, 207-215.

de Beurs, E., Beekman, A.T.F., van Balkom, A.J.L.M., Deeg, D.J.H., van Dyck, R., & van Tilburg, W. (1999). Consequences of anxiety in older persons: Its effect on disability, well-being and use of health services. *Psychological Medicine*, *29*, 583-593.

Diefenbach, G. J., Stanley, M. A., & Beck, J. G. (2001). Worry content reported by older adults with and without generalized anxiety disorder. *Aging and Mental Health*, *5*, 269-274.

Dugas, M. J., Gagnon, F., Ladouceur, R., & Freeston, M. H. (1998). Generalized anxiety disorder: A preliminary test of a conceptual model. *Behaviour Research and Therapy*, *36*, 215-226.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 277-299.

Fagan, P. J., Wise, T. N., Schmidt, C. W., Ponticas, Y., Marshall, R. D., & Costa, P. T., Jr. (1992). A comparison of five-factor personality dimensions in males with sexual dysfunction and males with paraphilia. *Journal of Personality Assessment*, *57*, 434-448.

Ferrari, J. R., Parker, J. T., & Ware, C. B. (1992). Academic procrastination: Personality correlates with Myers Briggs types, self-efficacy, and academic locus of control. *Journal of Social Behavior and Personality*, *7*, 495-502.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method of grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189-198.

Gallagher, D., Breckenridge, J. N., Steinmetz, J., & Thompson, L. W. (1983). The Beck Depression Inventory and research diagnostic criteria: Congruence in an older population. *Journal of Consulting and Clinical Psychology*, *51*, 945-946.

Gallagher, D., Nies, G., & Thompson, L. W. (1982). Reliability of the Beck Depression Inventory with older adults. *Journal of Consulting and Clinical Psychology*, *50*, 152-153.

Grembowski, D., Patrick, D., Diehr, P., Durham, M., Beresford, S., Kay, E., et al. (1993). Self-efficacy and health behavior among older adults. *Journal of Health and Social Behavior*, *34*, 89-104.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute, Inc.

Himmelfarb, S., & Murrell, S. A. (1983). Reliability and validity of five mental health scales in older persons. *Journal of Gerontology*, *38*, 333-339.

Hopko, D. R., Novy, D. M., Stanley, M. A., Beck, J. G., Averill, P. A., & Swann, A. C. (2001, November). *An empirical taxonomy of older adults diagnosed with generalized anxiety disorder.* Paper presented at the 35th annual Convention for the Advancement of Behavior Therapy, Philadelphia.

Kabacoff, R. I., Segal, D. L., Hersen, M., & Van Hasselt, V. B. (1997). Psychometric properties and diagnostic utility of the Beck Anxiety Inventory and the State-Trait Anxiety Inventory with older adult psychiatric outpatients. *Journal of Anxiety Disorders*, *11*, 33-47.

Maddux, J. E. (1995). *Self-efficacy, adaptation, and adjustment* (Plenum series in social/clinical psychology). New York: Plenum.

Maddux, J. E., & Gosselin, J. T. (in press). Self-efficacy. In M. Leary & J. Tangney (Eds.), *Handbook of self and identity.* New York: Guilford.

Marks, I. M., & Mathews, A. M. (1979). Brief standardized self-rating for phobic patients. *Behaviour Research and Therapy*, *17*, 263-267.

Marshall, G. N., & Lang, E. L. (1990). Optimism, self-mastery, and symptoms of depression in women professionals. *Journal of Personality and Social Psychology*, *59*, 132-139.

Marshall, G. N., Wortman, C. B., Kusulas, J. W., Hervig, L. K., & Vickers, R. R. (1992). Distinguishing optimism from pessimism: Relations to fundamental dimensions of mood and personality. *Journal of Personality and Social Psychology*, *62*, 1067-1074.

Martin, T. R., Flett, G. L., Hewitt, P. L., Krames, L., & Szanto, G. (1996). Personality correlates of depression and health symptoms: A test of a self-regulation model. *Journal of Research in Personality*, *31*, 264-277.

Mendes de Leon, C. F., Seeman, T. E., Baker, D. I., Richardson, E. D., & Tinetti, M. E. (1996). Self-efficacy, physical decline, and change in functioning in community-living elders: A prospective study. *Journal of Gerontology: Social Sciences*, *51B*, S183-S190.

Meyer, T., Miller, M., Metzger, R., & Borkovec, T. D. (1990). Development and validity of the Penn State Worry Scale. *Behaviour Research and Therapy*, *28*, 487-495.

Mroczek, D. K., Spiro, A., Aldwin, C. M., Ozer, D. J., & Bosse, R. (1993). Construct validation of optimism and pessimism in older men: Findings from the Normative Aging Study. *Health Psychology*, *12*, 406-409.

Novy, D. M., Frankiewicz, R. G., Francis, D. J., Liberman, D., Overall, J. E., & Vincent, K. R. (1994). An investigation of the structural validity of Loevinger's model and measure of ego development. *Journal of Personality*, *62*, 87-118.

Patterson, R. L., Sullivan, M. J., & Spielberger, C. D. (1980). Measurement of state and trait anxiety in elderly mental health clients. *Journal of Behavioral Assessment*, *2*, 89-97.

Quintana, S. M., & Maxwell, S. E. (1999). Implications of recent developments of structural equation modeling to counseling psychology. *The Counseling Psychologist*, *27*, 485-527.

Robinson-Whelen, S., Kim, C., MacCallum, R. C., & Kiecolt-Glaser, J. K. (1997). Distinguishing optimism from pessimism in older adults: Is it more important to be optimistic or not to be pessimistic? *Journal of Personality and Social Psychology*, *73*, 1345-1353.

Rodin, J. (1986). Aging and health: Effects of the sense of control. *Science*, *233*, 1271-1276.

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, *4*, 219-247.

Scheier, M. F., & Carver, C. S. (1993). On the power of positive thinking: The benefits of being optimistic. *Current Directions in Psychological Science*, *2*, 26-30.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, *67*, 1063-1078.

Seeman, T. E., Rodin, J., & Albert, M. (1993). Self-efficacy and cognitive performance in high-functioning older individuals. *Journal of Aging and Health*, *5*, 455-474.

Sharpe, P. A., Hickey, T., & Wolf, F. M. (1994). Adaptation of a general optimism scale for use with older women. *Psychological Reports*, *74*, 931-937.

Sherer, M., & Adams, C. H. (1983). Construct validation of the Self-Efficacy Scale. *Psychological Reports*, *53*, 899-902.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The Self-Efficacy Scale: Construction and validation. *Psychological Reports*, *51*, 663-671.

Smits, C.H.M., Deeg, D.J.H., & Bosscher, R. J. (1995). Well-being and control in older persons: The prediction of well-being from control measures. *International Journal of Aging and Human Development*, *40*, 237-251.

Snyder, A. G., Stanley, M. A., Novy, D. M., Averill, P. M., & Beck, J. G. (2000). Measures of depression in older adults with generalized anxiety disorder: A psychometric evaluation. *Depression and Anxiety*, *11*, 114-120.

Spielberger, C. D. (1983). *State-Trait Anxiety Inventory for adults: Manual*. Palo Alto, CA: Consulting Psychologists Press, Inc.

Stanley, M. A., Beck, J. G., & Glassco, J. D. (1996). Treatment of generalized anxiety in older adults: A preliminary comparison of cognitive-behavioral and supportive approaches. *Behavior Therapy*, *27*, 565-581.

Stanley, M. A., Beck, J. G., Novy, D. M., Averill, P. M., Swann, A. C., Snyder, A., et al. (1999, November). Cognitive behavioral treatment of GAD in older adults. In M. A. Stanley (Chair), *Assessment and treatment of anxiety in late-life*. Symposium conducted at the meeting of the Association for Advancement of Behavior Therapy, Toronto, Canada.

Stanley, M. A., Beck, J. G., & Zebb, B. J. (1996). Psychometric properties of four anxiety measures in older adults. *Behaviour Research and Therapy*, *34*, 827-838.

Stanley, M. A., Novy, D. M., Bourland, S. L., Beck, J. G., & Averill, P. M. (2001). Assessing older adults with generalized anxiety: A replication and extension. *Behaviour Research and Therapy*, *39*, 221-235.

Tinetti, M. E., Richman, D., & Powell, L. (1990). Falls efficacy as a measure of fear of falling. *Journal of Gerontology: Psychological Sciences*, *45*, P239-P243.

Tipton, R. M., & Worthington, E. L. (1984). The measurement of generalized self-efficacy: A study of construct validity. *Journal of Personality Assessment*, *48*, 545-548.

Waller, K. V., & Bates, R. C. (1992). Health locus of control and self-efficacy beliefs in a healthy elderly sample. *American Journal of Health Promotion*, *6*, 302-309.

Weissman, M. M., Myers, J. K., Tischler, G. L., Holzer, C. E., Leaf, P. J., Orvaschel, H., et al. (1985). Psychiatric disorders (*DSM-III*) and cognitive impairment among the elderly in a U.S. urban community. *Acta Psychiatrica Scandanavia*, *71*, 366-379.

Welch, D. C., & West, R. L. (1995). Self-efficacy and mastery: Its application to issues of environmental control, cognition, and aging. *Developmental Review*, *15*, 150-171.

Wetherell, J. L., Gatz, M., & Craske, M. G. (2001). *Treatment of generalized anxiety disorder in older adults*. Manuscript under review.

Williams, S. L. (1995). Self-efficacy and anxiety and phobic disorders. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment* (Plenum series in social/clinical psychology, pp. 69-107). New York: Plenum.

Woodruff, S. L., & Cashman, J. F. (1993). Task, domain, and general efficacy: A reexamination of the Self-Efficacy Scale. *Psychological Reports*, *72*, 423-432.

Melinda A. Stanley, Ph.D., is a professor of psychiatry and behavioral sciences at the University of Texas–Houston Health Science Center. She received her Ph.D. in psychology from Texas Tech University and completed a postdoctoral fellowship at the University of Pittsburgh Medical School. Her academic work focuses on the nature and treatment of anxiety in later life, with a most recent emphasis on identification and treatment of anxiety in primary care settings.

Diane M. Novy, Ph.D., is an associate professor in the Departments of Anesthesiology and Psychiatry and Behavioral Sciences. As an attending psychologist at the University Center for Pain Medicine and Rehabilitation, she integrates behavioral medicine and alternative medicine treatment strategies in the care of patients with chronic pain and medical conditions. Her research has focused on behavioral and alternative strategies for managing pain and affective distress. She also is interested in psychometric issues. For the past 9 years, she has served as a board member to the Texas Rehabilitation Commission.

Derek R. Hopko, Ph.D., is assistant professor at the University of Tennessee. He received his Ph.D. in psychology from West Virginia University and completed his residency and postdoctoral training at the University of Texas Medical School. His research focuses on the causes and correlates of anxiety disorders and treatment outcome as it pertains to behavioral therapy for major depression.

J. Gayle Beck, Ph.D., is a professor in the Department of Psychology at the State University of New York at Buffalo. Dr. Beck specializes in research on the psychopathology and treatment of anxiety disorders in adults, including older adults. She has published widely on Panic Disorder, Generalized Anxiety Disorder, and most recently, Posttraumatic Stress Disorder. A graduate of State University of New York at Albany (where she trained with David Barlow, Ph.D.), Dr. Beck recently completed a tour as editor of *Behavior Therapy* and currently serves on six editorial boards.

Patricia M. Averill, Ph.D., is an associate professor in the Department of Psychiatry at the Univeristy of Texas–Houston Medical School. She obtained her Ph.D. in clinical psychology from the University of Houston and completed her internship at UT–Houston Medical School and a postdoctoral fellowship in Pain Management at UT–Houston Department of Anesthesia. Her research interests include the adult anxiety disorders, dual diagnosis, and program evaluation.

Alan Swann, M.D., is Pat R. Rutherford Jr. Professor and Vice Chair for Research in the Department of Psychiatry, University of Texas Medical School at Houston. He has carried out research, patient care, and teaching in affective disorders for more than 20 years. He is part of a group that integrates treatment with basic and clinical research in mood disorders. His clinical research focuses on treatment of affective disorders, especially prediction of treatment response and development of more objective measures of disease severity and its change during treatment; preclinical human research mainly concerns the neurobiology of behavior, such as impulsivity and motivation; and basic research focuses on pharmacological and developmental aspects of mechanisms like behavioral sensitization to stimulants. He has presented extensively on these topics and has more than 200 publications.

Factor and Subtest Discrepancies on the Differential Ability Scales

Examining Prevalence and Validity in Predicting Academic Achievement

Shoshana Y. Kahana
Eric A. Youngstrom
Case Western Reserve University

Joseph J. Glutting
University of Delaware

*Past literature has largely ignored the population frequency of multivariate factor and subtest score discrepancies. Another limitation has been that statistical models imperfectly model the clinical assessment process, whereby significant discrepancies between both factors and subtests are included in predictions about an individual's academic achievement. The present study examined these issues using a nationally representative sample (*N *= 1,185) completing the Differential Ability Scales. Results indicate that approximately 80% of children in a nonreferred sample show at least one statistically significant ability discrepancy. In addition, the global estimate of cognitive ability was the most parsimonious predictor of academic achievement, whereas information about ability discrepancies did not significantly improve prediction. Findings suggest that when predicting academic achievement, there is little value in interpreting cognitive scores beyond the global ability estimate.*

Keywords: cognitive testing, predictions of academic achievement, global measure of intelligence, factor and subtest interpretation

The interpretation of cognitive abilities has significant implications for the prognostic evaluation of an individual. General cognitive ability predicts criteria such as scholastic achievement (Jensen, 1998), years of education (Jensen, 1998), and work-related success (Ceci & Williams, 1997; Jensen, 1998; Kaufman, 1994; Neisser et al., 1996). Many practicing clinicians, however, place less emphasis on a general cognitive ability construct and opt instead to interpret the more discrete measures of cognitive ability tests, such as factor and subtest scores, to glean information that is thought to be diagnostically useful and pertinent to treatment.

The current standard of clinical practice employs the "top-down approach" to test score interpretation, a methodology that advocates the use of factors (Donders, 1996; Kamphaus, 1993; Kaufman, 1994; McGrew & Flanagan, 1998; Naglieri, 1993; Sattler, 1992, 2001) and subtests (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Gregory, 1999; Kamphaus, 1993; Kaufman, 1979, 1994; Kaufman & Lichtenberger, 1999; Prifitera, Weiss, & Saklofske, 1998; Sattler, 1992, 2001) to construct clinical formulations about the strengths and deficits associated with an individual's performance. For example, the Processing Speed Index on the Wechsler Intelligence Scale for Children–Third Edition (WISC-III) (Wechsler, 1991) may relate to

Please address correspondence and reprint requests to Shoshana Y. Kahana, Case Western Reserve University, Department of Psychology, 11220 Bellflower Road, Cleveland, OH 44106-7123; e-mail: syk4@po.cwru.edu.

Assessment, Volume 9, No. 1, March 2002 82-93

attentional problems associated with information-processing deficits such as those manifest in attention-deficit hyperactivity disorder (Schwean, Saklofske, Yackulic, & Quinn, 1993). In addition, Lipsitz, Dworkin, and Erlenmeyer-Kimling (1993) found that the Comprehension subtest on the WISC-Revised (WISC-R) (Wechsler, 1974) significantly correlated with social competence for normal participants in childhood. A recent special issue of *School Psychology Quarterly* (2000, Vol. 15) demonstrates that interpretive approaches that incorporate both factor and subtest interpretation, such as the successive levels method and profile analysis, are still very much practiced and actively encouraged for clinicians (Carroll, 2000; Naglieri, 2000; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000; Pritchard, Livingston, Reynolds, & Moses, 2000; Riccio & Hynd, 2000; Stanton & Reynolds, 2000; see Watkins, 2000, for commentary). This top-down approach is based on the premise that multiple abilities have more predictive power than a single general cognitive factor.

