Soc Psychiatry Psychiatr Epidemiol (1999) 34: 399–409

ORIGINAL PAPER

G.E. Switzer · S.R. Wisniewski
S.H. Belle · M.A. Dew · R. Schultz

Selecting, developing, and evaluating research instruments*

Accepted: 20 April 1999

Abstract The goal of this paper is to provide researchers who are not experts in psychometric theory with a concise guide to instrument selection, development and evaluation. Issues of context – factors related to the setting or population in which an instrument will be used – and psychometrics – the functioning of an instrument within a given context – are reviewed and discussed. Finally, four categories or types of instruments, and the psychometric analyses that are necessary for establishing the reliability and validity of each type, are described.

Introduction

Although many excellent texts have been written on general principles of psychometric evaluation of instruments (e.g., Fleiss 1981; McDowell and Newell 1996; Nunnally and Bernstein 1994; Rosenthal and Rosnow 1991), few of these provide concise statements concerning the practical application of instrument selection and evaluation to the everyday issues encountered by researchers in all areas of social and behavioral science. Thus, the primary goals of this paper are to provide a "pocket-sized" guide that includes (a) an organizational structure for making decisions concerning instrument selection, development, and evaluation, and (b) a practically oriented discussion of the basic issues involved in such decisions. Our discussion is designed to provide researchers who are not experts in instrumentation with an overview of measurement issues, and to direct readers to more detailed texts on the topics covered here.

We focus initially on two broad types of considerations central to decisions about instrumentation: context and psychometrics (see Table 1). Context refers to factors exogenous to the assessment tool itself, such as characteristics of individuals to be assessed, the goals of the research endeavor, and constraints on data gathering capabilities. Psychometrics refers to the properties of the instrument as it functions within the context. Finally, we will discuss four classes or types of instruments and illustrate the application of principles discussed in the earlier sections to each class of instrument.

Although the examples included here focus primarily on mental and physical health issues generated in our research with elderly caregivers (Resources for Enhancing Alzheimer's Caregiver Health; REACH 1995), the concepts we cover may be more generally applicable to most self-report or expert-completed measures. They are not necessarily intended for performance-based assessments such as neuropsychology evaluations, intelligence tests, or academic achievement examinations.

Contextual issues in the selection and/or development of instruments

Participant characteristics

In selecting or developing an instrument, one of the primary considerations should be the characteristics of

G.E. Switzer (✉)
Departments of Medicine and Psychiatry,
University of Pittsburgh,
3811 O'Hara Street,
Pittsburgh, PA 15213, USA
Tel.: +1-412-624 2520, Fax: +1-412-383 1755

S.R. Wisniewski · S.H. Belle
Department of Epidemiology,
University of Pittsburgh,
Pittsburgh, Pennsylvania, USA

M.A. Dew
Departments of Psychiatry,
Epidemiology, and Psychology,
University of Pittsburgh,
Pittsburgh, Pennsylvania, USA

R. Schultz
Departments of Psychiatry, Psychology, and Sociology,
University of Pittsburgh,
Pittsburgh, Pennsylvania, USA

* This work was supported by grant AG13305 from the National Institute on Aging and the National Institute of Nursing Research, Bethesda, MD.

Table 1 Key contextual measurement issues

Contextual issues
Population characteristics
Age
Gender
Education level
Health status
Recent life experiences
Cultural context
Ethnicity
Cultural traditions and norms
Historical context
Language
Knowledge base
Beliefs, attitudes, values
Political and historical events
Research goals
Content of measurement
Specificity of measurement
Comparisons to normative groups
Administration issues
Feasibility
Format of instrument

the study participants (see Table 1). Recent studies have indicated that factors such as the respondent's age, gender, education level, physical and mental health status, and other recent life experiences (e.g., recent pregnancy and delivery, recent bereavement, traumatic life experience) affect responses to items. These factors may lead to under-endorsement or over-endorsement of items, biases in recalling events, and/or respondent difficulty in interpreting questions. For example, it has been argued that the Beck Depression Inventory produces falsely elevated ratings of depression among the elderly because of over-endorsement of certain items (e.g., body-image change; Talbott 1989). Other research has shown that many depression instruments may *underestimate* depression in the elderly, because older persons tend to deny depressive symptoms, or to attribute them to physical health problems (Maier et al. 1988). The respondent's gender may also affect responses. It has been argued that observed gender differences in depression may be at least partially due to a greater willingness of women to endorse symptoms included in these measures, rather than a true gender difference in depression levels (Miller et al. 1985).

Education level is also important to consider for many types of assessment. Mental status assessment instruments (e.g., the Short Portable Mental Status Questionnaire, the Mini-Mental State Examination) have been shown to over-estimate cognitive impairment among groups with little education and to under-estimate impairment among the highly educated (Berkman 1986; Brayne and Calloway 1990; Kay et al. 1985; Murden et al. 1991; Uhlmann and Larson 1991). Other instruments assessing mental health that include a relatively high proportion of somatic symptom items may be inappropriate for physically ill groups, in whom such symptoms may reflect medical status rather than emotional distress (Dew, in press; Williams and Richardson 1993). Reporting of health problems may also be affected by other health-related behaviors such as visiting the hospital; current chronic illness is increasingly under-reported as the length of time since the last hospital visit increases (Cannell et al. 1977; Madow 1973).

Cultural context

A second important issue to consider is the cultural appropriateness of the instrument for the study population. Most instruments used in social and behavioral research are based on middle-class, Western European/North American assumptions, values, and norms, and thus may not be entirely appropriate for other cultural groups. For example, many of the classic symptoms of schizophrenia as defined by the *Diagnostic and Statistical Manual IV* (DSM; American Psychiatric Association 1994; e.g., delusions, hallucinations, disorganized speech) are part of the religious ceremonies or daily spiritual experiences of many cultural groups (Eaton 1980). Conversely, it appears that some mental disorders – for example, "ataques de nervios" among Puerto Ricans – are recognized only among non-European cultures (Guarnaccia et al. 1990). Culture-bound assumptions may pervade virtually all mental and physical health instruments. Consequently, it is important to determine whether the instrument has been used successfully with the particular cultural/ethnic groups included in the sample.

