JOURNAL OF PERSONALITY ASSESSMENT, *80*(3), 217–222

STATISTICAL DEVELOPMENTS AND APPLICATIONS

Being Inconsistent About Consistency: When Coefficient Alpha Does and Doesn't Matter

David L. Streiner

Baycrest Centre for Geriatric Care
Department of Psychiatry
University of Toronto

One of the central tenets of classical test theory is that scales should have a high degree of internal consistency, as evidenced by Cronbach's α, the mean interitem correlation, and a strong first component. However, there are many instances in which this rule does not apply. Following Bollen and Lennox (1991), I differentiate between questionnaires such as anxiety or depression inventories, which are composed of items that are manifestations of an underlying hypothetical construct (i.e., where the items are called *effect indicators*) and those such as Scale 6 of the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1943) and ones used to tap quality of life or activities of daily living in which the items or subscales themselves define the construct (these items are called *causal indicators*). Questionnaires of the first sort, which are referred to as *scales* in this article, meet the criteria of classical test theory, whereas the second type, which are called *indexes* here, do not. I discuss the implications of this difference for how items are selected, the relationship among the items, and the statistics that should and should not be used in establishing the reliability of the scale or index.

In a previous article (Streiner, 2003), I discussed that one of the major tenets of classical test theory is that all of the items in a scale tap a single domain or attribute such as anxiety or achievement motivation. A consequence of this assumption is that the scale has a high degree of internal consistency, reflecting strong correlations among the items. However, there are a number of indexes that do not fall into this mold. As everyone who has had a child since 1953 knows, newborns are evaluated on a five-item Apgar Scale (Apgar, 1953), which rates the infant's heart rate, respiration, muscle tone, reflex response, and skin color as being either 0, 1, or 2. Whereas the correlation among these items may be high for healthy infants, the relationship among the items breaks down for those with serious medical problems. Neonates with neurological difficulties may score low on muscle tone and reflex response but have no problems in the other three areas. On the other hand, those with cardiac problems will score 0 or 1 for heart rate and skin color, but 2 on the remaining items. In fact, many questionnaires used to assess activities of daily living (ADL), instrumental ADL, and quality of life (QOL) appear to fit this latter model closer than they do the classical test model in that the correlations among the items may be low, often deliberately so. In this article, I discuss these two different models of test construction, when each should be used, and the test statistics that should and should not be used with each.

One difficulty in discussing this, though, is terminology. Many terms have been used to describe a collection of items or questions—*scale, test, questionnaire, index, inventory,* and a host of others—with no consistency from one author to another. For example, both the Apgar Scale (Apgar, 1953) and the Hamilton Depression Rating Scale (Hamilton, 1967) are called *scales,* although the former consists of unrelated items, whereas the latter falls into the more traditional model of highly correlated items; similar examples can be found for the other terms. To simplify matters in this article, I refer to questionnaires that are composed of theoretically correlated items as *scales* and those that consist of unrelated items as *indexes,* recognizing that counterexamples of each term can readily be found.[1]

[1]Feinstein (1987) used the term *clinimetric scales* to refer to what are called indexes in this article and similarly proposed the term *clinimetrics,* in contrast to *psychometrics,* to reflect the fact that different statistical approaches are used with them. Although the terms

As with most things in life, this is somewhat of an oversimplification. No scale is comprised of items that are perfectly correlated with each other and as pointed out in the previous article (Streiner, 2003), even very highly correlated items are usually avoided because they result in unnecessary redundancy and length at the cost of breadth of scope of the instrument. Similarly, there is always some degree of correlation among the items in an index if for no other reason than Meehl's (1990) sixth law that everything is correlated with everything else (Does anyone remember his other nine laws? If you do not, see footnote 2). However, the distinction holds in terms of the underlying theoretical models and in the way items are chosen.