Despite the potential clinical promise of differentiating among cognitive ability dimensions, the utility of these more discrete indices is dubious. In the past decade, several studies have demonstrated the questionable reliability and validity of interpreting the more specific measures of cognitive ability for predicting child functioning (Beebe, Pfiffner, & McBurnett, 2000; Glutting, Youngstrom, Ward, Ward, & Hale, 1997; McDermott & Glutting, 1997; Riccio, Cohen, Hall, & Ross, 1997; Youngstrom & Glutting, 2001; Youngstrom, Kogos, & Glutting, 1999). More specifically, proponents of the top-down approach and subtest analysis (e.g., Kamphaus, 1993; Kaufman, 1979, 1994; Sattler, 1992, 2001) have yet to prove the criterion-related and incremental validity of discrete indices in predicting academic criteria beyond that offered by a global measure of cognitive ability. In practical terms, the primary standard for validity of diagnostic, score-based interpretations should be the degree to which they accurately predict future performance or prescribe a clear intervention (Glutting et al., 1997; Glutting, Watkins, & Youngstrom, in press; Gough, 1971).

Current tests of ability and achievement have attained a level of precision whereby it is possible to identify reliable discrepancies between cognitive ability indices that occur at high rates in the population and, as a result, are likely to be diagnostically uninformative. Because contemporary interpretive practice encourages multiple comparisons across sets of factor and subtest scores, it is important to know the overall base rate of discrepancies (or the frequency of discrepancies between multiple variables) in the population. This information has not been previously reported for the Differential Ability Scales (DAS) (Elliot, 1990a). Thus, the first purpose of this article is to examine the overall frequency of significant ability discrepancies between factor and subtest scores from the DAS.

In addition, the current literature on cognitive interpretations has utilized research designs that fail to accurately model the clinical assessment process. Previous investigations of the incremental validity of the more discrete cognitive ability scores (e.g., Youngstrom et al., 1999) were limited in that they usually selected or emphasized one discrete index (e.g., factor), as opposed to simultaneously considering several lower level indices, in predicting academic achievement. In reality, however, current clinical interpretive practice typically encourages psychologists to compare many ability variables present on an IQ-test protocol (e.g., possible factor-score comparisons, possible subtest-score comparisons) (Kaufman, 1994; Sattler, 1992, 2001). For example, assessment authorities such as Kaufman (1994) and Sattler (1992, 2001) provided interpretive tables that include common or shared abilities tapped by various subtests as well as what specific subtest discrepancies might indicate. More specifically pertaining to the DAS, Sattler (2001) clearly outlined the approaches to profile analysis for the DAS. In addition to evaluating the global estimate of intelligence and factor scores, the clinician is encouraged to evaluate within-factor differences, evaluate differences between subtests and the mean core *T*-score, and compare each subtest *T*-score with the other subtest *T*-scores.

It appears that the clinician is advised to interpret if not all then certainly most potential discrepancies between subtests. Whereas past studies have found that factor IQs retain no general advantage over conventional IQs in predicting academic performance (Glutting et al., 1997; Youngstrom et al., 1999), no research to date has examined whether there is any predictive utility to factor and subtest discrepancies for the prediction of academic achievement on the DAS. Thus, the second purpose of this study is to model and examine the validity of current clinical assessment, whereby (a) both factors and subtests are used in predicting academic achievement and, more important, (b) through the use of interaction terms, to examine whether significant discrepancies between discrete indices are valid in predicting academic achievement (i.e., do they provide any incremental contributions beyond the global estimate of cognitive ability in predicting academic achievement). Many advocates of factor and subtest analysis acknowledge that global estimates of cognitive ability will make accurate predictions, particularly for children with evenly developed cognitive abilities (e.g., Kamphaus, 1993; Kaufman, 1994; Sattler, 1992). The question is whether children who display discrepancies across specific cognitive abilities will perform differently, on average, from those who do not display either strengths or weaknesses in academic achievement.

The DAS serves as a good instrument for this study. Research has consistently shown that both the general cognitive ability scores and the more discrete verbal ability and nonverbal reasoning conceptual abilities are measured well by the DAS (Keith, 1990). The DAS was a priori designed to measure multiple factors of cognitive ability in the belief that such measurement would provide clinically relevant information pertaining to diagnosis and treatment. In addition, the DAS contains multiple achievement tests that can serve as particularly effective criterion measures for the concurrent prediction of academic achievement (Elliot, 1990b, 1990c).

The present investigation did not consider all possible patterns of discrepancy on the DAS. We chose to examine those discrepancies that current clinical interpretive practice considers either theoretically motivated or empirically supported for predicting clinical or academic problems. Sattler (2001) provided a list of illustrative hypotheses for verbal ability and nonverbal reasoning discrepancies, verbal ability and spatial ability discrepancies, and for comparisons of the Verbal, Nonverbal Reasoning, and Spatial Ability subtests. To choose factors that were related to specific achievement criteria, we also examined the DAS manual's table of correlations (Table 9.37) between ability and achievement for all children in the standardization sample. We chose those factors that had the highest correlations with the achievement tests, such that nonverbal reasoning correlated the highest with basic number skills (r = .59), verbal ability and nonverbal reasoning with spelling (both correlated r = .49), and verbal ability with word reading (r = .59). Thus, the final choice of discrepancies to include in the models predicting achievement blended clinical recommendations with existing data about ability-achievement correlations.

METHOD

Participants

Participants were 1,185 children who completed the DAS during the national standardization of the Adjustment Scales for Children and Adolescents (McDermott, 1994). The sample was configured according to the 1988 to 1990 U.S. census and was stratified for age, gender, race/ethnicity, parent education, family structure, national region, and community size. Although the DAS standardization sample included preschoolers, the present study concentrated entirely on school-aged children, who ranged in age from 6 to 17 years, with a mean of 11.54 (*SD* = 3.41) years. Children from kindergarten through 12th grade were included, with 3 participants marked as "other." There was a roughly equal representation of gender across grade, with 595 males and 590 females participating. The ethnic breakdown of the sample was such that 69.7% of the participants were Caucasian, 14.9% African American, 11.8% Hispanic, and 3.6% from other ethnic groups. In accordance with the 1988 to 1990 U.S. census, three fourths of the sample (74.8%) came from families in which there was both a mother and father. Finally, 44.8% of the participants were from a major metropolitan area, 35% from a minor metropolitan area, and 19.4% from a rural area.

Procedure

Two hundred and twenty-five individuals with formal training in individual assessment of cognitive ability administered the DAS and achievement subtests. Administrators were either independently qualified to administer such tests or were appropriately supervised. The majority of the administrators were trained in DAS administration by project staff members at workshops conducted during the fall of 1986 and were required to submit two satisfactory practice cases before being permitted to conduct testing. Central project staff managers reviewed all case protocols as they were received to maintain adequate standardization and administration procedures (Elliot, 1990a).

Measures

Cognitive ability. The interpretive hierarchical structure of the DAS begins with a global measure of intelligence called the General Conceptual Ability (GCA) (alpha = .95), a measure of general cognitive ability, analogous to *g*, the construct underlying intelligence. In general, the DAS uses a relatively small number of core subtests that have high *g* loadings for the calculations of the GCA score. The subtests cover a range of abilities and processes, including verbal ability and nonverbal reasoning, visual and auditory memory, language expression and comprehension, perceptual-motor skills, speed of information processing, and school achievement in essential areas (Elliot, 1990a). The GCA shows good convergent validity with other measures of general ability, including WISC-R full scale IQ (FSIQ) (r = .84) and the Stanford-Binet Intelligence Scale–Fourth Edition (Thorndike, Hagen, & Sattler, 1986) composite score (r = .88) (Elliot, 1990a).

The GCA is composed of three factor scores: Verbal Ability, Nonverbal Reasoning, and Spatial Ability.[1] Each factor score is defined by two subtests. The Word Definitions (alpha = .83) and Similarities (alpha = .79) subtests compose the Verbal Ability factor. The Nonverbal Reasoning factor is formed by the Matrices (alpha = .82) and

TABLE 1
Descriptive Statistics for Cognitive Ability and Achievement Measures (*N* = 1,185)

Variable	*Mean*	*Standard Deviation*	*Range*
Cognitive ability[a]			
GCA	100.56	14.67	55-145
Verbal factor	100.15	14.54	55-140
Nonverbal Reasoning factor	100.44	14.99	60-142
Spatial factor	100.68	14.48	56-143
Cognitive subtests[b]			
Matrices	50.50	10.00	20-80
Similarities	50.19	9.77	20-80
Sequential and Quantitative Reasoning	50.45	9.90	20-79
Pattern Construction	50.54	9.77	21-80
Recall of Digits	50.18	9.84	20-80
Recall of Objects	50.14	10.33	20-80
Recall of Designs	50.87	9.65	20-79
Achievement criteria[a]			
Basic Number Skills	100.44	14.53	57-145
Word Reading	101.17	15.01	55-145
Spelling	100.50	14.88	55-145

NOTE: GCA = General Conceptual Ability.
a. Standard score metric, *M* = 100, *SD* = 15.
b. *T*-score metric, *M* = 50, *SD* = 10.

Sequential and Quantitative Reasoning (alpha = .85) subtests. The Recall of Designs (alpha = .84) and Pattern Construction (alpha = .91) subtests constitute the Spatial Ability factor.

The Verbal Ability factor is a measure of complex verbal mental processing that includes acquired verbal concepts, verbal knowledge, reasoning, and a general knowledge base. It is strongly associated with other related measures, reporting a correlation of .84 with the WISC-R verbal IQ and a .72 with the WISC-R FSIQ (Elliot, 1990a). The Nonverbal Reasoning factor represents nonverbal and inductive reasoning and requires complex mental processing. It too shows good convergent validity with other measures, correlating .75 with the WISC-III Perceptual Organization Index (POI) and .78 with the WISC-III performance IQ (PIQ) (Elliot, 1990a). The Spatial Ability factor is related to visualization, spatial orientation, and visual-motor tasks. It shows strong associations with other performance-oriented tasks, correlating .82 with both the WISC-III PIQ and POI. The GCA and factor indices are expressed as standard scores, with population means of 100 and standard deviations of 15.

Achievement criteria. The DAS also consists of three individual achievement scales that were conormed with the ability subtests. They include Basic Number Skills (alpha = .87), Spelling (alpha = .92), and Word Reading (alpha = .92). Basic Number Skills focuses on the concepts and skills that underlie basic competence in arithmetic calculation. Spelling examines the child's ability to produce correct spellings and includes a range of phonetically regular and irregular words. Word Reading is an achievement test of the recognition and oral reading of single words.

In general, each of these tests shows a consistently moderate to good pattern of convergent and divergent validity with other individually and group administered achievement tests (*r* = .43-.68) (Elliot, 1990a). In addition, the three DAS achievement scales demonstrate moderately positive correlations with school performance, specifically teacher-assigned school grades in Mathematics, Spelling, and Reading (Elliot, 1990a). The tendency of school grades to be less reliable than scores on standardized group achievement tests reduces the level of correlation between the two. Nevertheless, these correlations are higher than those obtained by other frequently used achievement tests, such as the Wechsler Individual Achievement Test (Psychological Corporation, 1992) (*r* = .23-.46).

RESULTS

Descriptive Statistics

Table 1 presents the descriptive statistics for all of the measures used in subsequent analyses. It is worth noting that all ability and achievement variables showed approximately normal distributions (all skews between –.05 and +.14, all kurtoses between –.40 and +.08), and there were no univariate outliers with unusually extreme scores.

Frequency of Discrepant Cognitive Abilities

To estimate multivariate prevalence, we calculated critical discrepancy scores for each of the following pairings: at the factor score level, Verbal Ability versus Nonverbal Reasoning, Verbal Ability versus Spatial Ability, and Nonverbal Reasoning versus Spatial Ability and at the subtest level, Word minus Similarities, Design minus Pattern Construction, Matrices minus Sequential and Quantitative Reasoning.[2] It is important to note that other pairings are possible (e.g., Similarities versus Matrices), but these were not included because current practice does not emphasize direct comparison of subtests that fall on different ability factors (cf. Sattler, 1992). We then determined the percentage of participants showing discrepancies exceeding each critical value. As reported in Table 2, 373 participants (31.5%) showed a reliable difference between Verbal Ability and Nonverbal Reasoning scores, whereas Verbal and Spatial Ability discrepancies were somewhat more common, with 474 youths (40.0%) exhibiting discrepancies. Finally, we calculated the cumulative number of dis-

TABLE 2
Multivariate Frequency of Cognitive Ability Discrepancies (N = 1,185)

	90% Confidence (Clinical Standard of Practice)		*95% Confidence*	
Number of Discrepancies	*Percentage of Sample*	*Cumulative Percentage*	*Percentage of Sample*	*Cumulative Percentage*
6	0.2	0.2	0.0	0.0
5	2.4	2.6	0.7	0.7
4	11.1	13.7	5.2	5.9
3	26.8	40.5	19.2	25.1
2	28.2	68.7	30.0	55.1
1	19.7	88.4	25.1	80.3
0	11.6	100.0	19.7	100.0

crepancies exhibited by each participant. Table 2 indicates that a staggering 80% of the nonreferred sample showed at least one discrepancy between factor or subtest scores when using a conservative 95% confidence interval approach. When using the 90% confidence interval recommended for clinical applications (e.g., Kaufman, 1994; Sattler, 1992), 88.4% of youths exhibited at least one discrepancy.

Incremental Value of Specific Cognitive Abilities in Predicting Achievement

Multiple regression analyses tested the hypotheses that DAS factor or subtest scores would significantly improve prediction of academic achievement criteria even after controlling for GCA. These analyses are consistent with contemporary assessment practice that often incorporates information about the presence of a clinically interpretable discrepancy in underlying cognitive abilities. Particularly in situations where cognitive abilities are not evenly developed, discrepancies between the more discrete indices of measurement, such as factors and subtests, will often become the focal point or main piece of interest in determining a clinician's interpretation of the child. Authorities argue that when a child shows a substantial difference between Verbal Ability and Nonverbal Reasoning, for example, then GCA would not be the optimal predictor of reading achievement (e.g., Kaufman, 1994; Sattler, 1992, 2001). Instead, the child's Verbal Ability should be the more accurate predictor because it more purely reflects the cognitive ability most involved in reading.

Given the high frequency of statistically significant differences in the standardization sample (see Table 2), we defined clinically interpretable discrepancies as those having a bivariate prevalence rate of less than 10%. A number of writers have emphasized the differences between statistical significance and clinical rarity (Cahan, 1986; Glutting, McGrath, Kamphaus, & McDermott, 1992). Differences between scores may be statistically significant but not especially unusual or particularly meaningful in the population. The statistical significance of a discrepancy refers to the probability that the results are not merely a chance occurrence but does not describe how frequently a discrepancy of a given magnitude occurs in the normal population or whether the differences are so large that they are considered abnormal or rare.