Historical context

The effects of historical and political events on measurement issues are rarely discussed, but may be as critical as any of the other contextual issues discussed here, especially for classes of measures that have been used for several years. Societies as a whole experience changes in knowledge, beliefs, attitudes, language, and values that, in turn, may affect how individuals interpret items. There are several recent examples of responses to changes in historical context. For example, tests of IQ have a long history of revision and updating to accommodate the fact that the knowledge base of society has shifted over time, and that, on average, individuals are becoming more educated and adept at answering the types of questions that have been used as indicators of IQ. Measures of health behaviors also require continual revision as our knowledge about health indicators improves. Until a few years ago, questions about smoking, alcohol consumption, and diet, which are currently regarded as central to health assessment, were rarely included in health questionnaires. Thus, the language and other implicit assumptions of a given measure should be part of the initial considerations in instrument selection, especially for instruments that are several years old.

Research goals

It may seem obvious to suggest that the goals of a specific research effort should guide instrument selection, but there are multiple considerations in this regard. When assessing global health status, for example, it is critical to determine whether it is most important to measure symptoms (e.g., Were you short of breath?), performance (e.g., Would you have trouble running the length of a football field?), feeling-states (e.g., I feel I am a burden to people), general quality of life (e.g., In general, how satisfying is your life?), or some combination of these. Health measures vary greatly in their relative emphasis on physical, emotional, and social health, and the extent to which information reported across these domains is based on perceptions and feeling-states, or on symptom frequencies.

A second consideration is whether a general (generic) measure or a specific measure should be used. The relative advantages and disadvantages of generic versus specific measures are currently being discussed prominently in the physical health, psychiatric, and quality of life literatures. Disorder-specific health measures (e.g., Arthritis Impact Measurement Scales; Meenan et al. 1992) enhance the ability to discover fine-grained distinctions among individuals suffering from the disorder under consideration, but may not be adequate if comparisons of status across individuals with different disorders is central to the research goals. Finally, decisions about instrumentation may be based on the importance of making comparisons across studies, or with normative samples. If it is desirable to make such normative comparisons, it will be critical to utilize a measure that has been used extensively in other populations, even if it does not address the full range of issues important to the project.

Administration issues

Researchers often have several choices about how to gather information from respondents. An initial consideration should be the feasibility of using a particular instrument with the population of interest. Feasibility issues include the burden to potential respondents, and the financial cost per subject of gathering the information. Respondents may be reluctant to complete a lengthy interview or survey, both because of the time involved and perceptions that they will be asked to give confidential or sensitive types of information. Groups receiving medical or psychiatric treatment, for example, depending on the nature or severity of their illnesses, may have more difficulty in completing certain types of assessments such as self-administered questionnaires. Reluctance to participate may be addressed with careful explanation of the study procedures and how the data will be used, assurances of anonymity, and with monetary, or other types of incentives offered to participants. An initial cost consideration is that of the instrument itself; many established measures are copyrighted and the authors may charge a fee each time the instrument is administered or scored. Another important consideration is the cost of the assessment modality, and of the person who will administer the assessment. Clinician interviewers are most costly, followed by trained lay interviewers and research assistant administered questionnaires/interviews; self-administered questionnaires are the least costly.

In terms of the format of data gathering, in-person interviews are generally the most costly mode of assessment, followed by telephone interviews, and self-administered questionnaires. Although self-report forms may be the least costly to administer, this method is limited by the respondent's ability to read and understand questions, greater potential for non-response bias, and difficulties in presenting complicated question sequences. Telephone interviews may provide a middle ground in terms of cost and quality of information gathered. They also have been shown to yield highly reliable data if the interviewers are carefully trained and supervised (Aneshensel et al. 1982a, b; Aneshensel and Yokopenic 1985; Fenig et al. 1993; Paulsen et al. 1988; Wells et al. 1988). The use of computers to aid in recording responses to both interviews and self-administered questionnaires has also become more prevalent, and seems to provide a reliable, valid, and highly efficient means of assessing some attributes (e.g., Brugha et al. 1996; Dignon 1996; Erdman et al. 1992; Kobak et al. 1993; Steer et al. 1994; Thornicroft 1992; for a review, see Kobak et al. 1996).

Issues in the psychometric evaluation of instruments

In this section, we discuss the general meaning of instrument reliability and validity – the two primary concerns of psychometric evaluation – and methods for examining whether measures meet these minimum psychometric requirements (see Table 2 for a summary of key psychometric issues).

Reliability

The score or value obtained by an individual on a measure traditionally has been viewed as comprising two components: an underlying "true" score, and error caused by imprecision in measurement (McDowell and Newell 1996; Nunnally 1978). Reliability of a measure refers to the measure's ability to detect the true score rather than measurement error. A perfectly reliable instrument would detect only the true score. The concept of reliability is based on two central considerations:

1. Do items purportedly belonging to a scale actually assess a single construct, and
2. Do scales measuring a single construct produce consistent estimates of that construct across multiple measurements.

Table 2 Key psychometric measurement issues

Psychometric issues
Reliability
Internal-consistency
Multiple measurement consistency
Test-retest
Alternate form
Split-half
Inter-rater
Validity
Content
Criterion
Construct
Factor analytic
Group differences
Within-subject variation across time
Correlations with other measures
Internal consistency
Explication of process

The first consideration is usually labeled ''internal-consistency reliability'' and is most commonly assessed with Cronbach's alpha, which provides an estimate of the extent to which items covary, or ''hang-together'' as a common unit (Cronbach 1951). Alpha ranges from 0.00 to 1.00, with higher scores indicating greater internal-consistency of the scale. Alpha is sensitive to the number of items in a scale and typically increases as the number of items increases; the incremental improvements in alpha resulting from adding items to the scale may be relatively large up to about 10 items, and then begin to diminish (Shrout and Yager 1989). Comparisons between individuals, such as those necessary in case-finding, require high reliability (above 0.90). Research focused on group comparisons and research in the early stages does not require as extremely high reliability. It has been suggested that a good standard for the latter two situations is to obtain reliability coefficients of 0.50–0.80 (Helmstadter 1964; Nunnally 1978; Ware 1984). Attempting to achieve reliability coefficients above 0.80 may require considerable time and money, and may lead to redundancy among items in the measure (Boyle 1985; McDowell and Newell 1996; Nunnally 1978). The Kuder-Richardson-20 is similar to Cronbach's alpha, but used for dichotomous scales.