CONSTRUCTION OF SCALES

The Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) is perhaps the epitome of an atheoretical, "dust bowl empiricism" approach to the development of a tool to measure personality traits. Basically, a large number of potential items derived from a variety of sources were gathered. Then, one by one, each item was tested to determine if a larger proportion of people in a criterion group answered it in a given direction as compared to the comparison group. If so, it ended up in the target scale without any regard to the content of the item. This resulted in items that may discriminate well but in many cases whose relationship to the underlying trait remains a source of bafflement. For example, it is a complete mystery why not having hay fever or asthma is indicative of depression, but that is what this item purportedly taps.

More recent tests, such as the Personality Research Form (Jackson, 1984) or the Personality Assessment Inventory (Morey, 1991) were developed in light of Cronbach and Meehl's (1955) classic article on hypothetical constructs. *Hypothetical constructs* (which are similar to what statisticians call *factors* or *latent traits*) refer to attributes that cannot be seen directly but only inferred by the hypothesized effects they have on observable behaviors. For example, we cannot see or measure intelligence (according to some professors and conservative political theorists because it exists in such limited quantities). Rather, we observe and thus can measure the extent of people's vocabulary, their mathematical ability, ease in working out puzzles, and knowledge about the world around them. We hypothesize that the correlation among the measures is a result of them all being reflections of the same underlying trait of intelligence.

There are a number of implications of this approach. First, it assumes that there is a "universe" of potential items that can tap the construct, and the ones that appear on a test are a sample from this universe. Second, because the items are all drawn from the same universe and measure the same construct, they should be correlated with each other to varying degrees. Using the pictorial conventions of structural equation modeling (SEM), this is shown in Figure 1. Note that the construct (depicted as a circle) has arrows coming from it leading to the observed variables or items drawn as rectangles. The lambdas (λ_i) in the diagram indicate the expected effect of the construct on the items (these are called *factor loadings* in factor analysis and *path coefficients* in SEM). This diagram reflects the assumption that the level or amount of the observed variables (e.g., the person's score on a vocabulary test) is caused by the hypothesized underlying trait. These observed (or measured) variables are often referred to as *effect indicators* (Bollen & Lennox, 1991; Fayers & Hand, 1997; Fayers, Hand, Bjordal, & Groenvold, 1997). Following the notation of Bollen and Lennox (1991), the relationship between the effect indicators (y_i) and the hypothetical construct or latent trait (η) can be written as

$$y_i = \lambda_i \eta + \varepsilon_i, \quad (1)$$

where the ε is the error term, which has a mean of zero and is not correlated with η.

As Bollen and Lennox (1991) pointed out, if we standardize all of the variables to have a mean of 0 and standard devi-

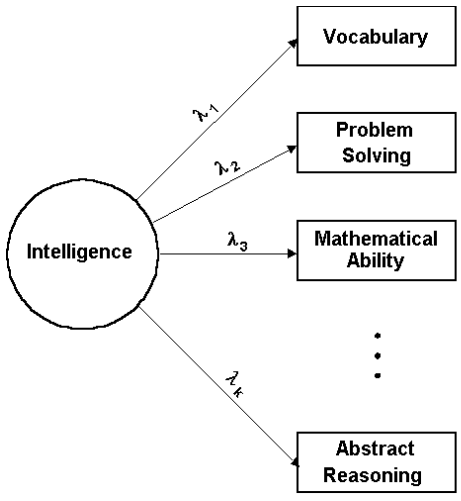

FIGURE 1 Measured variables as effect indicators.

still appear in the medical literature, they thankfully have not been adopted by psychologists and educational test constructors, as they reflect a narrow conceptualization of test theory.