Following this first set of results and employing statistical significance as a guideline, clinicians would identify some form of ability discrepancy or generate an interpretive hypothesis for close to 80% of the children in the United States. Thus, we decided that it might be more meaningful to look at those discrepancies that were both statistically significant and rare. We created two dummy codes for each participant, which were scored 0 if he or she did not show a rare or unusual discrepancy between factor scores (Verbal Ability–Nonverbal Reasoning or Verbal Ability–Spatial Ability) and scored 1 if he or she did. We also created dummy codes reflecting the presence or absence of rare subtest scatter between the two subtests composing the Verbal Ability factor (i.e., Word Definitions and Similarities, n = 105 with significant discrepancies or 8.9% of sample), the Nonverbal Reasoning factor (i.e., Matrices and Sequential and Quantitative Reasoning, n = 82 or 6.9% of sample), and the Spatial Ability factor (i.e., Pattern Construction and Recall of Designs, n = 88 or 7.4% of sample).

Next, we subtracted 100 from the GCA and the factor scores and 50 from each subtest score for each participant (centering them around their population means, as recommended for regression models using interaction terms). Interaction terms were created by multiplying each dummy code by its respective ability score. Regression models then tested the incremental validity of the factor scores, subtests, and discrepancy information using a block entry approach. For all three achievement criteria, the first block entered the GCA by itself. This model represents the most parsimonious approach, basing predicted academic achievement simply on general cognitive ability. The second block entered three factor-score predictors as a set, chosen on theoretical grounds for each achievement criterion. For Basic Number Skills, the second block of predictors included the centered Nonverbal Reasoning factor score, the dummy code indicating whether Nonverbal Reasoning was significantly different from Verbal Ability, and the interaction of the two. For Spelling and Word Reading achievement, the predictors were Verbal Ability, the dummy code for significant Verbal–Nonverbal Reasoning discrepancy, and the interaction term. The regression analyses were also repeated for each achievement criterion using Verbal and Spatial Ability discrepancies, resulting in a total of six regression models. Nonverbal Reasoning–Spatial Ability discrepancies were not examined because we are not aware of clinical guidelines for their interpretation, especially in the context of predicting academic achievement.

The interaction term was the most interesting component of the second block of predictors, for three reasons. Most important, the interaction operationalized the clinical view that the particular ability factor was most likely to be important for individuals with a marked strength or weakness on that ability. Conversely, previous research (e.g., Youngstrom et al., 1999) has already established that the main effect for each factor score does not provide substantial improvements in the prediction of academic criteria. Finally, the dummy codes were not expected to show significant effects because they blended children who showed weaknesses with those who showed strengths on a particular factor (whereas the interaction term included information about each child's specific performance on the factor, along with whether this performance constituted a clinically interpretable strength or weakness). The main effect and dummy code were still included in the regression model even though they were conceptually uninteresting because that is the recommended practice for using multiple regression to test interactions (Aiken & West, 1991; Cohen & Cohen, 1983).

The third block for the regression models entered the centered subtest scores, dummy codes for significant subtest discrepancies, and Subtest × Discrepancy interaction terms.[3] The third block simultaneously entered all of the subtests and discrepancies for the factor scores used in the second block.[4] For example, the regression model predicting Word Reading entered the centered scores for Word Definitions and Similarities, the dummy code for the discrepancy, and the interaction between both subtests and the discrepancy. Again, the interaction terms were the only components of conceptual interest in each block; the main effects were included only to follow established guidelines for regression analysis. The inclusion of the main effects did not substantially reduce power to detect interactions, as the main effects in even the most complicated regression models only consumed 5 degrees of freedom, leaving 1,175 degrees of freedom for model testing.

Table 3 presents the results for the regressions that were grouped by achievement criterion. Regression coefficients are only reported for variables making a significant unique contribution or for the interaction terms, which are included because of their conceptual importance within the study. As the findings indicate, the GCA was the only variable to contribute significant, unique variance to the prediction of achievement in all of the various regression models. Nonverbal Reasoning and Spatial Ability both made small but statistically significant incremental predictions for Basic Number Skills, as can be seen by looking at the tests of their respective main effects. Similarly, Verbal Ability provided a small but significant incremental improvement in predicted Word Reading and Spelling. Reported statistical probabilities are two-tailed and should be compared with a Bonferroni-adjusted critical value of $p < .0083$ to maintain overall alpha < .05 across six models.

Table 3 also reports the part correlations between each predictor and the achievement criterion. Squaring the part *r* indicates how much variance in achievement was uniquely explained by the factor score. The largest part *r* was .18, for the Verbal Ability factor predicting Word Reading, showing that less than 4% of the variance in any achievement criterion was uniquely accounted for by an ability score. None of the subtests provided any significant predictive increment when entered in Block 3, and none of the mean centered factor scores continued to provide unique information once the constituent subtests were entered into the regression model. This is not surprising because the factor scores would be highly collinear with the subtests. Also as expected, the dummy coded discrepancy variables did not provide any new information in predicting the achievement criteria (all *p*s > .05).

None of the Cognitive Ability × Discrepancy interactions were statistically significant when compared with a Bonferroni-adjusted critical value of $p < .0024$ to maintain overall alpha < .05 for 21 comparisons. Even though none met statistical criteria for interpretation, all 21 interaction terms are reported in Table 3. They represent the most direct test of the clinical hypothesis that prediction of achievement criteria should be adjusted when there are clinically significant discrepancies between the factor scores. Examination of Table 3 shows that the largest part *r* associated with an interaction term is –.07, meaning that adjustments made on the basis of clinically significant ability discrepancies improved prediction of achievement criteria by, at best, 0.49%. These null findings are disappointing; however, they are unlikely to result from low statistical power. The tests of the interaction terms are based on 1, 1175 *df*, and using a two-tailed alpha of .05, statistical power was .80 to detect partial correlations of .082 or larger (Buchner, Faul, & Erdfelder, 1996).

Some researchers might take issue with the blocked hierarchical regression approach that we employed. Specifically, the high *g* loadings of the subtests and factors make it almost impossible to find additional unique contributions if the added variables bring in shared *g* as well as shared error variance. Obviously, there is a high degree of multicollinearity among the predictors as a consequence of global ability being derived from the underlying factor and subtest scores. However, in situations where variables are all highly interrelated, we would contend that more things (such as factor and subtest scores) will nearly al-

TABLE 3
Tests of Incremental Validity of Factor and Subtest Scores in Predicting Achievement Criteria for Individuals With Discrepant Cognitive Abilities

Criterion and Analysis	*Block*	*Predictor*	R^2 Δ	B	*Part* r
Basic Number Skills	1. GCA only	GCA	.347***	.58***	.59
Verbal versus Nonverbal Reasoning	2a. Factor discrepancies	3 *df*	.017***		
		GCA		.36***	.17
		Nonverbal Reasoning		.25***	.11
		Nonverbal Reasoning × Discrepancy		–.04	–.01
	3a. Subtest discrepancies	5 *df*	.010**		
		GCA		.34***	.15
		Nonverbal Reasoning × Discrepancy		–.06	–.02
		Matrices × Discrepancy		.00	.00
		Sequential/Quantitative × Discrepancy		.11	.02
Verbal versus Spatial[a]	2b. Factor discrepancies	3 *df*	.015***		
		GCA		.77***	.42
		Spatial		–.24***	–.12
		Spatial × Discrepancy		.12	.04
Word Reading	1. GCA only	GCA	.356***	.61***	.61
Verbal versus Nonverbal Reasoning	2a. Factor discrepancies	3 *df*	.037***		
		GCA		.31***	.16
		Verbal		.37***	.18
		Verbal × Discrepancy		–.10	–.03
	3a. Subtest discrepancies	5 *df*	.011***		
		GCA		.31***	.17
		Verbal × Discrepancy		–.09	–.03
		Word Definitions × Discrepancy		–.11	–.02
		Similarities × Discrepancy		.07	.02
Verbal versus Spatial	2b. Factor discrepancies	3 *df*	.037***		
		GCA		.35***	.19
		Verbal		.31***	.15
		Verbal × Discrepancy		.09	.03
	3b. Subtest discrepancies	5 *df*	.012**		
		GCA		.35***	.19
		Verbal × Discrepancy		.09	.03
		Word Definitions × Discrepancy		–.11	–.02
		Similarities × Discrepancy		.07	.02
Spelling	1. GCA only	GCA	.273***	.54***	.52
Verbal versus Nonverbal Reasoning	2a. Factor discrepancies	3 *df*	.018***		
		GCA		.32***	.17
		Verbal		.27***	.14
		Verbal × Discrepancy		–.24	–.07
	3a. Subtest discrepancies	5 *df*	.012**		
		GCA		.32***	.17
		Verbal × Discrepancy		–.23	–.06
		Word Definitions × Discrepancy		–.18	–.04
		Similarities × Discrepancy		.08	.02
Verbal versus Spatial	2b. Factor discrepancies	3 *df*	.015***		
		GCA		.38***	.20
		Verbal		.19***	.09
		Verbal × Discrepancy		.07	.03
	3b. Subtest discrepancies	5 *df*	.013**		
		GCA		.38***	.20
		Verbal × Discrepancy		.06	.02
		Word Definitions × Discrepancy		–.17	–.04
		Similarities × Discrepancy		.09	.02

NOTE: GCA = General Conceptual Ability. All analyses are based on $N = 1{,}185$. Final univariate tests of significance are based on 1, 1175 *df*. All interaction terms are reported regardless of statistical significance. Only main effects making a significant unique contribution to prediction are reported.

a. There were no subtest discrepancies that were considered appropriate to use in predicting Basic Number Skills when there was a Verbal Ability–Spatial Ability split. The subtests that were most logically connected to Basic Number Skills were subsumed in the Nonverbal Reasoning factor.

***$p < .001$, compare with a Bonferroni-adjusted critical value of $p < .0024$ to maintain overall alpha < .05 for 21 comparisons.

**$p < .05$.

ways predict as well, or even marginally better, than one thing (global ability). This phenomenon is precisely the rationale for why such multicollinearity is a violation of parsimony and not an asset (Glutting et al., in press). Nevertheless, we decided to examine if the factor indices would outpredict the GCA when each was entered alone into a regression equation. As Table 4 indicates, even when entered alone, in almost all cases the GCA will still outperform more discrete indices. Single-factor scores did not outpredict the GCA for any achievement criterion, with the exception of the Verbal Ability factor accounting for .002 more of the variance than GCA for Word Reading.

Alternate Tests of Clinical Interpretive Model

The regression models presented above represent a statistically conventional approach to evaluating interpretive strategy but might not exactly duplicate the logic recommended in clinical procedures. The regression approach uses the GCA to predict every child's achievement, then adjusts predictions based on one of the factor scores, and finally further adjusts scores when a significant discrepancy is present between the ability factor scores (by means of the interaction term). Similarly, information from the subtests is integrated after already controlling for the GCA, the factor score, the presence or absence of a cognitive discrepancy, and the interaction between factor and discrepancy. As Table 3 makes clear, these regression models become complicated, and in the final stages they are using 10 different pieces of information about the child to optimize prediction of achievement criteria.

Some authorities recommend the substitution of the factor score for the global cognitive ability estimate in cases where the cognitive ability scores actually are significantly disparate (e.g., Kaufman, 1994). In cases where there is reliable and statistically significant scatter among the factor scores, the regression approach would continue to utilize the GCA and augment it with additional information, whereas clinical authorities would opt to avoid interpreting the GCA and instead to replace it with one of the factor scores. The logic is that (a) if there are marked differences in the cognitive abilities measured by the factor scores, then the global measure of ability is a potentially misleading aggregate and (b) achievement criteria are likely to be predicted more accurately by a specific cognitive ability that is more directly related to the particular achievement task. For example, if a child has a GCA of 94 but a Verbal Ability factor score of 77 and a Nonverbal Reasoning score of 107, clinical authorities would generally recommend ignoring the GCA and using the Verbal Ability factor to predict reading achievement. Similarly, clinicians might consider using Nonverbal Reasoning to predict the child's likely performance in mathematics. The important point is that clinical guidelines are oriented toward the selection of the optimal piece of information rather than the combination of multiple pieces of information according to regression weights (especially not involving interaction terms). Although the use of formulas and other aids, such as signal detectability, the Bayesian approach, or decision analysis, can help to model decisions by a variety of regression approaches (Kleinmuntz, 1990), clinicians typically do not rely on such formulas and instead select one test score from several that are available and base their interpretations on that score.

To test the efficiency of actual clinical practice, we selected the optimal predictor variable for each achievement criterion. If an individual did not show any significant discrepancies among factor scores, then we used the GCA as the predictor. If the youth showed Verbal Ability strengths or weaknesses (as compared to either Nonverbal Reasoning or Spatial Ability), then we switched to using Verbal Ability as the predictor for Word Reading and Spelling criteria. If a participant showed Nonverbal Reasoning–Verbal Ability or Spatial-Verbal Ability discrepancies, then we used Nonverbal Reasoning or Spatial Ability as the predictor of Basic Number Skills. If the youth showed discrepancies with both Nonverbal Reasoning and Spatial Ability, then we used the average of Nonverbal Reasoning and Spatial Ability as the predictor.

Table 5 shows the correlation between each "clinically optimized" index and the achievement criterion. Table 5 also presents the correlation between the GCA and the same achievement criterion and the correlation between the GCA and the clinically optimized index. It is possible to test whether the correlations with achievement are significantly different from each other, using the *t* test of dependent correlations (as per Cohen & Cohen, 1983). For Word Reading and Spelling, there is no difference in the performance of the GCA versus the clinically optimized index. For Basic Number Skills, the clinically optimized index actually performs significantly worse than the GCA alone.

TABLE 4
Percentage of Variance Explained by GCA (Alone) and Factor (Alone) in Predicting Academic Achievement

Achievement Criterion	*GCA*	*Factor*
Basic Number Skills	34.7	17.6 (spatial) 32.9 (nonverbal)
Word Reading	35.6	35.8 (verbal)
Spelling	27.3	24.6 (verbal)

NOTE: GCA = General Conceptual Ability.

TABLE 5
Correlations Between Achievement Criteria and General Conceptual Ability or Clinically "Optimized" Index

		r*(Clinical*	*Nuisance*	t *Test of*
Achievement Criterion	r*(GCA)*	*Choice)*	r	*Difference*
Word Reading	.597	.604	.848	0.55
Basic Number Skills	.589	.488	.850	–7.84****
Spelling	.523	.504	.848	1.40

NOTE: *r*(GCA) = correlations between achievement criteria and General Conceptual Ability; *r*(Clinical Choice) = correlations between achievement criteria and clinically "optimized" index; nuisance *r* is the correlation between GCA and the respective clinically "optimized" index. All *t* tests are based on 1, 182 *df.*
****$p < 1.1 \times 10^{-14}$.

DISCUSSION

Our results suggest that statistically significant discrepancies are quite frequent. The average child in the sample exhibited two "significant" discrepancies at the factor or subtest level. Nearly 4 children in 5 demonstrated at least one ability discrepancy when using a conservative 95% confidence approach, and more than 88% of children showed at least one discrepancy when employing the 90% standard recommended in many textbooks. What is particularly striking about this finding is that it comes from a representative, nonreferred sample. If the results are to be taken at face value, then it stands to follow that nearly 80% of average American children will be diagnosed as having some kind of discrepancy, either classified as a deficit or a strength. Thus, when incorporating information about discrepancies, it is important to consider the multivariate base rate in the sample population.