The second reliability consideration – consistency across multiple measurements – has several variations, including test-retest, alternate form, split-half, and intra- and inter-rater reliability, and is based on the assumption that many human attributes are relatively stable in the short term. Thus, reliable instruments should produce consistent estimations of such attributes across multiple measurements administered in relatively close temporal proximity. It should be noted that for intervention or longitudinal research, the optimal measure would produce consistent results in the short term, but also have high sensitivity to changes that may take place longitudinally and/or during an intervention (Kraemer 1992). Test-retest reliability is obtained by reassessing individuals with the same measure at a second time point after the initial measurement. There are some serious limitations in using test-retest methods to estimate reliability, explicated in detail by Nunnally (1978). For example, the selection of the second administration point is based on the competing goals of minimizing the chance that respondents will remember and attempt to duplicate their responses from the first administration (implying that a longer time interval should be used), and minimizing the chance that any true change in the attribute will have occurred between the two administrations (implying that a shorter time interval is needed). Although judgements about when to administer the retest must be based on the specific instrument under consideration, testing experts suggest that an interval of 2–4 weeks from initial administration may be most appropriate (Nunnally 1978).

To overcome the liability of respondents' recalling their previous responses inherent in test-retest reliability, alternate form (a second, similar version of the instrument) and split-half methods of establishing reliability were developed. Conceptually, both methods are based on the idea that high correlations between two different versions of a measure is evidence that a construct is being assessed reliably. In alternate form reliability, the two versions are administered at separate sittings and high correlations between the two versions is taken as evidence of reliability. The split-half method assesses the degree of correlation between two halves of an instrument (often odd versus even items) or between all possible pairs of items administered at a single administration of the instrument. Intra and inter-rater reliability is similar to these methods, but is appropriate for data involving researchers' judgements (e.g., ratings by interviewers, observational assessments), rather than by respondent self-report. Reliability of assessments conducted by a single rater at different timepoints, or two different raters or judges, is typically evaluated with the Intraclass r for continuous variables and Kappa for dichotomous or ordinal-level variables; high correlations or agreement scores are taken as evidence of measurement reliability.

Although to this point, we have described reliability techniques as involving the family of Pearson correlations and related measures of association, limitations in how such associational coefficients can be interpreted has led some researchers to advocate use of alternate methods for evaluating reliability (see Bartko and Carpenter 1976, for an excellent review of reliability assessment; Bland and Altman 1986; McDowell and Newell 1996). The central limitation of the Pearson correlation is that it reports *association* – how accurately one score can be predicted from the other – rather than *agreement* – whether the two scores are identical. Several alternative methods of evaluating reliability based on agreement, including graphically plotting and examining the differences between pairs of scores (Bland and Altman 1986), calculating an intraclass correlation for continuous scales (an ANOVA-based approach also used with judges' ratings as noted above), and

calculating Kendall's index of concordance for ordinal scales (Deyo et al. 1991), have been developed to address the limitations of the Pearson correlation. The more sophisticated approaches – the intraclass correlation and Kendall's index of concordance – indicate the degree of similarity between the two scores rather than the relative position of individuals on the first and second administration. The intraclass correlation has many variations and can be used to evaluate reliability among pairs of scores obtained by a variety of methods: test-retest, alternate form, and different raters (Shrout and Fleiss 1979). The equivalent statistic for nominal-level or dichotomous variables is the Kappa coefficient (Cohen 1960).

Validity

Validity is most often defined as the extent to which an instrument measures what it was intended to measure (Anastasi 1982). However, it is important to note that instruments may fail validity criteria for one purpose but be valid measures of a different construct (e.g., the Health Opinion Survey, developed to assess mental health, may be a better indicator of generalized stress; Butler and Jones 1979) *or* may be valid indicators of constructs in addition to the one for which they were originally intended (e.g., measures of physical functioning that are also useful as quality of life indicators). In addition, instruments that may be valid in one context (i.e., population, culture, historical period, administration format), may not be valid in another context; validity is always context specific.

Because validity is context specific, validating a measure must be viewed as a process of accumulating evidence that supports the meaningfulness of the measure rather than a discrete endpoint at which validity is proven (Stewart and Ware 1992). Three broad types of validity are most often cited as central to any validity argument: content, criterion, and construct. Extended discussions of these types of validity exist elsewhere (e.g., Helmstadter 1964; McDowell and Newell 1996; Nunnally 1978; Stewart and Ware 1992), and we will define and discuss each type briefly here. However, we should note that *reliability* of an instrument is a necessary but not sufficient condition for establishing the instrument's validity (Nunnally 1978). If an instrument is not assessing something consistently, the meaningfulness of the measure is called into question even before validity arguments can be addressed.

Content, or face validity concerns the extent to which items in a measure accurately reflect the full breadth of the construct of interest. Nunnally (1978) suggests that if we imagine a sampling universe of all possible items that might identify a construct, content validity is established by demonstrating that a representative set of items has been selected for our measure. Validity of content is usually established by having experts in the field, and subjects or patients from the population for whom the instrument would be appropriate, review the instrument and provide critical evaluations of content; there are no formal empirical tests that will verify that content validity has been established. Recently, focus groups and in-depth interviews have gained popularity as methods for gathering content validity information for instruments in the early stages of development. Although evidence of content validity may provide the least powerful validity argument, such evidence is a prerequisite for establishing other types of validity.