[2]Meehl (1990) proposed 10 "obfuscating influences" (p. 195) that render research literature reviews "well-nigh uninterpretable" (p. 195). They are (1) a loose derivative chain between the theory and the experiment; (2) problematic auxiliary theories; (3) problematic *ceteris paribus* ["all other things being equal"] clause; (4) experimenter error; (5) inadequate statistical power; (6) the cited "crud factor"; (7) pilot studies [similar to what Rosenthal (1979) called the "file drawer problem"]; (8) selective bias in submitting reports that show significant differences; (9) selective editorial bias toward publishing positive findings; and (10) using tests that have not been validated for the use to which they are put.

ation of 1, then the correlation between items y_1 and y_2 is $\lambda_1\lambda_2$. Because each item should be positively correlated with the construct (i.e., all the λs are positive[3]), this means that all of the items must be correlated with each other, resulting in a high degree of internal consistency of the scale as measured by Cronbach's α or the mean interitem correlation. Turning this around, a component analysis performed on these items should result in a strong first component, with all of the remaining ones accounting for only a small amount of the total variance. Note that there are no curved, double-headed arrows between the variables, which in the convention of SEM would indicate correlations among them. This is because any covariances between the items are assumed to exist only because of their relationship to the underlying construct (Fayers & Hand, 2002) and that other factors that lead to covariation among the items, such as response style, are incorporated in the latent variable.

A third implication, which has direct consequences for scale construction, is that the specific items that appear on a scale, or the specific subscales that comprise a test battery, are somewhat arbitrary. If both sweatiness and fearfulness are effect indicators of anxiety and if they are highly correlated with each other, it is not necessary for both to be present on an inventory designed to measure the extent of anxiety (as opposed to measuring which specific aspects of anxiety may or may not be present). Whatever component of anxiety may be missed by omitting certain items will be picked up by the other ones that are on the scale. One may want to include both sweatiness and fearfulness to improve the face validity of the test, as some people may question whether it is really measuring anxiety if one of the items is absent, but the presence of the second item may do nothing with regard to improving the construct validity of the scale as long as enough other items are present to tap the full range of anxiety.

Note that these attributes of a scale—positive correlations among the items and the interchangeability of the items—pertain only to unidimensional scales (Bollen & Lennox, 1991). If our theoretical model of anxiety, for example, pictures it as having different facets (e.g., cognitive, affective, behavioral, physiological) that are themselves correlated only poorly (Antony, 2001), then these criteria for scale development would be applied to the subscales tapping each facet individually and not to the scale as a whole. That is, the anxiety scale would be seen to more resemble an inventory composed of four unidimensional subscales.

[3]This assumes that all of the items are scored in the same direction (e.g., higher scores reflecting more of the trait) and that items that are scored in the opposite direction have been reversed.

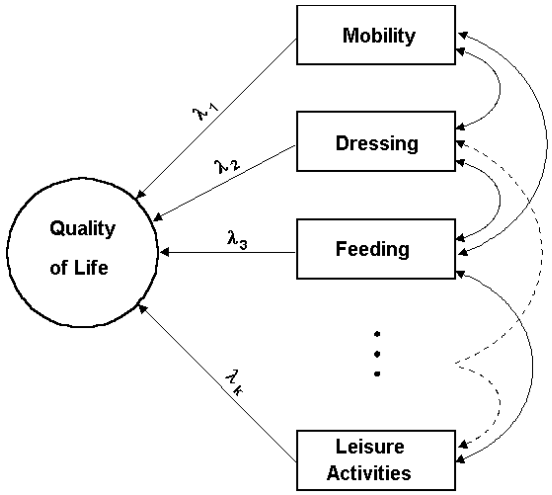

FIGURE 2 Measured variables as causal indicators.