Multiple regression analyses showed that the GCA was the most parsimonious and robust predictor of all three forms of academic achievement. The global cognitive estimate continued to make a unique contribution to achievement prediction even when combined with a factor score, subtest scores, and information about the presence of discrepancies. Multiple regression analyses also failed to detect any instance where interactions between factor scores and reliable discrepancies (or subtests and discrepancies) significantly improved prediction of achievement. This finding was surprising because it represents a statistically sophisticated means of testing the clinical impression that specific abilities should outperform global ability estimates when an individual's specific abilities are not evenly developed. The failure to find a significant interaction is unlikely to be due to low statistical power, as the sample size was large enough to afford power > .90 to detect small effect sizes (e.g., $r = .10$). Null findings are also unlikely to be attributable to sampling bias, as the data come from the standardization sample of a published measure conforming closely to the 1988 to 1990 U.S. census data (McDermott, 1994).

In reality, clinicians do not combine multiple scores from within a test battery to best estimate a child's achievement performance but rather use interpretive guidelines to select one test score as optimal for a particular purpose. When comparing the individualized predictor approach to prediction using the GCA, the clinically optimized predictor performed either equally well as the GCA or significantly worse (in the case of Basic Number Skills). We would argue that "ties" should be awarded to the GCA: Reliance on the GCA is the simplest prediction model, and it concentrates clinical attention on the most reliable and well-validated score (for purposes of predicting academic achievement) from the test battery.

The failure to detect statistically significant improvements in prediction of achievement, even when considering cases that show dramatic (i.e., 30+ point) discrepancies in underlying cognitive abilities, is surprising. Global ability estimates have drawn considerable criticism, and interpretive systems have devoted considerable energy to creating auxiliary or alternative assessment systems. What could explain the apparent robustness of the GCA and the relatively modest performance of the factor scores? Several possibilities include the following: The factor indexes do not contain sufficient unique variance to provide accurate estimates of individual specific abilities as distinct from general ability (Youngstrom & Frazier, 2000), the specific cognitive abilities measured do not possess significant incremental validity for the achievement criteria studied here (although it is an empirical question whether they might demonstrate incremental validity for other criteria), statistically significant differences in ability are apparently commonplace in the population, and finally, a substantial proportion of the apparently significant discrepancies may either be Type I errors (Silverstein, 1993) or artifacts of administrative error (Hanna, Bradley, & Holen, 1981). Another important consideration is the psychometric limitations of difference scores. Differences between correlated scores can show very low levels of reliability, even when the two tests being compared are highly reliable. This reflects the fact that difference scores will possess minimal true score variance and almost entirely reflect measurement error (cf. Caruso, 2001). Relying on a single score, however, such as the GCA, obviates this problem because there is no need to interpret any differences between scores.

In addition to the psychometric limitations, other compelling factors need to be considered when debating the use of more discrete indices. In scientific investigations, the law of parsimony, or having an explanation that invokes as few principles as necessary, is ideal. Interpretations that employ factor and subtest scores are not consonant with this principle—unless they demonstrate marked incremental validity. In addition, the use of more discrete measures significantly contributes to the length of the test administration process (Camara, Nathan, & Puente, 1998). Given the current public policy and managed health care milieu, there are very often severe time constraints and eroding reimbursement rates surrounding psychological assessments (Camara, Nathan, & Puente, 2000; Eisman et al., 1998; Groth-Marnat, 1999). In that sense, the gains from interpreting the more dubious discrete indices might not be commensurate with the expenditures involved in their use. Using more discrete indices to predict academic achievement, even in more specific content areas, leads to more complex models that provide meager dividends (e.g., Grossman & Johnson, 1982; Mishra, 1983).

There are several limitations to the generalizability of the present study. First, the findings are based solely on the DAS as a measure of cognitive ability and academic achievement. Although the DAS has good psychometric properties (e.g., good validity, high reliability), these results would benefit from further cross-validation using additional well-established measures of ability and achievement. Glutting et al. (1997) have provided some initial support for the current findings with similar results for the WISC-III sample. Second, the study only emphasizes the criteria of academic achievement. Given the reliability of factor scores and their inclusion in most top-down approaches, it might be helpful to examine whether the more discrete measures significantly or meaningfully relate to other criteria of interest besides achievement. Factor scores might become more important predictors of educational-vocational criteria as people begin to pursue more specialized educational and vocational training in young adulthood (Lubinski & Benbow, 2000). Third, it is important to recognize that all of the analyses examined here employ concurrent measures. It might be useful to explore whether factor scores provide any incremental predictive advantage beyond the GCA with longitudinal data (e.g., Moffitt, Caspi, Harkness, & Silva, 1993). Fourth, the sample is based on U.S. census information from more than a decade ago. As compared to the 1988 to 1990 U.S. census information, the 2000 census indicates that there are some significant differences in certain demographic characteristics of the U.S. population. For example, of the total population, there is a 6.5% decrease in the Caucasian population, whereas the Hispanic population has increased by 3.5%, so that they currently compose more than 10% of the total population (U.S. Department of Commerce, 2000). These demographic changes may alter the generalizability of the current findings to certain ethnic groups. Finally, the results of the current study are based on a nonclinical population, and it is unclear how well they would generalize to other settings.

The present study has important clinical implications. Results strongly suggest that clinical assessments should concentrate on the most global assessment of cognitive ability when addressing referral questions pertaining to academic achievement. This accords well with other research that has suggested that general intelligence is the most potent and parsimonious predictor of academic performance for K-12 students (Lubinski & Benbow, 2000). Interpretation of discrepancies between factors and subtests does not significantly help in predicting academic achievement, even in specific content areas, and results in models that are more complex, confusing, and time consuming.

We are certainly not implying that a general estimate of cognitive ability is the only piece of information, or even the most important data point, in addressing the clinical needs of children. It appears unlikely, however, that assessment of specific cognitive strengths and weaknesses on the DAS, using either factor or subtest scores, will uncover diagnostic information that could lead to more effective educational interventions. This view is consistent with the largely negative literature on aptitude by treatment interactions (Cronbach & Snow, 1977). If clinicians are intent on accurately predicting children's performance, then it appears that a brief and reliable assessment of general cognitive ability would be sufficient. Evaluators can use an instrument that was specifically designed and normed as a brief ability test, such as the Wechsler Abbreviated Scales of Intelligence (Psychological Corporation, 1999) or the Wide Range Intelligence Test (Glutting, Adams, & Sheslow, 2000). Examiners might also consider using a new, shorter version of the DAS that would include only three subtests. Either the Word Definitions (alpha = .83) or Similarities (alpha = .79) subtests could signify the Verbal Ability factor. These subtests are also compelling because each also loads on highly with *g*, as evidenced by correlations of $r =$ .74 and $r = .75$, respectively, with the GCA. The Sequential and Quantitative Reasoning subtest could well represent the Nonverbal Reasoning factor, both because of its high internal consistency (alpha = .85) and correlation with GCA ($r = .79$). Similarly, the Pattern Construction subtest would strongly represent the Spatial Ability factor, with alpha = .91 and a correlation of $r = .77$ with the GCA. An important caveat to consider is that future research will have to validate this shorter DAS form before it would be appropriate for clinicians to use. Ultimately, any abbreviated ability battery would free up time and resources for diagnostic assessment activities that are more likely to identify remediable weaknesses and to prescribe successful interventions (e.g., for referral questions involving reading difficulty, tests such as the Woodcock Reading Mastery Tests–Revised) (Woodcock, 1987).

NOTES

1. Elliot (1990a) called the three factors “clusters” in the manual, even though exploratory and confirmatory factor analyses were the empirical basis for generating the scales.

2. Discrepant scores were calculated using the geometric mean or difference score formula, where differences required for statistical significance are based on the standard errors of measurement for each index scale as well as the *z* score under the normal curve that is associated with the desired significance level. The formula is Difference Score = $Z \sqrt{SE_{ma}^{2} + SE_{mb}^{2}}$.

3. See Note 2.

4. In the present study, we used interaction terms to examine if there was a differential predictive relationship of academic achievement between those individuals who demonstrated clinically significant or rare discrepancies between factor and subtest scores and those who did not. An alternative approach would be to take all participants with the discrepancy or discrepancies of interest and compare their academic scores with those of a subgroup having the same General Conceptual Abilities but no discrepancies. Supplemental analyses were conducted using this paired matching technique. Specifically, youths showing rare or clinically significant factor discrepancies (less than 5% population prevalence) were matched with controls, drawn from an epidemiological sample of 1,400, on overall cognitive ability and demographics. Three academic achievement criteria were used (Word Reading, Number Skills, Spelling) with four groups showing ability discrepancies (Verbal Ability > Nonverbal Reasoning, Nonverbal Reasoning > Verbal Ability, Verbal Ability > Spatial Ability, Spatial Ability > Verbal Ability) and matched controls. The *n*s for each group ranged from 67 to 75, and *t* values fell between –3.406 and 1.941. Results indicate that no means showed reliable differences when compared with a Bonferroni-adjusted critical value of $p < .0042$ (correcting for 12 comparisons—three achievement scores with four matched groups), with the possible exception of strengths on Verbal Ability as compared to Nonverbal Reasoning being associated with modestly higher Word Reading ($p < .001$). These results are convergent with and serve to validate the findings of the present study.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park: Sage.

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review*, *29*, 52-64.

Beebe, D. W., Pfiffner, L. J., & McBurnett, K. (2000). Evaluation of the validity of the Wechsler Intelligence Scale for Children-Third Edition Comprehension and Picture Arrangement subtests as measures of social intelligence. *Psychological Assessment*, *12*, 97-101.

Buchner, A., Faul, F., & Erdfelder, E. (1996). *G–Power: A priori, post-hoc, and compromise power analyses for the Macintosh* (Version 2.1.1). Trier, Germany: University of Trier.

Cahan, S. (1986). Significance testing of subtest score differences: The rules of the game. *Journal of Psychoeducational Assessment*, *4*, 273-280.

Camara, W., Nathan, J., & Puente, A. (1998). *Psychological test usage in professional psychology: Report of the APA practice and science directorates*. Washington, DC: American Psychological Assocation.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, *31*, 141-154.

Carroll, J. B. (2000). Commentary on profile analysis. *School Psychology Quarterly*, *15*, 449-456.

Caruso, J. C. (2001). Increasing the reliability of the fluid/crystallized difference score from the Kaufman Adolescent and Adult Intelligence Test with reliable component analysis. *Assessment*, *8*, 155-166.

Ceci, S. J., & Williams, W. M. (1997). Schooling, intelligence, and income. *American Psychologist*, *52*, 1051-1058.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Donders, J. (1996). Factor subtypes in the WISC-III standardization sample: Analysis of factor index scores. *Psychological Assessment*, *8*, 312-318.

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., et al. (1998). *Problems and limitations in the use of psychological assessment in contemporary health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, part II*. Washington, DC: American Psychological Association.

Elliot, C. D. (1990a). *Differential Ability Scales: Introductory and technical handbook*. San Antonio, TX: Psychological Corporation.

Elliot, C. D. (1990b). The nature and structure of children's abilities: Evidence from the Differential Ability Scales. Conference on Intelligence: Theories and practice (1990, Memphis, Tennessee). *Journal of Psychoeducational Assessment*, *8*, 376-390.

Elliot, C. D. (1990c). The nature and structure of the DAS: Questioning the test's organizing model and use. *Journal of Psychoeducational Assessment*, *8*, 406-411.

Glutting, J. J., Adams, W., & Sheslow, D. (2000). *Wide Range Intelligence Test manual*. Wilmington, DE: Wide Range.

Glutting, J. J., McGrath, E. A., Kamphaus, R. W., & McDermott, P. A. (1992). Taxonomy and validity of subtest profiles on the Kaufman Assessment Battery for Children. *Journal of Special Education*, *26*, 85-115.

Glutting, J. J., Watkins, M., & Youngstrom, E. A. (in press). Multifactored and cross-battery assessments: Are they worth the effort? In C. R. Reynolds & R. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (2nd ed.). New York: Guilford.

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. L. (1997). Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment*, *9*, 295-301.

Gough, H. (1971). Some reflections on the meaning of psychodiagnosis. *American Psychologist*, *26*, 106-187.

Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice*. Boston: Allyn & Bacon.

Grossman, F. M., & Johnson, K. M. (1982). WISC-R factor scores as predictors of WRAT performance: A multivariate analysis. *Psychology in the Schools*, *19*, 465-468.

Groth-Marnat, G. (1999). Financial efficacy of clinical assessment: Rational guidelines and issues for future research. *Journal of Clinical Psychology*, *55*, 813-824.

Hanna, G. S., Bradley, F. O., & Holen, M. C. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our heads out of the sand. *Journal of School Psychology*, *19*, 370-376.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence*. Boston: Allyn & Bacon.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York: John Wiley.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: John Wiley.

Kaufman, A. S., & Lichtenberger, E. O. (1999). *Essentials of WAIS-III assessment*. New York: John Wiley.

Keith, T. Z. (1990). Confirmatory and hierarchical confirmatory analysis of the Differential Ability Scales. *Journal of Psychoeducational Assessment*, *8*, 391-405.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin*, *107*, 296-310.

Lipsitz, J. D., Dworkin, R. H., & Erlenmeyer-Kimling, L. (1993). Wechsler Comprehension and Picture Arrangement subtests and social adjustment. *Psychological Assessment*, *5*, 430-437.

Lubinski, D., & Benbow, C. P. (2000). States of excellence. *American Psychologist*, *55*, 137-150.

McDermott, P. A. (1994). *National profiles in youth psychopathology: Manual of Adjustment Scales for Children and Adolescents*. Philadelphia, PA: Edumetric and Clinical Science.

McDermott, P. A., & Glutting, J. J. (1997). Informing stylistic learning behavior, disposition, and achievement through ability subtests—or more illusions of meaning? *School Psychology Review*, *26*, 163-175.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Boston: Allyn & Bacon.

Mishra, S. P. (1983). Validity of WISC-R IQs and factor scores in predicting achievement for Mexican-American children. *Psychology in the Schools*, *20*, 150-154.

Moffitt, T. E., Caspi, A., Harkness, A. R., & Silva, P. A. (1993). The natural history of change in intellectual performance: Who changes? How much? Is it meaningful? *Journal of Child Psychology and Psychiatry*, *14*, 455-506.

Naglieri, J. A. (1993). Pairwise and ipsative comparisons of WISC-III IQ and index scores. *Psychological Assessment*, *5*, 113-116.

Naglieri, J. A. (2000). Can profile analysis of ability tests work? An illustration using the PASS theory and CAS with an unselected cohort. *School Psychology Quarterly*, *15*, 419-433.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77-101.

Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. A. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly*, *15*, 376-385.

Prifitera, A., Weiss, L. G., & Saklofske, D. H. (1998). The WISC-III in context. In A. Prifitera & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 1-39). New York: Academic Press.

Pritchard, D. A., Livingston, R. B., Reynolds, C. R., & Moses, J. A., Jr. (2000). Modal profiles for the WISC-III. *School Psychology Quarterly*, *15*, 400-418.

Psychological Corporation. (1992). *Wechsler Individual Achievement Test manual*. San Antonio, TX: Author.

Psychological Corporation. (1999). *Wechsler Abbreviated Scale of Intelligence manual*. San Antonio, TX: Author.

Riccio, C. A., Cohen, M. J., Hall, J., & Ross, C. M. (1997). The third and fourth factors of the WISC-III: What they don't measure. *Journal of Psychoeducational Assessment*, *15*, 27-39.

Riccio, C. A., & Hynd, G. W. (2000). Measurable biological substrates to verbal-performance differences in Wechsler scores. *School Psychology Quarterly*, *15*, 386-399.

Sattler, J. (1992). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Sattler, J. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Author.

Schwean, V. L., Saklofske, D. H., Yackulic, R. A., & Quinn, D. (1993). WISC-III performance of ADHD children. In B. A. Bracken & R. S. McCallum (Eds.), *Wechsler Intelligence Scale for Children* (3rd ed., pp. 56-70). Brandon, VT: Clinical Psychology Publishing.