Criterion or correlational validity is the extent to which the measure correlates with a "gold standard" of the intended construct. The gold standard (or criterion) can be another accepted measure of the same construct, or in rare cases, observed behavior, characteristic, or attribute that the measure is designed to assess (e.g., self-reported physical functioning validated against observer ratings of actual physical capabilities). Criterion validity is typically established by examining the correlation of each item and/or the full scale with the criterion score or behavior. Low correlations – either item-criterion or scale-criterion – suggest that particular items, or the scale as a whole, may not measure the intended construct. (Note that this conclusion rests on the assumption that an appropriate criterion has been selected.) Criterion validity can be further divided into concurrent validity – the intended construct and criterion are assessed simultaneously – and predictive validity – the intended construct is measured first and then used to predict the criterion.

As noted by Helmstadter (1964), construct validity is the most recent addition to ideas about required validity evidence (APA Committee 1952, 1954; Cronbach and Meehl 1955), and requires that an instrument be (a) viewed as measuring an underlying construct, and (b) tested to see whether its hypothesized or theoretical relationships with other variables can be established. Factor analytic techniques are one way of exploring and/or confirming whether a group of items comprises a single unified construct, multiple components of a single construct, or multiple divergent constructs. Factor analysis is useful in determining whether a group of items hypothesized to assess a construct actually do cluster together when they are analyzed with items from other scales, and whether items *within* a measure describe a unified versus a multicomponent construct. Factor analysis should be undertaken in the early stages of examining an instrument to help determine the relationship among items, and to provide evidence of construct validity (Nunnally 1978). Factor analyses may be confirmatory – if a priori hypotheses about which items will load together are specified – or exploratory – if no such hypotheses are made. Structural equation modeling is also increasingly used as a highly sophisticated and flexible means of conducting confirmatory factor analysis (Ullman 1996).

Cronbach and Meehl (1955) outline five additional ways that construct validity can be established. First, *group differences* may be examined; groups of

individuals expected to differ – based on additional characteristics (e.g., ethnicity, gender) – should score differently on the measure. Second, *within-subject variation* measured across time should indicate minimal changes for trait-like variables and more substantial changes for state-like variables. Third, strong *correlations with other measures* of the same construct (convergent validity), and weak correlations with measures of other constructs (discriminant validity) should be observed. The multitrait-multimethod approach (Campbell and Fiske 1959) is one framework that can be used to examine such interrelationships among items and scales purported to assess different psychosocial domains. Fourth, the *internal consistency* of an instrument or subscale provides evidence that a single construct is being assessed. Finally, it is important to conduct a thorough examination and *explication of the assessment process* in which all the steps necessary to answer a certain item are analyzed to eliminate alternate hypotheses about observed patterns of responses (e.g., response set, social desirability).

Issues in selecting, developing, and evaluating four classes of instruments

Instruments used in most research efforts can be divided into four broad categories based on the extent to which the full instrument, or items within the instrument, have been used in other research and have well-established psychometric properties. The following section is organized around these four categories, which we have labeled (a) established measures, (b) modified measures, (c) hybrid measures, and (d) new measures (see Fig. 1). Established measures are those that have been used in more than one research setting and have exhibited good reliability and validity in each of these settings – published measures that do not meet these criteria should be treated as new measures. Modified measures have been modified in some way (e.g., shortened, altered response categories) to fit the research goals. Hybrid measures combine items from more than one source to assess a single construct. New measures are those that are newly developed with a specific research goal in mind.

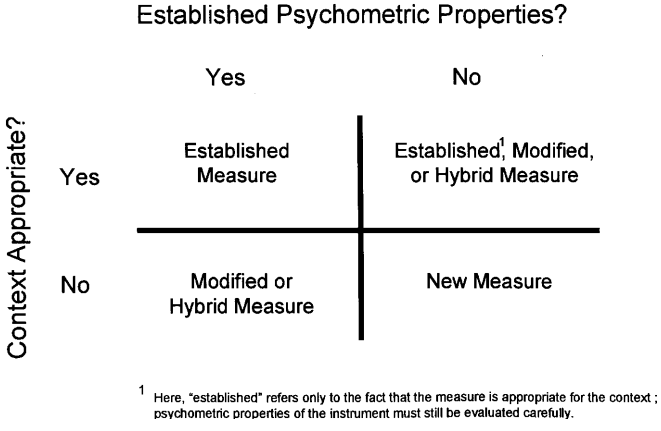

Fig. 1 Basic template for measure selection

Two questions should guide the search for an appropriate study measure.

1. Do appropriate established measures exist? If so, the issues in the Established measures section below should guide instrument selection.
2. Do measures that are nearly appropriate for the study goals exist? If so, a modified or hybrid measure should be considered. If no appropriate or nearly appropriate measures exist, creation of a new measure may be justified.

Established measures

As noted at the outset of this discussion, the two primary considerations in evaluating whether or not an existing measure is suitable for a particular research endeavor are contextual and psychometric. These two sets of issues are linked in the sense that using an inappropriate measure for a given context (e.g., unclear wording, gender biased) will lead to psychometric liabilities such as poor reliability or validity. Thus, the first consideration should be whether an established measure meets the contextual considerations we have outlined. If some characteristics of the population to be studied (e.g., age, culture, historical period) or the administration format are significantly different from those that were used to establish the psychometric properties of the instrument, pilot tests should be conducted to establish the psychometrics of the instrument in the new population.

An excellent example of efforts to establish the psychometric properties of a new instrument – and the potential pitfalls that are inherent in such efforts – is the 20-item Center for Epidemiological Studies Depression Scale (CES-D; Radloff 1977). The CES-D has been widely used to assess general depressive symptomatology or distress as a combination of affective, somatic, and interpersonal symptoms. Initial evaluations of the instrument conducted among English-speaking, middle-class, Anglo individuals of various ages yielded evidence of good reliability and validity (Hertzog et al. 1990). However, subsequent studies of CES-D characteristics among diverse ethnic groups including American Indians (Manson et al. 1990) and Hispanics (Guarnaccia et al. 1989), *and* comparisons of men and women (Guarnaccia et al. 1989; Stommel et al. 1993), suggest that the factor structure and/or the operation of individual items differed across ethnic groups and by gender. Such differences in the characteristics of a measure when it is applied to new populations have serious implications for construct validity, and have been addressed by researchers in a variety of ways. For example, it may be possible to identify a different measure of the construct – in this case a different measure of distress – that operates similarly across the population groups of interest. Al-

ternatively, items that are biased may be eliminated or altered – Stommel et al. (1993) used a 15-item version of the CES-D to reduce gender bias – and the newly discovered factor structure may be used in the analyses – alternative CES-D factors have been used in analyses involving minority populations (e.g., Guarnaccia et al. 1989).