CONSTRUCTION OF INDEXES

Now consider a different hypothetical construct, QOL. Like intelligence or anxiety, it cannot be observed directly but only inferred from other variables that can be measured. However, as seen by the arrows in Figure 2, the relationship between the observed variables and the construct flow in the opposite direction. In this case, it is the measured variables (which are now referred to as *causal indicators*) that influence the latent variable rather than the latent variable affecting the indicators.[4] Again following Bollen and Lennox (1991), one can write this as:

$$\eta = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_k x_k + \zeta, \quad (2)$$

where η again is the latent trait, λs are the path coefficients, *x*s are the measured variables, and ζ is a "disturbance" term equivalent to ε in Equation 1 and with the same properties—an expected mean of zero and uncorrelated with the *x*s. In Equation 1, the indicators are called *y,* whereas they are referred to as *x* in Equation 2 to emphasize the point that they can be seen as the dependent variable in the first case but the

[4]In many ways, *causal* is a poor choice of terms. The observed variable(s) influence the value of the construct but do not in all instances cause it to occur. For example, skin color on the Apgar Scale (Apgar, 1953) is an indicator of the neonate's condition but cannot be said to cause it. On the other hand, death of a spouse, which appears on the Holmes and Rahe (1967) scale, the Schedule of Recent Events, discussed later, is causal in affecting a person's stress level. As one anonymous reviewer pointed out, the difference may be due to the coherence of the construct being measured. The Apgar Scale is a screening device derived from items that have been found empirically to correlate with outcome, whereas Holmes and Rahe's scale is based on Cannon's (1939) theory of homeostasis. A better term may be *defining characteristic* or *defining indicator,* but the term *causal indicator* has achieved some degree of permanence at this time.

predictors in the second. A second important point is that the latent trait η is to the right of the equal sign in Equation 1, reflecting its role as a predictor; but it is to the left of the equal sign in Equation 2, as it is now the variable to be predicted from the others.

Finally, the curved arrows between the items in Figure 2 (not all of which are shown) indicate that there may be some correlation among them. However, the nature of these correlations is very different with causal indicators than with effect indicators. In the latter case, as was pointed out earlier, the correlations were due solely to the relationships of the items with the underlying construct, and their magnitude is the product of the λs for each pair of items. For causal indicators, however, covariances may or may not exist among them, irrespective of their relationship with the construct (Fayers & Hand, 2002); using Bollen and Lennox's (1991) notation, $\text{CORR}(x_1, x_2) = ?$. Furthermore, as was seen with the example of the Apgar Scale (Apgar, 1953), and as I discuss in more detail later, the magnitude of the correlations may change radically from one population to another: positive with healthy children and absent or even negative with sickly ones.

I illustrate the difference with some examples. If people are able to engage in more leisure activities because of a job change, their QOL will improve even though there has been no change in their ability to get around, get dressed, or feed themselves. Conversely, a focal stroke that affects a person's mobility will decrease his or her QOL even in the absence of change in the other domains. Indeed, each of the domains themselves may be measured with a tool that is more like an index than a scale. For example, within the realm of dressing, one woman may have rotator cuff damage that prevents her from reaching behind to do up her bra but does not interfere with her ability to put on stockings or a skirt. Another person may have severe arthritis of the fingers, affecting his or her ability to button a shirt but not to slip on a sweater.

In a different area, Holmes and Rahe (1967) developed a measure called the Schedule of Recent Events (SRE) based on the homeostatic theory of stress. Holmes and Rahe proposed that any change in a person's life requires adaptation and that too many changes predispose people to stress and hence makes them susceptible to illness. Their index consists of a list of 43 or so recent events (the actual number varies from one version to another), ranging from getting a parking ticket, to buying a home, to the death of a spouse, with a weight of 500 assigned to getting married. The hypothetical construct, or latent trait, of stress is a result of these purported stressors; it is not the case that the construct causes them. Hence, in the terminology of this article, this would be called an index composed of causal indicators rather than a scale (which is how Holmes and Rahe, 1967, referred to it) of effect indicators.

Similarly, global measures of symptomatology such as the General Health Questionnaire (Goldberg, 1979) or the Langner Scale of Psychophysiological Strain (Langner, 1962) may be closer to indexes than scales. Whereas disorder-specific scales such as those for anxiety or depression consist of items that are manifestations of the underlying trait, global indexes consist of a shopping list of symptoms that arise from a number of different disorders. Not only may the specific items not be related, some may in fact be mutually exclusive (e.g., anorexia and sudden weight increase).