Silverstein, A. B. (1993). Type I, Type II, and other types of errors in pattern analysis. *Psychological Assessment*, *5*, 72-74.

Stanton, H. C., & Reynolds, C. R. (2000). Configural frequency analysis as a method of determining Wechsler profile types. *School Psychology Quarterly*, *15*, 434-448.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale–Fourth Edition*. Chicago: Riverside.

U.S. Department of Commerce. (2000). *Profile of general demographic characteristics for the United States* (Current Population Reports). Washington, DC: Bureau of the Census.

Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly*, *15*, 465-479.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children–Revised Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children–Third Edition*. San Antonio, TX: Psychological Corporation.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests–Revised: Examiner's manual*. Circle Pines, MN: American Guidance Service.

Youngstrom, E., & Frazier, T. W. (2000, December). *Evidence and implications of over-factoring on commercial tests of cognitive ability*. Paper presented at the Annual Meeting of the International Society for Intelligence Research, Cleveland, OH.

Youngstrom, E. A., & Glutting, J. J. (2001). *Individual strengths and weaknesses on factor scores from the Differential Ability Scales: Validity in predicting concurrent achievement and behavioral criteria*. Manuscript submitted for publication.

Youngstrom, E. A., Kogos, J. L., & Glutting, J. (1999). Incremental efficacy of Differential Ability Scales factor scores in predicting individual achievement criteria. *School Psychology Quarterly*, *14*, 26-39.

Shoshana Y. Kahana is currently a graduate student in the clinical psychology doctoral program at Case Western Reserve University. She holds a B.A. from the University of Pennsylvania in Psychology. Her current research interests include the interpretation of cognitive and academic achievement performance in addition to the relationship between maternal affect and ratings of child functioning.

Eric A. Youngstrom, Ph.D., is currently an assistant professor at Case Western Reserve University. His research interests are in measurement and clinical assessment as well as the appropriate interpretation of individual test data. In addition, his research also examines the use of clinical instruments and the integration of multiple sources of data to assess individuals' emotional experiences.

Joseph J. Glutting, Ph.D., is currently a professor in the School of Education at the University of Delaware. His reseach interests include issues related to school psychology, psychoeducational assessment, and educational measurement.

Confirmatory Factor Analysis of Single- and Multiple-Factor Competing Models of the Dissociative Experiences Scale in a Nonclinical Sample

Gary D. Stockdale
Betty E. Gridley
Deborah Ware Balogh
Thomas Holtgraves
Ball State University

Previous research on the Dissociative Experiences Scale (DES) has demonstrated that (a) dissociation is quantifiable in both clinical and nonclinical samples and (b) a three-factor structure (amnesia, depersonalization, and absorption) is tenable for clinical samples. The factor structure for nonclinical samples is less clear, with one- and multiple-factor solutions proposed. To clarify the DES factor structure in nonclinical samples, confirmatory factor analyses were conducted on (a) one-, two-, three-, and four-factor first-order models and (b) two bifactor (hierarchical) models of DES scores for two samples of nonclinical university students. Results of $\Delta\chi^2$ *and goodness-of-fit indices support the three-factor (first-order) model as best fitting of the data for these samples. The utility of this DES model for screening both dissociative pathology and elevated normal dissociative behavior in clinical and nonclinical populations is discussed.*

Keywords: dissociation, factor analysis, DES, CFA

Dissociation describes complex human behavior that results in the disruption of normally integrated memory, thought, behavior, sensation, or affect (American Psychiatric Association, 1994, p. 766; Braun, 1988b). An individual can express dissociation in three domains of behavior, composed of (a) a personality trait (Kihlstrom, Glisky, & Angiulo, 1994) manifested primarily by normal, benign experiences (e.g., daydreams and automatisms); (b) a defense mechanism used to buffer against stress and trauma, including war and natural disasters (e.g., Cardeña & Spiegel, 1993) and interpersonal abuse (e.g., B. Sanders & Giolas, 1991); and (c) both symptoms and syndromes of psychopathology, occurring in certain clinical and personality disorders (American Psychiatric Association, 1994). Moreover, the dissociative construct is clearly multifaceted, although not often explicitly described as such by researchers and theoreticians. For example, in the Behavior, Affect, Sensation, and Knowledge (BASK) theoretical model of dissociation (Braun, 1988a), depersonalization is described as an alteration in the sensation of oneself, whereas amnesia is primarily a disruption in self-knowledge. Thus, for the BASK model, alterations in one or more of the four components conceptually describe different facets of the dissociative construct.

In addition to theoretical conceptualizations of dissociation and its subsumed facets, several self-report scales have been recently developed to empirically quantify dissociative phenomena among individuals (Bernstein-Carlson & Putnam, 1986; Riley, 1988; S. Sanders, 1986). The Dissociative Experiences Scale (DES) (Bernstein-

Correspondence concerning this article and requests for offprints should be addressed to Gary D. Stockdale, Department of Psychology, University of California–Davis, One Shields Avenue, Davis, CA 95616-8686.

Assessment, Volume 9, No. 1, March 2002 94-106

Carlson & Putnam, 1986; Carlson & Putnam, 1993) is a self-report measure of dissociation frequently used for research on dissociative behavior and as a dissociative-pathology screening instrument in both clinical and nonclinical populations. One area that has been widely investigated by researchers who use the DES is the structure (facets or factors) of dissociative phenomena manifested by different populations. To date, exploratory factor analyses (EFAs) of DES scores have produced a panoply of results and interpretations dependent on participant characteristics (clinical vs. nonclinical), statistical procedures (principal components, Tobit factor analysis, etc.), and researchers' theoretical orientations.

For clinical and mixed clinical-nonclinical samples, three-factor solutions of DES scores generally have been the norm (Carlson et al., 1991; Ross, Ellason, & Anderson, 1995), although a four-factor solution has been identified in a substance abuse sample (Dunn, Ryan, & Paolo, 1994), and a one-factor (Marmar et al., 1994) as well as a four-factor (Amdur & Liberzon, 1996) solution have been offered for two separate groups diagnosed with post-traumatic stress disorder (PTSD). For strictly nonclinical samples, factor analytic interpretations of DES scores have been more equivocal with one-factor (Fischer & Elnitsky, 1990; Holtgraves & Stockdale, 1997), three-factor (Ross, Joshi, & Currie, 1991; B. Sanders & Green, 1994), four-factor (Ray & Faith, 1995), and seven-factor (Ray, June, Turaj, & Lundy, 1992) solutions proposed.

Although these DES exploratory factor analytic studies have produced diverse outcomes and interpretations, three general themes transverse all results. First, for nonclinical samples, a large first factor accounting for a substantial amount of the total variance was identified in all studies. Second, for nonclinical samples where multiple factors were proposed, the first factor was always identified as absorption or a synonymic variant. Third, for both clinical and nonclinical samples where three or more factors were proposed, the factors invariably included absorption, amnesia, and depersonalization. The one- and three-factor findings corroborated, in part, a factor analysis of DES scores across many protocols done by Waller (1995) where the Tobit procedure (Muthén, 1989) was used to account for the positively skewed response profiles of the DES items. This analysis supported a large, general dissociation factor that subsumed absorption along with separately identified, but not reliably measurable, amnesia and derealization subordinate factors (Waller, 1995). Thus, no definitive consensus has yet been reached among researchers and theoreticians on the number of factors actually measured by the DES, particularly for nonclinical samples.

What has not been done—or at least reported—in factor-analytic studies of DES scores is a confirmatory factor analysis (CFA) of competing models, with each model having a different number of factors. CFA offers researchers statistical techniques for optimal model selection not achievable in EFA. For example, in EFA the decision on the number of factors to retain is often subjective, based on the researcher's mental weighing of often disparate information such as scree tests, eigenvalues greater than one, and percentage of total variance explained by the factors. In CFA, by contrast, statistical tests and formulae, such as change in χ^2 and fit indices (e.g., goodness-of-fit index [GFI]), are used to directly and quantitatively compare preset models of, say, differing number of factors. Thus, CFA resolves conundrums such as deciding which one of two apparently tenable models best fits the data or, contrarily, which model is the best alternative when no model robustly supports the data. Based on the superior attributes of CFA compared to EFA, the research presented herein elucidates the factor structure of the DES using CFA methodology on the DES scores of two independent, university-aged samples of nonclinical participants.

METHOD

Participants

The DES was administered to two samples of midwestern university students. The first sample was a convenience sample of 971[1] university undergraduates (69% female, age *M* = 19.2 years, *SD* = 2.53 years, range 18-45 years) who gave informed consent and then voluntarily and anonymously participated in the study in small groups of 5 to 40 individuals to satisfy a course requirement of introductory psychology at Ball State University, Muncie, Indiana. Participants completed the DES that was embedded in a battery of nine self-report personality measures used for a more comprehensive study conducted from 1995 through 1998. The second sample was 400 undergraduates (56% female) who completed the DES and other self-report behavioral measures (different measures from the first sample) from 1993 to 1994 to satisfy the same course requirement. The 971-participant sample was used to develop all the reported psychometrics; the 400-participant sample was used only as an independent confirmation sample for the final confirmatory factor-analytic competing-models comparisons.

The DES

The DES is a 28-question, paper-and-pencil, self-report scale used to measure dissociation and screen for dissociative pathology (Bernstein-Carlson & Putnam, 1986; Carlson & Putnam, 1993). In its present format, respondents indicate what percentage of the time they had experienced the dissociative situation described in each question by circling 1 of 11 response options: 0% to 100% in increments of 10%. The DES usually takes less than 15 minutes to administer and score.

The DES is the most frequently used and best validated of the self-report scales of dissociation (Kluft, 1993; Waller, 1995). As of 1996, the reported studies using the DES have included scores of more than 12,000 participants (van Ijzendoorn & Schuengel, 1996). The DES has adequate temporal reliability, with test-retest correlation coefficients between .79 and .84 (Carlson & Putnam, 1993). Internal consistency is good to excellent with reported split-half coefficients from .83 to .93 (Bernstein-Carlson & Putnam, 1986; Pitblado & Sanders, 1991) and an average alpha coefficient of .93 across six studies (Waller, 1995).

Validity-related evidence of the DES has supported the dissociative construct and criteria of dissociative experiences and disorders. Criterion-related validity has been established by demonstrating that the gradient of scores on the DES is related to the severity of dissociation, from dissociative identity disorder (DID, formerly multiple personality disorder) at the uppermost end of scores to normal adults at the lowest end (Bernstein-Carlson & Putnam, 1986; Carlson & Putnam, 1993). Although the DES generally is not used to diagnose dissociative disorders, scores above 40 are highly indicative of DID, and scores above 30 are indicative of severe dissociative pathology often present in PTSD and certain severe personality disorders (Allen, Coyne, & Console, 1997; Carlson & Putnam, 1993).

Convergent-related validity of the DES to other self-report measures of dissociation and related behavioral phenomena (e.g., Perceptual Alteration Scale, Tellegan Absorption Scale) (S. Sanders, 1986; Tellegan & Atkinson, 1974) was clearly demonstrated in a meta-analysis by van Ijzendoorn and Schuengel (1996) where an overall correlation of .67 was reported across 26 studies with a total *N* of 5,916. Furthermore, factor analyses of DES scores have supported the construct validity of the DES vis-à-vis dissociation and many of its putative facets (e.g., absorption, amnesia, and depersonalization) (Carlson & Putnam, 1993), although, as mentioned previously, there is at least a modicum of confusion concerning the composition of the factor structure.

Data Analyses

Descriptive statistics and EFA were conducted using the Statistical Package for the Social Sciences (SPSS) 9.0 for Windows (SPSS Inc., 1998), and CFAs were conducted using AMOS 4 Graphics (Arbuckle, 1999). The *M*s and *SD*s of the 28 DES items for the 971-participant sample are listed in Appendix A, and the DES interitem correlation matrix for this sample is listed in Appendix B.

RESULTS AND CONCLUSIONS

EFA

Principal axis extraction and oblimin rotation[2] were used to investigate the initial factor structure (Gorsuch, 1997) of the 28 DES items for the 971-participant sample. Preliminary examination of the initial EFA solution—where factors with eigenvalues greater than one were retained—indicated one- to four-factor structures all were tenable, based on various factor-retention criteria. According to the eigenvalue-greater-than-one criterion (Kaiser, 1960), four factors qualified for inclusion; however, the scree-plot criterion (Cattell, 1978, p. 7) suggested a two-factor structure of the DES. Moreover, a single factor was also certainly reasonable based on the disproportionately large amount of initial (i.e., before extraction) variance accounted for by the first factor compared to subsequent factors: 39.8% for the first factor (initial eigenvalue 11.15) versus 7.3% for the second factor (eigenvalue 2.05), 4.1% for the third factor (eigenvalue 1.14), and 3.8% for the fourth factor (eigenvalue 1.05).

Next, to determine the factor structures and factor loadings of the DES items for the one-, two-, and three-factor structures, principal-axis extraction and oblimin rotation were again conducted for the 971-participant sample but with the number of factors preset to one, two, and then to three. (The corresponding four-factor solution values were obtained in the initial, eigenvalue-greater-than-one solution.) Here, after extraction, the cumulative percentages of total variance[3] for the four solutions were one factor, 37.7%; two factor, 43.3%; three factor, 45.8%; and four factor, 47.7%. Individual factor statistics of the one-, two-, three-, and four-factor solutions are listed in Table 1, where again it can be seen that each solution is viable. The single-factor solution accounted for a large proportion of the variance, yet the two-, three-, and four-factor solutions moderately increased the total variance explained by the DES scores. Also, all the interfactor correlations in each of the two- and three-factor solutions were substantively large, yet with no two factors overcorrelated (i.e., a coefficient above .80) that would indicate two factors measuring essentially the same construct (Whitley, 1996). Moreover, for solutions of three factors or less, the smallest individual item loading on the apposite rotated factor in the pattern matrix was .33 for one item in the two-factor solution and .35 for one item in the three-factor solution. Thus, for the one-, two-, and three-factor solutions, no item had less

TABLE 1
Exploratory Factor Analysis of DES Scores: Principal Axis Extraction and Oblimin Rotation

Factor Solution	*Number of DES Items*	*Eigenvalue (Extracted)*	*Percentage of Variance (Extracted)*	*Correlation to Factor 1*	*Correlation to Factor 2*	*Correlation to Factor 3*
One factor						
1	28	10.54	37.66			
Two factor						
1	15	10.61	37.86			
2	13	1.53	5.45	.67		
Three factor						
1	16	10.63	37.96			
2	6	1.56	5.56	.55		
3	6	0.63	2.26	.55	.56	
Four factor						
1	12	10.64	38.02			
2	7	1.57	5.62	.55		
3	7	0.66	2.34	.54	.65	
4	2	0.48	1.71	.43	.25	.25

NOTE: DES = Dissociative Experiences Scale. N = 971 university undergraduates. Determinant of correlation matrix = .0001; Kaiser-Meyer-Olkin Measure of Sampling Adequacy = .95; Bartlett's Test of Sphericity, approximate χ^2 = 11,334, df = 276, p < .001.

than 10% of its total variance accounted for by a single factor. (See Appendix A for factor loadings of each DES item for the three-factor solution.)

Further examination of the initial EFA solution suggested the fourth factor might be extraneous. An extraneous factor is one that has a paucity of items or that minimally adds to the total variance explained by the scale scores. In our four-factor solution, the initial eigenvalue of the fourth factor was 1.05, very close to the Kaiser (1960) cutoff criterion of 1.00. Moreover, only two DES items loaded on the fourth factor in the pattern matrix—the doubleton-item factor eschewed by Gorsuch (1983) and others—and both these items had factor loadings less than .30.