Evidence of criterion-related validity, if available, should also be evaluated carefully. For example, Roberts et al. (1990) found that although CES-D scores were highly associated with diagnosed depression – as assessed by the Diagnostic Interview Schedule – in Anglo and English-speaking Mexican-American populations, there was poor agreement between the CES-D and diagnosed depression in Spanish-speaking Mexican-Americans. Although the appropriateness of the criterion (in this case the DSM) should always be considered seriously, intergroup differences between the measure and the criterion should raise validity concerns.

After an established measure is selected and used in a research effort, evaluating its psychometric properties in the current research effort is still critical, but may entail a less rigorous process than for the other types of measures. At minimum, initial psychometric analyses should include an evaluation of the internal consistency of the measure, as well as analyses designed to verify the factor structure (e.g., confirmatory factor analysis). In the case of the CES-D, for example, the three-factor structure should be verified, and overall and subscale internal consistency values should be computed and reported. A conservative approach would also suggest that some evidence of construct validity is necessary. Findings such as the evidence of the unstable factor structure of the CES-D across some research settings suggest that assuming validity may not always be warranted. If novel intergroup comparisons are part of the research goals (e.g., by gender or ethnic group), it is important to conduct the psychometric analyses described here *within* each group of interest. Divergent factor structures or internal consistency coefficients imply that the measure is not equivalent across groups and that differences among groups should be interpreted with caution.

Modified measures

Perhaps the first issue in modifying an established measure should be to explicate a detailed rationale for the alterations to be made. Shortening a measure substantially, changing the response categories, or altering the item stems may have serious psychometric consequences for the scale. Comparisons with studies employing the original version of the scale may not be valid. In other words, depending on the extent of the modifications, a modified measure may be only moderately superior to a newly created measure in terms of the ability to rely on previously reported psychometric work. The primary advantages of modifying a measure over developing a new measure are that there are some assurances that this set of items has operated as an indicator of a unified construct in the past, and that clarity of item wording and content has been demonstrated.

Justifications for modifying a measure are not limited to, but may include

1. Original measure is too long for the current research purpose
2. Original response categories are not expected to produce sufficient variation
3. Original response categories are too broad or inclusive, and
4. Original item wording is unclear or not relevant to the current population.

In presenting findings based on a modified measure, it is important to describe the original measure, outline the steps that were taken to alter the measure, and discuss any anticipated differences in the performance of the modified measure.

Many published examples are available of measures that have been modified in some way after the original psychometric work was conducted. One of the most common modifications is to reduce the number of items in a measure to reduce respondent burden. The 20-item Short Form Health Survey, assessing six health-related domains, is an example of the process of selecting and evaluating items from longer health surveys (Stewart et al. 1988; Ware et al. 1992). When the 20-item version was criticized for being too limited in scope, a 36-item, eight-domain version of the Short Form was developed and evaluated (Ware et al. 1993; Ware and Sherbourne 1992). Finally, in response to calls for an abbreviated instrument, a 12-item, two-domain version of the instrument was developed (Ware et al. 1996). Each step in the refinement of the instrument was fully reported, and psychometric evaluations were described (Jenkinson et al. 1997; Ware et al. 1996).

Substantial psychometric work is needed to assess the reliability and validity of modified measures. Depending on the extent of the modifications, the full range of reliability tests may need to be conducted. In addition, at least some validity work is necessary. Content validity and evaluation of internal factor structure may be especially important for measures that are reduced in length from their original versions, to ensure that the same number, and full breadth of, the original constructs are represented. Conversely, evaluating construct validity to determine whether the measure seems to support anticipated theoretical relationships may be more of a concern when there are additions to, or modifications in, item wording (e.g., RMBPC).

Hybrid measures

Hybrid measures – created by combining items from more than one established scale, or by combining items from an established scale with newly created items – are one step further removed from their original psycho-

metric properties than are measures that have been modified. When existing scales do not adequately cover all the issues of interest, or have questionable psychometric properties, creating a composite measure from more than one scale or developing new items to supplement a scale may be justified.

As with the modified measures, the rationale for creating a hybrid measure should be developed with the foreknowledge that previous psychometric work with these items may no longer be valid. In presenting work involving a hybrid measure, it is important to provide the following information:

1. Description of the original measure(s)
2. Inadequacies in existing measures that led to the creation of a hybrid measure
3. Steps in selecting or creating items
4. Modifications that were made to item stems or response categories, and
5. How the hybrid measure is expected to function differently from existing measures (e.g., it will assess the same construct but with better psychometric properties, or will assess a broader construct).

An example from our research with caregivers of individuals with Alzheimers disease or dementia is the Caregiver Health and Health Behaviors Form (REACH, 1995). Because no single existing measure assessed the full range of health and health-related behaviors important to our caregiver cohort, we drew items from several measures to examine specific aspects of health that might be particularly affected by the caregiving role. For example, items concerning perceived physical health and stress-related health symptoms were selected from the SF-36 (Ware and Sherbourne 1992; Ware et al. 1993), comorbidity items were selected from the AHEAD study Health Retirement Study; (Asset and Health Dynamics Among the Oldest Old 1993), and health behavior items were selected from the Nutrition Screening Initiative (NSI; Posner et al. 1993).