The implications of this for developing an index are almost completely opposite to the ones I discussed for constructing a scale. Most important, there is no assumption that the individual items need be correlated with each other. Some may be positively related (difficulty buttoning one's shirt because of arthritis is likely accompanied by problems tying one's shoes), some may be negatively correlated (e.g., mania and lethargy), and others may not be related at all (for example, the items on the Goldberg, 1979, General Health Questionnaire and the Holmes & Rahe, 1967, SRE indexes). Consequently, it would be inappropriate to use statistics that are based on the assumption of homogeneity of the items, such as coefficient α, the mean interitem correlation, or factor analysis. In fact, a high value of α or a very strong first component may point to deficiencies in the index and that rather than tapping a broad set of causal indicators, the actual items may be too narrowly focused. Indeed, the SRE has been criticized on the grounds that the correlations among some items are too high (Zimmerman, 1985).

The use of the word *may* in the previous sentences is not simply a reflexive tic reflecting the inability of an academic to avoid saying, "On the other hand …." Rather, it points to a related issue: that the correlations among the items and the factor structure as a whole are much more highly dependent on the sample than is the case with scales. This is especially true for indexes that tap symptoms, treatment side effects, and the like. As an example, consider an index of ADL that consists of items such as the ability to tie one's shoelaces, to pick up small objects (e.g., coins), and to climb a flight of stairs. Because all of these activities are adversely affected by rheumatoid arthritis, one would expect that the interitem correlations would be high, and they would emerge as part of a common factor if the index had been given to these patients. However, people with a hip dysplasia would find it difficult to climb the stairs but have no trouble with the first two items, whereas those with back problems may find it hard to bend to tie their shoes and also have problems with stair climbing, but their physical limitation would not involve grasping small objects. Therefore, the internal structure of the index depends on the group being studied, and inconsistencies from one study to the next do not necessarily mean that the index is unreliable (Fayers & Hand, 1997).

When developing an index, the choice of the specific items is much more important than is the case in the construction of scales. Because the items may be uncorrelated, it cannot be assumed that what is missed if one item is omitted will be covered by the others that remain. For example, if the Holmes and Rahe (1967) SRE did not include an item tapping divorce, then the stress caused by this life event will be missed completely.

Thus, slight differences among indexes purported tapping the same construct may yield very different results regarding the magnitude of the construct and even its dimensionality. For this reason, Bollen and Lennox (1991) stated that "With causal indicators, we need a census of the indicators, not a sample. That is, all indicators that form η_1 should be included" (p. 308). The problem that naturally arises is in arriving at this census. This is an issue of the content validity of the index and is highly dependent on our underlying theory of the construct and prior research. There is nothing beforehand to tell us whether all relevant areas have been included and only a lack of construct validity and critical articles from our colleagues to alert us to the fact afterwards.

A different aspect of content validity is that it is method specific. Free-response performance measures of personality such as the Thematic Apperception Test (Murray, 1943) or the Rorschach (Rorschach, 1942), create a census of indicators from the summary scores that are derived. For example, the Intellectualization Index on the Rorschach combines AB, Art, and AY. Although this may be a total census of the indicators that are available from this instrument, it is inherently limited by what the Rorschach can and cannot measure. This problem in the content coverage of indexes is not limited to free-response techniques; all instruments are limited in this regard. Questionnaires may offer more flexibility regarding their content, but their content validity is still never perfect, and there is no guarantee that a respondent will answer a question honestly or even at all.