Considering all four EFAs of the 971-participant sample, then, it was concluded that from a quantitative perspective the one-, two-, and three-factor solutions of the DES scores were defensible. Furthermore, the four-factor solution, although less defensible, could not be summarily rejected because it added incrementally almost as much to total variance explained as did the three-factor solution (1.9% vs. 2.5%). A similar conclusion was entertained when substantive meaning of the factors was considered. First, the single-factor structure identified a large, general, and common dissociation factor. Second, based on experiences described by each of the DES items (see Table 4), the three-factor structure satisfactorily demarcated the DES items into an absorption factor, an amnesia factor, and a depersonalization factor. Third, except for one item, the two-factor structure mimicked the three-factor structure with Factors 2 and 3 combined into a single factor accounting for both amestic experiences and feelings of altered reality. Thus, additional analysis clearly was needed to determine the optimum factor structure of the DES from a quantitative perspective. To that end, CFA was employed.

CFA

Maximum likelihood CFA of the DES scores for the 971-participant sample was conducted comparing the one-, two-, three-, and four-factor competing first-order models, the factor structure of each having been generated in EFA. Moreover, because the first-order factors were substantially intercorrelated and because the first extracted factor accounted for a preponderance of the explained variance in each solution, two hierarchical bifactor models were also run to evaluate the role of a general factor in the explanation of total DES variance accounted for by the participants' responses (Yung, Thissen, & McLeod, 1999). By comparing appropriate fit statistics of a bifactor model to the parallel first-order factor model, it is possible to determine whether a unique hierarchical general factor or, alternatively, the intercorrelations among the first-order factors best account for general common variance among the scale items.

CFA analysis of the four first-order models showed (a) a substantial reduction in total χ^2, (b) a reduction in χ^2/df, (c) a reduction in root mean square error of approximation (RMSEA) (Steiger, 1990), and (d) higher fit index values (e.g., GFI, Tucker-Lewis Index [TLI]) as the number of factors increased from one to three, and the reversal, or an increase in χ^2 and χ^2/df for the four-factor model compared to the three-factor model (see Table 2). The χ^2s of the four models can be compared because each multifactor model

TABLE 2
DES Confirmatory Factor Analysis Fit Statistics Comparing One, Two, Three, and Four First-Order Factor and Two Bifactor Competing Models

Model	χ^2	df	χ^2/df	*RMSEA*	*GFI*	*PGFI*	*PCFI*	*TLI*
Independence	13,417	378	35.49	.189	.20	.19	.00	.00
One factor	3,152	350	9.01	.091	.75	.65	.73	.77
Two factor	2,220	349	6.36	.074	.85	.73	.79	.85
Three factor	1,973	347	5.69	.070	.87	.74	.80	.86
Four factor	1,997	344	5.81	.070	.87	.73	.80	.86
Bifactor 2	2,602	340	7.65	.083	.84	.70	.74	.81
Bifactor 3	2,542	325	7.82	.084	.84	.67	.71	.80

NOTE: DES = Dissociative Experiences Scale; RMSEA = root mean square error of approximation; GFI = goodness-of-fit index; PGFI = parsimony goodness-of-fit index; PCFI = parsimony comparative fit index; TLI = Tucker-Lewis Index. *N* = 971 university undergraduates.

is nested within similar models of fewer factors, and all models are nested within the one-factor model (Keith, 1997). The nesting is demonstrated by setting the intercorrelations among the factors of the multiple-factor models to one. By doing so, the two-factor model collapses to the one-factor model, and the three-factor model collapses to either the two- or the one-factor model. Figure 1 illustrates the outcome of this procedure for the three- versus one-factor model. The decreases in χ^2 ($\Delta\chi^2$) from the one-factor model to the two-factor model ($\Delta\chi^2$ 952, 1 *df*) and then from the two-factor model to the three-factor model ($\Delta\chi^2$ 247, 2 *df*) are both substantial and statistically significant (all $\chi^2 p$s < .001). Again, the $\Delta\chi^2$ from the three-factor model to the four-factor model increased ($\Delta\chi^2$ 24, 3 *df*, $p < .001$), illustrating a decrease in model fit for the four-factor model. These statistics advocate the three-factor model to be the best fitting model of those compared and provide one line of evidence for the superiority of the three-factor model.

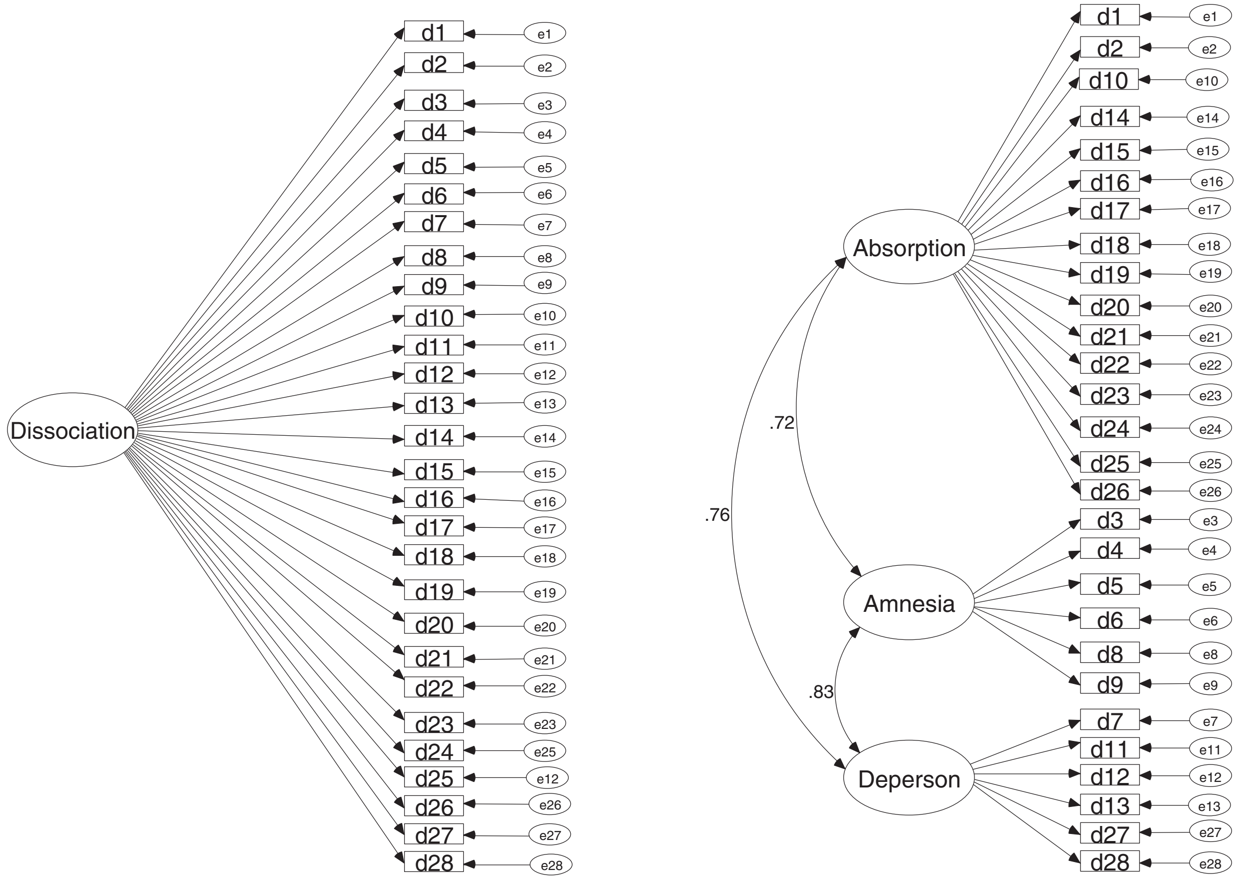
The superiority of the three-factor model over models of fewer factors is also supported by CFA fit statistics (see Table 2). The GFI statistic compares each model with a perfectly fitting or just-identified model, whereas the comparative fit index (CFI) and TLI compare each model to the independence model, a model with no assumed relationships among the measured variables. The TLI is unrelated to sample size, and the CFI estimates the (sample) model fit to the population. The parsimony goodness-of-fit index and the parsimony comparative fit index are also reported in Table 2. These two indexes account for model parsimony by introducing penalties for model complexity, that is, models having more parameters and fewer degrees of freedom (Keith, 1997)—the four- and three-factor models in our analysis. For all the fit indexes reported in Table 2, values closer to 1.00 represent a better fitting model. All these values are largest for the three- and four-factor models when compared to the other models in the analysis.

The RMSEA is better (i.e., lower) for the three- and four-factor models (.070 for each) than for either the two- or one-factor models (.074 and .091, respectively). However, an RMSEA value of .07 is above the recommended maximum value of .05 needed for attaining a close-fitting model in relation to degrees freedom (Browne & Cudeck, 1993). And, indeed, also not attained are fit-statistic (GFI, TLI, etc.) values above the recommended minimum of .90 suggested for a psychometrically reasonable model fit (Crowley & Fan, 1997)—although the GFI and TLI are larger than .85 for the three- and four-factor models. The purpose of our research was model comparison with an extant scale, not scale development. Accordingly, modification procedures, either in EFA or CFA, to improve the model fit were not investigated. Even so, the statistics of the untampered three-factor model are meritorious across several CFA criteria and clearly superior to the corresponding statistics for the models of fewer factors.

Superiority of the three-factor model over the one-factor model is further demonstrated by the magnitude of the standardized regression weights generated in CFA. Table 3 shows that except for two DES items (22 and 26) in the absorption factor, the item standardized regression weights are all larger for the three-factor model than for the one-factor model. This is particularly true for the Amnesia and Depersonalization factors where the increase is substantially larger than for the Absorption factor. Because the Amnesia and Depersonalization factors represent the essence of the three-factor solution—otherwise they would be subsumed within the larger first factor—their differentially larger standardized regression weights support the soundness of conceptually separating Amnesia and Depersonalization from a single, general dissociation factor.

To further test the role of a large general factor in the accounting of the DES variance, two bifactor models (Yung et al., 1999) were generated and evaluated against the first-order three-factor model. In the first model, the Bifactor 2 model, all 28 items load on the dissociation general factor in the second factor layer, whereas the 6 amnesia and 6 depersonalization items load on their respective first-order factors in the first factor layer. In the second model, the

FIGURE 1
One-Factor and Three-Factor Competing Models of the DES With Factor Structures Identified by DES Item Numbers and Factor Intercorrelations



NOTE: DES = Dissociative Experiences Scale.

Bifactor 3 model, all 28 items load on the dissociation general factor in the second factor layer, whereas the 6 amnesia, 6 depersonalization, and 16 absorption items load on their respective first-order factors in the first factor layer. Furthermore, in bifactor hierarchical models (used here), the second-order factor does not load on the first-order factors. The two models used in our analysis are illustrated in Figure 2.

In bifactor analysis, the factors in each layer are orthogonal to each other; thus, the general factor can be compared directly to first-order factors for variance partitioning and (quantitative) factor identification. However, the purpose of these analyses for our research was to determine whether a model with a general factor offered a superior explanation to the first-order (only) factor models. As can be seen from Table 2, neither bifactor model offered an improvement over the first-order three-factor model when the same comparative fit statistics and indices used in the first-order (only) comparisons are examined (e.g., $\Delta\chi^2$, GFI, etc.). Thus, a general, separate dissociation general factor with an additional breadth layer (Humphreys, 1981) of first-order factors was rejected as the optimal solution, and the common variance among all 28 DES items was deemed best explained by the intercorrelations among the three first-order factors.

Before advocating a particular model as an optimal solution based on factor analysis, it is prudent to evaluate the descriptive statistics and internal consistency of the proposed factors. Table 4 reports this information for the one- and three-factor (first-order) models. The one-factor model is included (a) to compare the three-factor model to the full-scale values and (b) because a general dissociation factor is still arguably plausible from a theoretical perspective, that is, apart from the statistical analyses reported herein. Regarding internal consistency, the Cronbach's alpha of each retained factor must be robust if the factor is to have substantive meaning when used as a subscale to measure some facet of the construct. Referring to Table 4, it is evident that each factor of the three-factor model (as well as the full scale, or single factor) generated adequate reli-

TABLE 3
DES Items and Standardized Regression Weights for Three- and One-Factor Confirmatory Factor Analysis Solutions

	SRW	
DES Item	*Three Factor*	*One Factor*
Absorption		
1. Did not remember all or part of a car or bus trip	.49	.47
2. While listening, did not hear all or part of a conversation	.59	.54
10. Accused of lying when person thought truth was told	.59	.58
14. Remembered past event vividly, seemed like reliving it	.61	.57
15. Not sure if past events actually happened or were just dreamed	.72	.65
16. Experienced being in a familiar place as strange and unfamiliar	.74	.73
17. Absorbed in TV or movie story, unaware of surrounding events	.63	.58
18. So involved in fantasy or daydream that it felt real	.72	.68
19. Able to ignore pain	.45	.44
20. Stared into space, thought of nothing, unaware of time	.65	.61
21. Talked out loud to self when alone	.55	.51
22. Acted differently in different situations, like two people	.64	.65
23. Did difficult things easily	.55	.52
24. Not sure if something happened or just thought it had happened	.75	.71
25. Evidence of doing something but did not remember doing	.76	.74
26. Found writings not remembered as having written	.68	.70
Amnesia		
3. Found self in place but no memory of having got there	.71	.60
4. Dressed in clothes but not remembering having put them on	.76	.60
5. Found new things among belongings but not remembering buying them	.73	.61
6. Approached by strangers who said they know you	.54	.53
8. Sometimes did not recognize friends or family	.71	.63
9. No memory of some important personal events (e.g., graduation)	.57	.54
Depersonalization		
7. Felt and watched self as if looking at another person	.74	.66
11. Did not recognize self when seen in a mirror	.71	.63
12. Felt other people and objects were not real	.75	.65
13. Felt body was not one's own	.76	.64
27. Heard voices inside head who have told one what to do	.61	.60
28. People and objects appeared distant and unclear, seen through a fog	.74	.69

NOTE: DES = Dissociative Experiences Scale; SRW = standardized regression weight. *N* = 971 university undergraduates.