The psychometric work necessary for evaluating hybrid measures may equal that required for any category of measure discussed here. Advantages of utilizing items that come from well-established measures include the fact that most items have been evaluated for clarity, and the fact that tentative comparisons of responses to individual items as assessed in previous studies may be possible. Disadvantages include the fact that slight modifications to item stems or response categories are almost always necessary to enhance the flow of the items, and response categories of items from different scales are seldom similar. Dissimilar response categories that are retained in the hybrid measure may be confusing to respondents and make the necessity of transforming item distributions highly likely. For example, the health behavior items discussed above are dichotomous, while the stress-related symptoms are on a three-point scale, making item transformations necessary prior to computing the scale.

Because hybrid measures present items in a novel combination, often with some alterations in wording, heavy emphasis should be given to preliminary analyses in order to evaluate whether the items belong together in a scale. Item distributions, and inter-item and item-scale correlations should be carefully examined. In addition, given that the factor structure of this particular set of items will not have been previously evaluated, factor analysis should play a prominent role in the early analyses. At this stage, the results of the factor analysis can be used to make judgements about which items to retain or eliminate and about how (and whether) subscales will be computed.

At minimum, more than one technique for establishing the reliability of the hybrid measure should be utilized. As noted above, content validity concerns may have already been addressed in the creation of the hybrid measure and, in fact, may have been the primary justification for creating the measure in the first place. However, construct validity – and, if possible, criterion validity – should be evaluated.

New measures

The creation of a new measure should be undertaken only as a last resort, after a search for existing measures of the construct of interest has been conducted. The willingness of researchers to create new measures has led to an explosion of published instruments assessing similar constructs (e.g., more than 500 assessing depression, more than 3000 assessing health status), many with virtually no reported psychometric properties (Health and Psychosocial Instrumentation database, HaPI-CD; Behavioral Measurement Database Services 1997).

However, there are emerging research questions for which no appropriate instruments may exist; for example, physician attitudes about palliative care, acceptance of new forms of organ and tissue donation, and issues surrounding aging and caregiving. In other circumstances, although relevant instruments may exist, they may have poor psychometric properties that would be difficult to correct (e.g., questionable construct validity). In these situations, the creation of a new measure may be justified. Advantages of creating a new measure include the fact that researchers can

1. Conduct focus groups and reviews by experts to ensure that the content of the measure is specific to their research goals
2. Control item wording and response categories, and
3. Establish the length of the measure at the outset.

Disadvantages include

1. The intensive psychometric work that it is critical to conduct *prior* to analyzing (or even collecting) the data in terms of central study hypotheses
2. The possibility that a new measure may fail some critical reliability or validity criterion, and

3. The inability to compare results with any other previous research.

As a core measurement battery was being created for the REACH project, we reviewed the literature for a measure that would assess the burden of caring for a person with Alzheimers disease or dementia. Specifically, we were interested in the extent to which the care recipients demand a "vigilant" attitude in the caregiver. After identifying the construct of interest, a large number of items were generated, pilot tested, and reduced to a four-item scale asking about the length of time that the care recipient could be left alone, and the number of hours that the caregiver felt they must present and/or be actively involved in doing something for the care recipient. Data are being gathered currently with the Vigilance items, and a series of psychometric evaluations are planned.

In the process of developing and/or reporting on a newly created measure, comprehensive justifications for developing the measure – including a description of why the measure was necessary and the unavailability of appropriate established instruments – should be provided. The steps followed in generating ideas about specific items and the constructs they identify should be explained in detail. Ideally, a large pool of potential items should be generated on the basis of focus groups or expert opinion, pilot tested, re-evaluated, and reduced to form some final draft of the measure. As part of the process of identifying a construct and creating the measure, the purpose of the measure (i.e., proposed theoretical relationships) should also be clearly described. After data using the measure are collected, researchers should extensively evaluate the measure's reliability and validity.

For new measures especially, it is important to discuss how the instrument could be refined for application to other research questions. For example, items that may be altered or deleted should be identified, indications for additional psychometric work should be discussed, appropriateness of the measure for groups outside the normative sample should be addressed, and implications for the development of additional items or scales should be described.

Concluding comments

The value of any research effort rests to a high degree on the foundation of appropriate measurement. What constitutes "appropriateness" is a complex issue that has generated individual reports, full volumes, and entire journals devoted to the problem of ensuring that the assessment tools we use produce accurate information. Our primary aim in this paper was to reinforce the seriousness of measurement issues and to provide a basic template to guide nonexperts in the selection, development, and evaluation of study instruments. Because verification of the psychometric properties of an instrument – even those that are well established – is context-specific and must be re-established to some degree for every research effort, it is critical for all researchers to have a basic understanding of measurement issues.

References

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

American Psychological Association Committee on Psychological Tests (1952) Technical recommendations for psychological tests and diagnostic techniques: preliminary proposal. Am Psychol 7: 461–476

American Psychological Association Committee on Psychological Tests (1954) Technical recommendations for psychological tests and diagnostic techniques: preliminary proposal. Psychol Bull Suppl 51: 1–38

Anastasi A (1982) Psychological testing. Macmillan, New York

Aneshensel CS, Yokopenic PA (1985) Tests for the comparability of a causal model of depression under two conditions of interviewing. J Pers Soc Psychol 49: 1337–1348

Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA (1982a) Measuring depression in the community. Public Opin Q 46: 110–121

Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA (1982b) Telephone versus in person surveys of community health status. Am J Public Health 72: 1017–1021

Anthony JC, Folstein M, Romanoski AJ, Von Korff MR, Nestadt GR, Chahal R, Merchant A, Hendricks Brown C, Shapiro S, Kramer M, Gruenberg E (1985) Comparison of the lay diagnostic interview schedule and a standardized psychiatric diagnosis: experience in Eastern Baltimore. Arch Gen Psychiatry 42: 667–675

AHEAD: Asset and Health Dynamics among the Oldest Old (1993) Health Retirement Study. Institute for Social Research and the National Institute on Aging

Bartko JJ, Carpenter WT (1976) On the methods and theory of reliability. J Nerv Ment Dis 163: 307–317