CAUSE OR EFFECT

The differentiation of questionnaires into scales and indexes represent the two ends of a continuum. In reality, there are many that fall somewhere in between where it is difficult to determine if certain items are effect indicators or causal indicators. For example, an item such as "I have few close friends" could reflect the demoralization and anhedonia associated with depression; that is, it could be an effect indicator. On the other hand, the lack of friendship could itself be a cause of the dysphoric mood and thus be a causal indicator. In a different realm, the SRE has been criticized because some of the recent events that are purported to be causes of stress, such as changes in eating or sleeping patterns, may in fact be reactions to it (Zimmerman, 1985). Part of the problem is that psychological problems rarely fit the simple model of a cause leading to an outcome in a linear fashion. Rather, there are often feedback loops with symptoms of a disorder exacerbating the disorder itself. The anger and suspiciousness that are hallmarks of patients with paranoid disorders have the effect of driving people away, reinforcing their belief that others are hostile and angry with them and are to be avoided. Many similar examples can be found, especially in the domains of QOL and symptom scales.

Another aspect of the problem is that psychologists' knowledge of many disorders and psychological states is far from complete, resting primarily on correlations among variables so that it is difficult to determine what is a cause and what is an effect. For example, MacMillan et al. (1999) found an association between corporal punishment and childhood psychopathology. It is possible that spanking leads to psychological problems (the way the results were reported in the popular press), but it is just as likely that the physical punishment is a result of the parents' frustration dealing with a child who has behavioral problems, or that some third factor, such as socioeconomic status or parental education, results in both a greater tendency to use physical means of control and the higher prevalence of some forms of disorder. Without a clear model of the link between punishment and psychopathology, it is impossible to state a priori in which direction the arrow should point.

Models and theory are necessary because in the absence of experimental interventions, which are difficult if not impossible in the areas of personality assessment and ADL, it is extremely difficult to establish causation. In 1965, Hill proposed nine criteria that can be used in medicine for assessing the probability that there is a causal relationship between two factors. Among others, these include the strength and consistency of the association, its specificity, temporal relationship, and plausibility. However, even if all nine criteria are met, it cannot be said that causality has been proven, only that it is more likely than if fewer have been satisfied. More recently, Fayers and Hand (1997) offered another method. A matrix is calculated with the levels of the target item as the columns and those of an external criterion as the rows. Items that are effect indicators (i.e., a scale model) should have most of the people clustered along the main diagonal, reflecting the hypothesized moderate to strong correlation of a construct (measured by the criterion) with items tapping it. On the other hand, because each person may have a different pattern of causal indicators, more people should appear on the off-diagonal cells when items are used from an index model. The first difficulty with this approach, though, is that it is often difficult to find an uncontaminated criterion measure other than asking the person a global question such as "How would you rate your quality of life?" A second problem is that there is no statistical test that can be used to measure this, and the judgment becomes a very subjective one in interpreting the pattern of responding. Again, therefore we are forced to rely on theory to determine whether a measurement tool should be analyzed as if it were a scale or an index.

Just to complicate matters a bit, there are some questionnaires that are a combination of both scales and indexes. For example, the three Harris–Lingoes (Harris & Lingoes, 1968) subscales of Scale 6 of the MMPI (Pa_1 – persecutory ideas, Pa_2 – poignancy, and Pa_3 – naïvete) are each scales as I am using the term here. Based on over 1,000 cases, Miller and Streiner (1985) found that Cronbach's αs were .78, .54, and .70, respectively, reflecting moderate to high

internal consistency. However, whereas the correlation between Pa_1 and Pa_2 is moderate and positive (.54), those between Pa_1 and Pa_3 and between Pa_2 and Pa_3 are modest and negative (–.40 and –.27, respectively). Thus, the three subscales together constitute an index because clinically, when all are elevated, the patient has a classic paranoid personality structure.