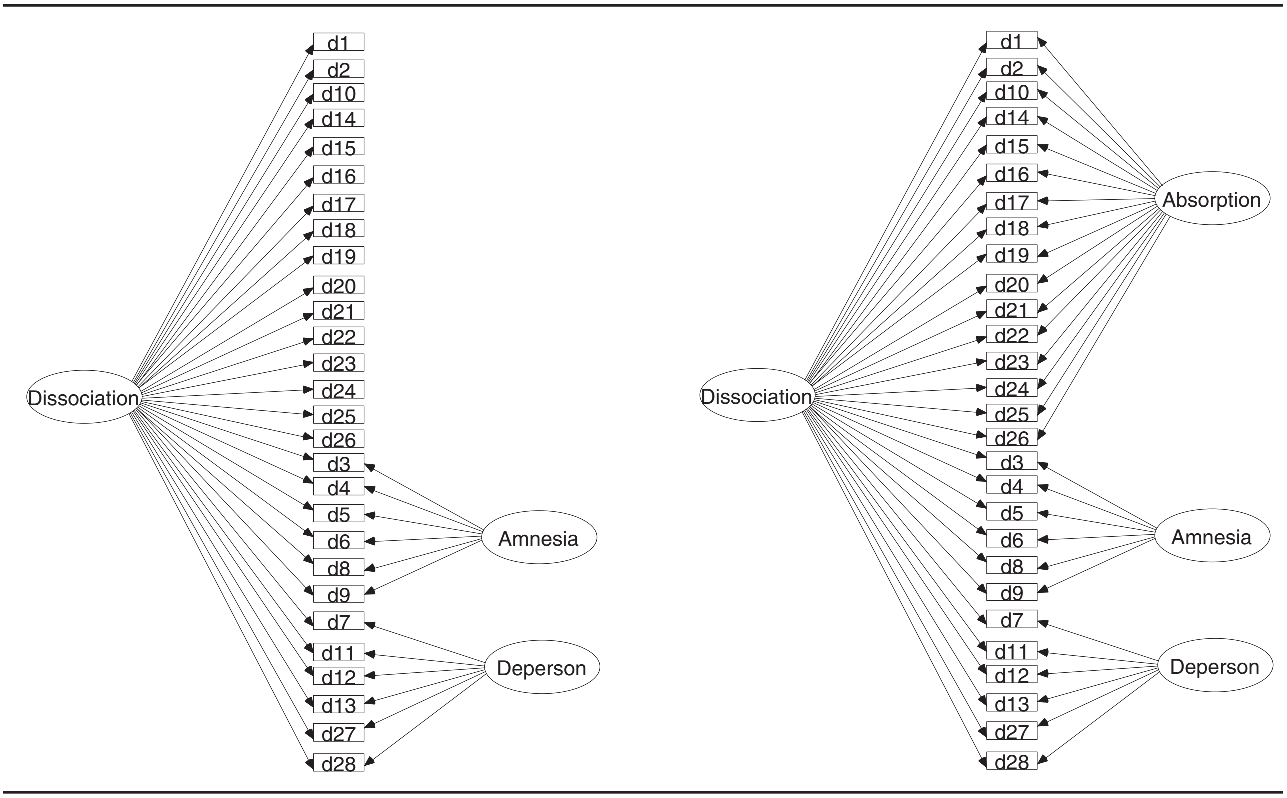
ability for this 971-participant sample. The smallest Cronbach's alpha value was .81 for the amnesia factor of the three-factor model, and this alpha slippage vis-à-vis the single factor certainly reflected, in part, the paucity of items for this factor, six. Also, the robustness of a factor vis-à-vis the total items of the scale should be reflected by higher interitem correlations for the item subsets of the factors rather than the interitem correlations for the full scale. Again, this was true in our analysis where the mean of interitem correlations increased from .37 for the single factor (full scale) to .40 for absorption, .44 for amnesia, and .51 for depersonalization. Consequently, ancillary statistics to EFA and CFA also support the superiority of the three-factor solution.

Finally, and perhaps most important, the superiority of the three-factor model was demonstrated by confirmation of an independent participant sample of 400 university students, the results shown in Table 5. The results of this second analysis were remarkably congruent with the 971-participant sample results reported in Table 2. Thus, confirmation of the optimum first-order three-factor model was achieved using a separate sample from the population of midwestern university undergraduates.

Several attributes of this second participant sample and the statistics generated in both EFA and CFA merit comment. First, the second sample is truly separate and independent, not a random subsample of the first sample. Second, EFA of the second sample (results not reported) indicated a similar but not identical factor structure to the first sample. Even so, when DES item scores of the second participant sample were analyzed by CFA using the competing model paradigm generated by the initial (EFA) factor structures of the first participant sample, the results were congruent and robust.

A third noteworthy comparison between the two participant samples is that the χ^2 and the χ^2/df of the second sample were demonstrably smaller than those for the first sample, indicating that the second sample's responses better fit the model than did the first sample's. This is strange indeed that a smaller, independent sample that was not used at all to define factor structures conferred a better χ^2 fit than the original model-generating sample. This paradoxical improvement in model fit clearly demonstrates how statistical power (here, sample size 971 vs. 400) affects χ^2 statistics. However, the $\Delta\chi^2$ statistic is not so affected when evaluating model improvement. This is because χ^2 is determined in part by sample size, but *df* is determined by the model, independent of sample size. Even so, $\Delta\chi^2$ of 108 (2 *df*) from the two-factor model to the three-factor model for the second sample was statistically significant ($\chi^{2,}\ p < .001$). Thus, for these two samples of nonclinical university undergraduates, it is concluded that the first-order three-factor structure of DES scores, com-

FIGURE 2
Bifactor 2 and Bifactor 3 Competing Models of the DES



NOTE: DES = Dissociative Experiences Scale. Error terms (e1-e28) for Items d1 to d28 are hidden from view to facilitate model clarity.

TABLE 4
Descriptive Statistics and Internal Reliability for the DES One- and Three-Factor Models

Model	*Number of Items*	*Item* M *(0-100)*	*Item* SD	*Reliability Cronbach's Alpha*	M *Interitem Correlation*
One factor					
Dissociation	28	15.55	11.78	.94	.37
Three factor					
Absorption	16	21.56	14.52	.91	.40
Amnesia	6	8.00	10.30	.81	.44
Depersonalization	6	7.07	11.50	.86	.51

NOTE: DES = Dissociative Experiences Scale. *N* = 971 university undergraduates.

pared to the four-, two-, and one-factor structures (as well as second-order bifactor structures), better represents the construct of dissociation measured by the DES.

Having unequivocally concluded that the first-order three-factor structure of the DES is optimum for these two participant samples, it is now essential to address the sample characteristics to determine what generalizations and applications of the three-factor structure can be applied to other individuals and populations. First, both samples were university undergraduates, primarily freshmen and sophomores, defined as nonclinical individuals. However, in this context, nonclinical does not preclude some individuals with diagnosable clinical and personality disorders and other individuals with subclinical psychopathology. Thus, the range of DES scores expressed by these nonclinical samples is clearly greater than the range for a participant sample devoid of all mental pathology. Individuals with certain mental disorders and symptoms clearly score higher on the DES than true normal individuals (e.g., Gleaves, Eberenz, Warner, & Fine, 1995; Putnam et al., 1996). Moreover, apart from intelligence and severe mental disorders, the characteristics of our samples can be assumed to be similar to the characteristics of the general population of the same age group and the same general geographical area.

Second, the typical age group of our samples, late adolescent, also very likely contributed to the factor analytic results of DES scores. It is well established that after peaking in adolescence, DES scores decrease with increasing age (Ross, Ryan, Anderson, Ross, & Hardy, 1989; Torem, Hermanowski, & Curdue, 1992). The full-scale DES mean score of 15.55 (*SD* = 11.78) for our first sample is very

TABLE 5
DES Confirmatory Factor Analysis Fit Statistics Comparing One, Two, Three, and Four First-Order Factor Models and Two Bifactor Competing Models

Model	χ^2	df	χ^2/df	*RMSEA*	*GFI*	*PGFI*	*PCFI*	*TLI*
Independence	6,052	378	16.01	.194	.19	.17	.00	.00
One factor	1,406	350	4.02	.087	.78	.67	.75	.80
Two factor	1,201	349	3.44	.078	.82	.70	.79	.84
Three factor	1,093	347	3.15	.073	.84	.71	.80	.86
Four factor	1,161	344	3.38	.077	.83	.70	.78	.84
Bifactor 2	1,376	340	4.05	.087	.81	.67	.74	.80
Bifactor 3	1,457	325	4.48	.093	.79	.63	.69	.77

NOTE: DES = Dissociative Experiences Scale; RMSEA = root mean square error of approximation; GFI = goodness-of-fit index; PGFI = parsimony goodness-of-fit index; PCFI = parsimony comparative fit index; TLI = Tucker-Lewis Index. *N* = 400 university undergraduates.

similar to DES mean scores of similar age groups found by other researchers (Murphy, 1994; Ross et al., 1989). By contrast, DES mean scores for nonclinical adult samples are approximately 7.00 (Carlson & Putnam, 1993; Putnam et al., 1996). But what do all these means mean for theoretical and practical applications? The DES has a possible score range from 0 to 100; therefore, sample means close to 0 are characteristic of a floor effect that probably generates decreased variance (smaller *SD*s) and increased positive skewness compared to samples with larger means. Both limited variance and skewness in any direction negatively affect factor analytic procedures (Muthén, 1989; Stevens, 1996). Thus, adolescent/young adult samples with DES mean scores of approximately 15.00 should generate more robust factor-analytic statistics than those generated using adult samples with mean DES scores of approximately 7.00.

GENERAL DISCUSSION

The purpose of our research was to define the optimum factor structure of the DES in a nonclinical university sample. The results of both our EFA and the results of past research using EFA to define the factor structure of the DES have resulted in the same sticky wicket: no definitive justification for either a one-factor or any of the multiple-factor solutions. Thus, we hypothesized that CFA would generate statistics capable of differentiating competing DES factor structures initially identified in EFA. For our sample, CFA successfully identified a first-order three-factor structure of the DES as optimal in three separate domains of model evaluation: χ^2 statistics, fit indices, and RMSEA. Moreover, the three-factor structure was further supported by (a) CFA of a separate, independent participant sample; (b) internal consistency statistics; and (c) logical, substantive factors for the three-factor structure based on the experiences described by each item in the factor subsets of the DES items. Finally, hierarchical structures represented in two separate second-order bifactor models failed to improve model fit of the DES using CFA.

Three caveats initially suggest caution in the interpretation of our results. First, the amount of total variance accounted for by the three factors is 45.8%; therefore, substantial variance in the DES remains unexplained by these common factors. Second, demographic characteristics of our sample limit the ecological validity of the results to nonclinical, university-aged individuals. However, it is believed that this is not a shortcoming of the model per se, rather, only of the relative factor loading values and item means generated by the sample. Third, the participants may have given biased responses to the DES questions, such as would occur in acquiescence or social desirability. Evidence against such responding biases is presented in a study by Beere, Pica, and Maurer (1996) who reported absolutely no relationship ($r = -.006$) between social desirability, as measured by the Marlowe-Crowne Social Desirability Scale, and DES scores in the responses of university undergraduates. Moreover, any tendency for a participant to acquiesce was squelched through an entreaty to respond as accurately as possible and assured anonymity. Thus, we purport that certain applications of our model are valid and useful across populations.

But of what substantial value and utility is the DES as measure of a multifaceted dissociative construct in any application for any population? Here, it is best to look to DES comparative research on dissociative pathology for a definitive answer. Two studies by Ross and his associates—one in the general population (Ross et al., 1991) and the other on a combined sample from five locations of individuals diagnosed with DID (Ross et al., 1995)—reported "virtually identical" (Ross et al., 1995, p. 229) three-factor solutions of the DES using principal components EFA. (Although not "virtually identical," interestingly, the factor structure reported by Ross et al. [1995] is remarkably similar to our three-factor structure.) Of major consequence in the Ross and associates' studies is that both the clinical and nonclinical samples scored similarly on the

Absorption factor, yet the clinical sample scored markedly higher than the nonclinical sample on the Amnesia and Depersonalization factors. These results parallel our results in that our nonclinical sample also had a large Absorption factor score and, by comparison, smaller Amnesia and Depersonalization scores (see Table 4 for factor scores of our study). Thus, elevated Amnesia and Depersonalization factor scores may be indicative of dissociative pathology, whereas Absorption factor scores, in general, may be regarded as representative of more benign dissociative behavior.

Waller, Putnam, and Carlson (1996) provided further evidence on the ability of the DES to differentiate pathological from nonpathological dissociation. Here, an argument is presented for a pathological dissociative taxon identified by elevated scores on DES Items 3, 5, 7, 8, 12, 13, 22, and 27. Tying all these results together, a strong case can be presented for the ability of the DES to differentiate pathological from nonpathological dissociation by examining item subset or factor scores. Precisely, DES Items 3, 5, 7, 8, 12, 13, and 27 (seven of eight in the taxon) are included in Ross et al. (1991, 1995) and in our Amnesia and Depersonalization factors. Item 22—acted differently in different situations, like two different people—is identified in the Absorption factor by both Ross et al. and us; thus, there is not uniform agreement about this item's discriminatory power for dissociative pathology. Overall, however, evidence from three sources converges to support the discriminant power of the DES to identify dissociative pathology by using subset(s) of DES items.

An analysis of differential DES item scores to produce DES total scores of 30, a suggested cutoff point for dissociative pathology (Carlson & Putnam, 1993), illustrates how the factor structure might work to discriminate an individual with dissociative pathology from a putatively extremely high dissociative-normal individual. If the factor scores were 30 across all three factors, or elevated for Amnesia and Depersonalization, then dissociative pathology would be highly suspect. However, if a score of, say, 40 was achieved on the Absorption factor, with scores of 15 or so achieved on the other two factors, yielding a DES total score of 30, then perhaps elevated normal dissociation would be indicated. (Granted, any combination of factor scores that yields a DES total score of 30 is indeed suspect for dissociative pathology; this example is used only to illustrate how differing factor scores might account for different dissociative profiles, with pathology indicated by elevated Amnesia and Depersonalization scores.)

Future research across diverse populations is needed to evaluate the usefulness of the three-factor model as a screening instrument for both dissociative pathology and variations in normal dissociative behavior. For example, one particular hypothesis that could be tested is whether the Amnesia and Depersonalization scales have utility for differential diagnosis within dissociative pathology. Perhaps dissociative experiences resulting from single-experience trauma (e.g., natural disasters or single-incident sexual abuse) are reflected more by depersonalization, whereas those resulting from chronic traumatization are reflected more by amestic experiences. In these situations, that is, where differential etiologies or severity of traumatic precursors affects ensuing dissociative behavior, the three-factor DES model identified by our research maintains the greatest potential for a discerning analysis.

APPENDIX A

Means, Standard Deviations, and Corrected Item-Total Correlations for the 28 DES Items and Factor Loadings of the Three-Factor Solution[a]

Item	M	SD	*Item-Total Correlation*	*Factor*	*Factor Loading*[b]
1	17.87	20.13	.47	Absorption	.43
2	34.79	22.31	.55	Absorption	.65
3	6.32	12.98	.57	Amnesia	.56
4	4.09	10.82	.54	Amnesia	.74
5	9.04	15.58	.56	Amnesia	.70
6	12.60	16.50	.50	Amnesia	.36
7	6.84	14.73	.61	Depersonalization	.44
8	7.14	13.43	.59	Amnesia	.57
9	8.80	15.71	.50	Amnesia	.42
10	17.59	18.78	.57	Absorption	.51
11	6.58	14.11	.58	Depersonalization	.50
12	8.61	15.56	.60	Depersonalization	.67
13	4.98	12.62	.59	Depersonalization	.64
14	26.41	25.12	.58	Absorption	.64
15	23.23	21.63	.65	Absorption	.73
16	15.12	18.80	.72	Absorption	.60
17	27.63	24.82	.58	Absorption	.70
18	20.68	23.87	.68	Absorption	.75
19	22.38	24.54	.44	Absorption	.36
20	22.72	23.12	.62	Absorption	.65
21	25.91	27.24	.51	Absorption	.55
22	16.69	21.52	.64	Absorption	.42
23	26.32	23.22	.52	Absorption	.52
24	22.63	21.30	.69	Absorption	.64
25	14.18	19.00	.71	Absorption	.60
26	10.89	17.13	.67	Absorption	.38
27	7.15	16.29	.57	Depersonalization	.35
28	8.25	16.44	.65	Depersonalization	.61

NOTE: DES = Dissociative Experiences Scale. *N* = 971 university undergraduates.

a. Principal axis exploratory factor analysis, number of factors preset to three.

b. Oblimin rotation, factor loading values from pattern matrix.