Beck AT, Steer RA (1990) Manual, Beck Anxiety Inventory. The Psychological Corporation, Harcourt Brace Jovanovich, San Antonio

Beck AT, Ward CH, Mendelsohn M, Mock J, Erbaugh J (1961) An inventory for measuring depression. Arch Gen Psychiatry 4: 561–571

Behavioral Measurement Database Services (1997) HaPI- Health and Psychosocial Instruments. BMDS, Pittsburgh

Berkman LF (1986) The association between educational attainment and mental status examinations: of etiologic significance for senile dementias or not? J Chronic Disord 39: 171–174

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1: 307–310

Boyle GJ (1985) Self-report measures of depression: some psychometric considerations. Br J Clin Psychol 24: 45–59

Brayne C, Calloway P (1990) The association of education and socioeconomic status with the Mini-Mental State Examination and the clinical diagnosis of dementia in elderly people. Age Ageing 19: 91–96

Brugha TS, Kaul A, Dignon A, Teather D, Wills KM (1996) Present state examination by microcomputer: objectives and experience of preliminary steps. Int J Methods Psychiatr Res 6: 143–151

Butler MC, Jones AP (1979) The Health Opinion Survey reconsidered: dimensionality,reliability, and validity. J Clin Psychol 35: 554–559

Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 56: 81–105

Cannell CF, Marquis KH, Laurent A (1977) A summary of studies of interviewing methodology. Vital Health Stat 2: 1–78
Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20: 37–46
Cooper JE, Kendell RE, Gurland BJ, Sharpe L, Copeland JRM, Simon R (1972) Psychiatric diagnosis in New York and London. Oxford University Press, London
Costa PT, McCrae RR (1985) The NEO Personality Inventory. Psychological Assessment Resources
Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16: 297–334
Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52: 281–302
Dew MA (1993) Assessment and prevention of expectancy effects in community mental health studies. In: Blanck PD (ed) Interpersonal expectations: theory, research and application. Cambridge University Press, New York
Dew MA (in press) Psychiatric disorder in the context of physical illness. In: Dohrenwend BP (ed) Adversity, stress and psychopathology. American Psychiatric Press, Washington, DC
Dew MA, Bromet EJ (1993) Epidemiology. In : Bellack AS, Hersen M (eds) Psychopathology in Adulthood. Allyn and Bacon, Needham Heights, pp 21–40
Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Controlled Clin Trials 12: 142S–158S
Dignon AM (1996) Acceptability of a computer administered psychiatric interview. Comput Hum Behav 12: 177–191
Eaton WW (1980) The sociology of mental disorders. Praeger, New York
Erdman HP, Klein MH, Greist JH, Skare SS, et al (1992) A comparison of two computer-administered versions of the NIMH Diagnostic Interview Schedule. J Psychiatr Res 26: 85–95
Fenig S, Levav I, Kohn R, Yelin N (1993) Telephone vs. face-to-face interviewing in a community psychiatric survey. Am J Public Health 83: 896–898
Fenig S, Bromet EJ, Jandorf L, Schwartz JE, Lavelle J, Ram R (1994) Eliciting psychotic symptoms using a semi structured diagnostic interview: the importance of collateral sources of information in a 1st admission sample. J Nerv Ment Dis 182: 20–26
Fleiss JL (1981) Statistical methods for rates and proportions. Wiley, New York
Fowler FJ (1984) Survey research methods. Sage, Beverly Hills
Friedman WJ (1993) Memory for the time of past events. Psychol Bull 113: 44–66
Guarnaccia PJ, Angel R, Lowe Worobey J (1989) The factor structure of the CES-D in the Hispanic Health and Nutrition Examination Survey: the influences of ethnicity, gender, and language. Soc Sci Med 29: 85–94
Guarnaccia PJ, Good BJ, Kleinman A (1990) A critical review of epidemiological studies of Puerto Rican mental health. Am J Psychiatry 147: 1449–1456
Helmstadter GC (1964) Principles of psychological measurement. Appleton-Century-Crofts, New York
Hertzog C, Van Alstine J, Usala PD, Hultsch DF, Dixon R (1990) Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. J Consult Clin Psychol 2: 64–72
Jenkinson C, Layte R, Jenkinson D, Lawrence K, Petersen S, Paice C, Stradling J (1997) A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? J Public Health Med 19: 179–186
Kay DWK, Henderson AS, Scott R, et al (1985) Dementia and depression among the elderly living in the Hobart community: the effect of the diagnostic criteria on the prevalence rates. Psychol Med 15: 771–788
Kirk SA, Kutchins H (1992) The selling of DSM: the rhetoric of science in psychiatry. Aldine de Gruyter, New York
Kobak KA, Reynolds WM, Greist JH (1993) Development and validation of a computer-administered version of the Hamilton Rating Scale. Psychol Assess 5: 487–492
Kobak KA, Greist JH, Jefferson JW, Katzelnick DJ (1996) Computer administered clinical rating scales: a review. Psychopharmacology 127: 291–301
Kramer M (1961) Some problems for international research suggested by observations on differences in first admission rates to the mental hospitals of England and Wales and the United States. Proc Third World Congr Psychiatry 3: 153–160
Kraemer HC (1992) Coping strategies in psychiatric clinical research. In: Kazdin AE (ed) Methodological issues and strategies in clinical research. American Psychological Association, Washington, D.C.
Lewis G (1994) Assessing psychiatric disorder with a human interviewer or a computer. J Epidemiol Community Health 48: 207–210
Madow WG (1973) Net differences in interview data on chronic conditions and information derived from medical records. Vital Health Stat 2: 1–58
Maier W, Philipp M, Heuser I, et al (1988) Improving depression severity assessment. I. Reliability, internal validity and sensitivity to change of three observer depression scales. J Psychiatr Res 22: 3–12
Manson SM, Ackerson LM, Wiegman Dick R, Baron AE, Fleming CM (1990) Depressive symptoms among American Indian adolescents: psychometric characteristics of the Center for Epidemiologic Studies Depression Scale (CES-D). J Consult Clin Psychol 2: 231–237
McDowell I, Newell C (1996) Measuring health: a guide to rating scales and questionnaires, 2nd edn. Oxford University Press, New York
McKinley RL (1989) Methods, plainly speaking: an introduction to item response theory. Meas Eval Counsel Dev 22: 37–57
Meenan RF, Mason JH, Anderson JJ, et al (1992) AIMS2: the content and properties of a revised and expanded Arthritis Impact Measurement Scales health status questionnaire. Arthritis Rheum 35: 1–10
Miller IW, Bishop S, Norman WH, et al (1985) The Modified Hamilton Rating Scale for Depression: reliability and validity. Psychiatry Res 14: 131–142
Millon T (1983) Millon Clinical Multiaxial Inventory, 3rd edn. National Computer Systems, Minneapolis
Murden RA, McRae TD, Kaner S, et al (1991) Mini-Mental State Exam scores vary with education in blacks and whites. J Am Geriatr Soc 39: 149–155
Murphy JM (1976) Psychiatric labeling in cross-cultural perspective. Science 191: 1019–1028
Murphy JM (1995) Diagnostic schedules and rating scales in adult psychiatry. In: Tsuang MT, Tohen M, Zahner GEP (eds) Textbook in Psychiatric Epidemiology. John Wiley, New York, pp 253–271
Murphy JM, Neff RK, Sobol AM, Rice JX, Olivier DC (1985) Computer diagnosis of depression and anxiety: the Stirling County Study. Psychol Med 15: 99–112
Nunnally JC (ed) (1978) Psychometric theory, 2nd edn. McGraw-Hill, New York
Nunnally JC, Bernstein IH (1994) Psychometric theory, 3rd edn. McGraw-Hill, New York
Paulsen AS, Crowe RR, Noyes R, Pfohl B (1988) Reliability of the telephone interview in diagnosing anxiety disorders. Arch Gen Psychiatry 45: 62–63
Posner BM, Jette AM, Smith KW, Miller DR (1993) Nutrition and health risks in the elderly: The nutrition screening initiative. Am J Public Health 83: 944–945
Radloff LS (1977) The CES-D Scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1: 385-401
REACH: Resources for Enhancing Alzheimer's Caregiver Health (1995-2000) National Institute on Aging and National Institute of Nursing Research
Roberts R, Rhoades HM, Vernon SW (1990) Using the CES-D Scale to screen for depression and anxiety: effects of language and ethnic status. Psychiatry Res 31: 69–83