CONCLUSIONS

In a previous article (Streiner, 2003), I pointed out that not all forms of reliability testing should, or even can, be used with all measurement tools. It makes no sense, for instance, to try to assess interrater reliability for self-report measures. The same can be said for measures of internal consistency. They are extremely useful in constructing scales that tap a unidimensional construct, but one should not assume that all measures must exhibit homogeneity among the items. Specifically, indexes, which are composed of causal indicators, most often do not have items that are correlated with each other. The blind use of coefficient α and other indexes of internal consistency, without considering whether they are appropriate for the measure, can lead to situations in which either a scale is wrongly dismissed for not being reliable or the indexes are unfairly criticized for not yielding useful results (e.g., Juniper, Guyatt, & King, 1994). Rather, one should recognize that different measurement tools rest on varying assumptions about the underlying nature of the relationships, and the statistics should mirror them.

REFERENCES

Antony, M. M. (2001). Assessment of anxiety and the anxiety disorders: An overview. In M. M. Antony, S. M. Orsillo, & L. Roemer (Eds.), *Practitioner's guide to empirically based measures of anxiety* (pp. 7–17). New York: Kluwer Academic/Plenum.

Apgar, V. (1953). A proposal for new method of evaluation of the newborn infant. *Current Research in Anesthesia and Analgesia, 32,* 260–267.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Cannon, W. B. (1939). *Wisdom of the body.* New York: Norton.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Fayers, P. M., & Hand, D. J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research, 6,* 139–150.

Fayers, P. M., & Hand, D. J. (2002). Causal variables, indicator variables and measurement scales: An example from quality of life. *Journal of the Royal Statistical Society (Statistics in Society), Series A, 165*(Part 2), 233–261.

Fayers, P. M., Hand, D. J., Bjordal, K., & Groenvold, M. (1997). Causal indicators in quality of life research. *Quality of Life Research, 6,* 393–406.

Feinstein, A. R. (1987). *Clinimetrics.* New Haven, CT: Yale University Press.

Goldberg, D. P. (1979). *Manual of the General Health Questionnaire.* Windsor, England: NFER.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology, 6,* 278–296.

Harris, R., & Lingoes, J. (1968). *Subscales for the Minnesota Multiphasic Personality Inventory.* Unpublished manuscript, Langley Porter Clinic, San Francisco.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory.* Minneapolis: University of Minnesota Press.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58,* 295–300.

Holmes, T. H., & Rahe, R. H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research, 11,* 213–218.

Jackson, D. N. (1984). *Personality Research Form manual.* Port Huron, MI: Research Psychologists Press.

Juniper, E. F., Guyatt, G. H., & King, D. R. (1994). Comparison of methods for selecting items for a disease-specific quality-of-life questionnaire—Importance versus factor-analysis. *Quality of Life Research, 3,* 52–53.

Langner, T. S. (1962). A twenty-two item screening score of psychiatric symptoms indicating impairment. *Journal of Health and Social Behavior, 3,* 269–276.

MacMillan, H. L., Boyle, M. H., Wong, M. Y., Duku, E. K., Fleming, J. E., & Walsh, C. A. (1999). Slapping and spanking in childhood and its association with lifetime prevalence of psychiatric disorders in a general population sample. *Canadian Medical Association Journal, 161,* 805–809.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66* (Monograph Supplement 1), 195–244.

Miller, H. R., & Streiner, D. L. (1985). The Harris–Lingoes subscales: Fact or fiction? *Journal of Clinical Psychology, 41,* 45–51.

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

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Rorschach, H. (1942). *Psychodiagnostics: A diagnostic test based on perception.* Berne, Germany: Huber.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha. *Journal of Personality Assessment, 80,* 99–103.

Zimmerman, M. (1985). Methodological issues in the assessment of life events: A review of issues and research. *Clinical Psychology Review, 3,* 339–370.

David L. Streiner
Baycrest Centre for Geriatric Care
Department of Psychiatry
University of Toronto
3560 Bathurst Street
Toronto, Ontario, M6A 2E1
Canada
E-mail: dstreiner@klaru-baycrest.on.ca

Received October 15, 2002
Revised November 21, 2002