APPENDIX B
Correlation Matrix for the 28 DES Items

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*
1	—																											
2	.45	—																										
3	.35	.30	—																									
4	.26	.22	.59	—																								
5	.22	.31	.53	.58	—																							
6	.16	.29	.38	.36	.41	—																						
7	.27	.25	.46	.52	.47	.42	—																					
8	.22	.25	.45	.53	.51	.41	.54	—																				
9	.28	.31	.38	.40	.42	.22	.39	.46	—																			
10	.25	.42	.31	.28	.34	.37	.32	.40	.36	—																		
11	.22	.30	.42	.48	.41	.28	.50	.42	.38	.35	—																	
12	.29	.29	.43	.46	.37	.28	.52	.37	.34	.29	.59	—																
13	.28	.25	.48	.56	.44	.29	.59	.40	.38	.29	.55	.61	—															
14	.25	.36	.29	.22	.27	.32	.31	.35	.19	.41	.32	.31	.27	—														
15	.33	.46	.36	.28	.31	.30	.35	.33	.29	.41	.34	.37	.32	.50	—													
16	.35	.42	.40	.35	.40	.43	.45	.46	.39	.45	.43	.44	.41	.48	.59	—												
17	.34	.42	.27	.25	.31	.30	.27	.31	.26	.39	.30	.26	.28	.40	.44	.47	—											
18	.36	.44	.35	.24	.31	.34	.39	.33	.28	.42	.35	.41	.37	.51	.57	.53	.58	—										
19	.23	.15	.25	.18	.20	.22	.32	.28	.25	.22	.25	.26	.27	.33	.28	.35	.27	.36	—									
20	.32	.48	.30	.25	.31	.34	.32	.34	.26	.41	.34	.37	.30	.42	.43	.43	.47	.55	.34	—								
21	.28	.32	.21	.20	.25	.23	.29	.25	.23	.30	.24	.28	.27	.36	.37	.37	.38	.39	.24	.40	—							
22	.32	.35	.37	.37	.35	.31	.42	.40	.39	.36	.39	.41	.42	.38	.39	.46	.37	.41	.32	.40	.43	—						
23	.24	.29	.27	.23	.25	.25	.29	.28	.24	.33	.25	.30	.21	.38	.35	.43	.35	.41	.39	.30	.30	.44	—					
24	.35	.45	.33	.37	.38	.34	.39	.40	.37	.40	.40	.42	.39	.41	.59	.54	.44	.48	.25	.40	.43	.45	.43	—				
25	.39	.42	.42	.41	.43	.38	.44	.44	.37	.44	.41	.42	.41	.38	.57	.55	.44	.49	.32	.48	.40	.45	.37	.72	—			
26	.30	.31	.39	.42	.46	.41	.48	.46	.42	.40	.38	.43	.47	.36	.42	.50	.37	.42	.29	.43	.36	.50	.37	.54	.60	—		
27	.30	.26	.40	.40	.37	.31	.40	.37	.30	.34	.42	.42	.39	.33	.33	.40	.30	.42	.35	.36	.27	.36	.28	.40	.40	.45	—	
28	.30	.27	.36	.41	.40	.34	.52	.42	.34	.39	.50	.57	.55	.35	.35	.51	.34	.47	.28	.43	.31	.53	.34	.41	.44	.52	.53	—

NOTE: DES = Dissociative Experiences Scale. *N* = 971 university undergraduates.

NOTES

1. The 971-participant sample was culled from 1,060 total participants via listwise deletion and elimination of patently faked or careless response sets. Faked or careless response sets were easy to identify because the Dissociative Experiences Scale was part of a larger battery of inventories in which several of these inventories had a different number of response options. Thus, for example, a participant who continued to mark a six-choice format (for Inventory A) when the response set changed to a four-choice format (for Inventory B) was deemed to have been faking or carelessly marking responses.

2. Principal components exploratory factor analysis with varimax rotation gave almost identical factor structures and item loadings.

3. Percentages of total variance presented for comparison purposes only. Sums of squared loadings cannot be added for total variance when factors are correlated.

REFERENCES

Allen, J. G., Coyne, L., & Console, D. A. (1997). Dissociative detachment relates to psychotic symptoms and personality decompensation. *Comprehensive Psychiatry*, *38*, 327-334.

Amdur, R. L., & Liberzon, I. (1996). Dimensionality of dissociation in subjects with PTSD. *Dissociation*, *9*, 118-124.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arbuckle, J. L. (1999). AMOS (Version 4) [Computer software]. Chicago: SmallWaters Corporation.

Beere, D. B., Pica, M., & Maurer, L. (1996). Social desirability and the Dissociative Experiences Scale. *Dissociation*, *9*, 130-133.

Bernstein-Carlson, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *Journal of Nervous and Mental Disease*, *174*, 727-735.

Braun, B. G. (1988a). The BASK model of dissociation. *Dissociation*, *1*(1), 4-23.

Braun, B. G. (1988b). The BASK model of dissociation: Part II—Treatment. *Dissociation*, *1*(2), 4-23.

Cardena, E., & Spiegel, D. (1993). Dissociative reactions to the San Francisco Bay Area earthquake of 1989. *American Journal of Psychiatry*, *150*, 474-478.

Carlson, E. B., & Putnam, F. W. (1993). An update on the Dissociative Experiences Scale. *Dissociation*, *6*, 16-27.

Carlson, E. B., Putnam, F. W., Ross, C. A., Anderson, G., Clark, P., Torem, M., et al. (1991). Factor analysis of the Dissociative Experiences Scale: A multicenter study. In B. G. Braun & E. B. Carlson (Eds.), *Proceedings of the Eighth International Conference on Multiple Personality and Dissociative States* (p. 16). Chicago: Rush Presbyterian.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum.

Crowley, S. L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality research. *Journal of Personality Assessment*, *68*, 508-531.

Dunn, G. E., Ryan, J. J., & Paolo, A. M. (1994). A principal components analysis of the Dissociative Experiences Scale in a substance abuse population. *Journal of Clinical Psychology*, *50*, 936-940.

Fisher, D. G., & Elnitsky, S. (1990). A factor analytic study of two scales measuring dissociation. *American Journal of Clinical Hypnosis*, *32*, 201-207.

Gleaves, D. H., Eberenz, K. P., Warner, M. S., & Fine, C. G. (1995). Measuring clinical and non-clinical dissociation: A comparison of the DES and QED. *Dissociation*, *8*, 24-31.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*, 532-560.

Holtgraves, T., & Stockdale, G. (1997). The assessment of dissociative experiences in a nonclinical population: Reliability, validity, and factor structure of the Dissociative Experiences Scale. *Personality and Individual Differences*, *22*, 699-706.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87-102). New York: Plenum.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flannagan, J. L. Genschaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 373-402). New York: Guilford.

Kihlstrom, J. F., Glisky, M. L., & Angiulo, M. J. (1994). Dissociative tendencies and dissociative disorders. *Journal of Abnormal Psychology*, *103*, 117-124.

Kluft, R. P. (1993). The editor's reflective pleasures. *Dissociation*, *6*, 1-2.

Marmar, C. R., Weiss, D. S., Schlenger, W. E., Fairbank, J. A., Jordan, B. K., Kulka, R. A., et al. (1994). Peritraumatic dissociation and posttraumatic stress in male Vietnam theater veterans. *American Journal of Psychiatry*, *151*, 902-907.

Murphy, P. E. (1994). Dissociative experiences and dissociative disorders in a non-clinical university student group. *Dissociation*, *7*, 28-34.

Muthén, B. O. (1989). Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology*, *42*, 241-250.

Pitblado, C. B., & Sanders, B. (1991). Reliability and short-term stability of scores on the Dissociative Experiences Scale. In B. G. Braun & E. B. Carlson (Eds.), *Proceedings of the Eighth International Conference on Multiple Personality and Dissociative States* (p. 179). Chicago: Rush Presbyterian.

Putnam, F. W., Carlson, E. B., Ross, C. A., Anderson, G., Clark, P., Moshe, T., et al. (1996). Patterns of dissociation in clinical and nonclinical samples. *Journal of Nervous and Mental Disease*, *184*, 673-679.

Ray, W. J., & Faith, M. (1995). Dissociative experiences in a college age population: Follow-up with 1190 subjects. *Personality and Individual Differences*, *18*, 223-230.

Ray, W. J., June, K., Turaj, K., & Lundy, R. (1992). Dissociative experiences in a college age population: A factor analytic study of two dissociation scales. *Personality and Individual Differences*, *13*, 417-424.

Riley, K. C. (1988). Measurement of dissociation. *Journal of Nervous and Mental Disease*, *176*, 449-450.

Ross, C. A., Ellason, J. W., & Anderson, G. (1995). A factor analysis of the Dissociative Experiences Scale (DES) in dissociative identity disorder. *Dissociation*, *8*, 229-235.

Ross, C. A., Joshi, S., & Currie, R. (1991). Dissociative experiences in the general population: A factor analysis. *Hospital and Community Psychiatry*, *42*, 297-301.

Ross, C. A., Ryan, L., Anderson, G., Ross, D., & Hardy, L. (1989). Dissociative experiences in adolescents and college students. *Dissociation*, *2*, 239-242.

Sanders, B., & Giolas, M. H. (1991). Dissociation and childhood trauma in psychologically disturbed adolescents. *American Journal of Psychiatry*, *148*, 50-54.

Sanders, B., & Green, J. A. (1994). The factor structure of the Dissociative Experiences Scale in college students. *Dissociation*, *7*, 23-27.

Sanders, S. (1986). The Perceptual Alteration Scale: A scale measuring dissociation. *American Journal of Clinical Hypnosis*, *29*, 95-102.

SPSS, Inc. (1998). SPSS graduate pack 9.0 for Windows [Computer software]. Chicago: Author.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Tellegan, A., & Atkinson, G. (1974). Openness to absorbing and self-absorbing experiences ("absorption"): A trait related to hypnotic susceptibility. *Journal of Abnormal Psychology*, *83*, 268-277.

Torem, M. S., Hermanowski, R. W., & Curdue, K. J. (1992). Dissociation phenomena and age. *Stress Medicine*, *8*, 23-25.

Van Ijzendoorn, M. H., & Schuengel, C. (1996). The measurement of dissociation in normal and clinical populations: Meta-analytic validation of the Dissociative Experiences Scale (DES). *Clinical Psychology Review*, *16*, 365-382.

Waller, N. G. (1995). The Dissociative Experiences Scale. In J. C. Conoley & J. C. Impara (Eds.), *Twelfth mental measurements yearbook* (pp. 317-318). Lincoln, NE: Buros Institute of Mental Measurement.

Waller, N. G., Putnam, F. W., & Carlson, E. B. (1996). Types of dissociation and dissociative types: A taxometric analysis of dissociative experiences. *Psychological Methods*, *1*, 300-321.

Whitley, B. E., Jr. (1996). *Principles of research in behavioral science*. Mountain View, CA: Mayfield.

Yung, Y., Thissen, D., & McLeod, L. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113-128.

Gary Stockdale, M.A., is presently a doctoral student in quantitative psychology at the University of California–Davis. His research interests include studying individual differences through structural equation modeling and related procedures.

Betty E. Gridley, Ph.D. is professor of psychology–educational psychology at Ball State University in Muncie, Indiana. She directs the M.A./Ed.S. programs in school psychology and teaches school psychology and measurement and statistics courses. Her major research interests are in applied multivariate statistics in assessment and with special populations.

Deborah Ware Balogh, Ph.D., is dean of the Graduate School and professor of psychological science at Ball State University. She is a fellow of the Society for Personality Assessment. Her primary research interests include adult psychopathology, assessment of personality disorders, and vulnerability to schizophrenia.

Thomas Holtgraves, Ph.D., is a professor of psychological science at Ball State University. His primary research examines the role of social psychological variables in the production and comprehension of language.

AUTHOR GUIDELINES

The editor invites articles that target empirical classification of normal and abnormal behaviors and personality characteristics, theory and measurement of diagnostic constructs, identification and clarification of mental disorders, and reliability and validity of clinical diagnoses and judgments. Research participants may represent diverse age and socioeconomic categories and both clinical and nonclinical populations. Research reviews and methodological papers will be considered, and article format may be varied to accommodate brief reports.

Manuscript Submission: Authors should use the *Publication Manual of the American Psychological Association* (5th edition, 2001) as a guide for preparing manuscripts for submission. All manuscript pages, including reference lists and tables, must be typed double-spaced. Supply four (4) copies of each manuscript to Robert P. Archer, Editor, *Assessment*, Department of Psychiatry and Behavioral Sciences, Eastern Virginia Medical School, 825 Fairfax Avenue (Hofheimer Hall 730), Norfolk, VA 23507. A masked blind review process may be requested. Authors requesting this option should submit manuscript copies that do not contain personal identification and provide a separate title page with authors' names and addresses.

Brief reports may be submitted for publication and should be limited to 1,000 words, including an abstract of 75 words or less. The brief report format should be used for carefully designed and executed investigations of research topics of specialized interest that make a significant contribution to the literature but do not require a full-length manuscript. In addition, comments may be submitted as brief manuscripts responding to articles published in the journal. Comments should be no longer than 1,000 words and will be selected for publication based on an editorial review process. Comments should be submitted no later than 6 months following the date of the issue containing the article on which the commment is based. Comments based on topics other than articles published in the journal will not be accepted.

If a manuscript is accepted for publication, authors will be asked to provide an electronic file copy of the manuscript and camera-ready figures. Authors submitting manuscripts to the journal should not simultaneously submit them to another journal, nor should manuscripts have been published elsewhere in substantially similar form or with substantially similar content.

Preparation of Manuscripts: Authors should carefully prepare their manuscripts in accordance with the following instructions.

Manuscripts should be as concise as possible, yet sufficiently detailed to permit adequate communication and critical review. Authors should follow the style of the *Publication Manual of the American Psychological Association* (5th edition, 2001). When possible, authors should prepare their manuscripts (including tables) on word-processing equipment that is capable of outputting files on disks or transmitting files by the Internet because mansucripts are copyedited and typeset from disks or files provided by authors. Authors will receive information for submitting the final copy of their manuscript by electronic means on final acceptance of their paper.

The first page of the paper should contain the article title, the names and affiliations of all authors, authors' notes or acknowledgments, and the names and complete mailing addresses of all authors. Please note the author to whom all correspondence, including proofs, should be sent. The second page should contain an abstract of no more than 150 words and five to seven keywords that will be published following the abstract. A separate page containing brief biographies for all authors should also be included.

The following sections should be prepared as indicated:

Tables. Each table should be fully titled, typed single-spaced on a separate page, and placed at the end of the paper. Tables should be numbered consecutively with Arabic numerals. Footnotes to tables should be identified with superscript lowercase letters and placed at the bottom of the table. All tables should be referred to in the text.

Figures. Copies of figures should be sent on first submission of a manuscript; original camera-ready and electronic figures will be requested when a manucript is accepted for publication. Electronic copies of figures can be submitted in one of the following formats: Microsoft PowerPoint or Word, Tagged Image File Format (.TIF), Encapsulated Postscript File (.EPS), Joint Photographic Experts Group (.JPG), or Portable Network Graphic Format (.PNG). All figures should be referred to in text. Each figure should appear on a separate page at the end of the paper, and all titles should appear on a single, separate page.

Endnotes. Notes should appear on a separate page before the References section. Notes should be numbered consecutively and each endnote should be referred to in text with a corresponding superscript number.

References. Text citations and references should follow the style of the *Publication Manual of the American Psychological Association* (5th edition, 2001).