Rosenthal R, Rosnow RL (1991) Essentials of behavioral research: methods and data analysis, 2nd edn. McGraw-Hill, New York
Shrout P (1994) The NIMH Epidemiologic Catchment Area Program: broken promises and dashed hopes? Int J Methods Psychiatr Res 4: 113–122
Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86: 420–428
Shrout P (1995) Reliability. In: Tsuang MT, Tohen M, Zahner GEP (eds) Textbook in psychiatric epidemiology. John Wiley, New York, pp 213–227
Shrout PE, Yager TJ (1989) Reliability and validity of screening scales: effect of reducing scale length. J Clin Epidemiol 42: 69–78
Steer RA, Rissmiller DJ, Ranieri WF, Beck AT (1994) Use of the computer administered Beck Depression Inventory and Hopelessness Scale with psychiatric inpatients. Comput Hum Behav 10: 223–229
Stewart AL, Ware JE Jr (eds) (1992) Measuring functioning and well-being: the Medical Outcomes Study approach. Duke University Press, Durham, NC
Stewart AL, Hays RD, Ware JE Jr (1988) The MOS Short-Form General Health Survey: reliability and validity in a patient population. Med Care 26: 724–735
Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R (1993) Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). Psychiatry Res 49: 239–250
Sue S, Fujino DK, Hu L, Takeuchi DT (1991) Community and mental health services for ethnic minority groups: a test of the cultural responsiveness hypothesis. J Consult Clin Psychol 59: 533–540
Talbott MM (1989) Age bias in the Beck Depression Inventory: a proposed modification for use with older women. Clin Gerontol 9: 23–35
Thornicroft G (1992) Computerised mental health assessments. In: Thornicroft G, Brewin CR, Wing J (eds) Measuring mental needs. University of London
Uhlmann RF, Larson EB (1991) Effect of education on the Mini-Mental State Examination as a screening test for dementia. J Am Geriatr Soc 39: 876–880
Ullman JB (1996) Structural equation modeling. In: Tabachnick BG, Fidell LS (eds) Using multivariate statistics, 3rd edn. Harper Collins, pp 709–811
Ware JE (1984) The General Health Rating Index. In: Wenger NK, Mattson ME, Furberg CD, Elinson J (eds) Assessment of quality of life in clinical trials of cardiovascular therapies. Le Jacq, New York, pp 184–188
Ware JE Jr, Sherbourne CD (1992) The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. Med Care 30: 473–483
Ware JE Jr, Sherbourne CD, Davies AR (1992) Developing and testing the MOS 20-item Short-Form Health Survey: a general population application. In: Stewart AL, Ware JE Jr (eds) Measuring functioning and well-being: the Medical Outcomes Study approach. Duke University Press, Durham, NC, pp 277–290
Ware JE Jr, Snow KK, Kosinski M, et al (1993) SF-36 Health Survey: Manual and Interpretation Guide. The Health Institute, New England Medical Center, Boston
Ware JE Jr, Kosinski M, Keller SD (1996) A 12-item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 34: 220–233
Wells KB, Burnam MA, Leake B, Robins LN (1988) Agreement between face-to-face and telephone-administered versions of the depression section of the NIMH Diagnostic Interview Schedule. J Psychiatr Res 22: 207–220
Williams JBW (1992) The structured clinical interview for DSM-III-R (SCID). II. Multisite test-retest reliability. Arch Gen Psychiatry 49: 630–636
Williams AC, Richardson PH (1993) What does the BDI measure in chronic pain? Pain 55: 259–266
World Health Organization (1992) International statistical classification of diseases and related health problems: ICD-10, vol I, 10th edn revised. World Health Organization, Geneva