Introduction to This Issue

This issue of The Behavioral Measurement Letter has three featured articles: one describes a process through which measurement instruments may be improved in various ways and outlines a model for revising instruments; another discusses self-estimated intelligence, gender differences in overestimations and underestimations of intelligence, and possible explanations for and consequences of these differences; the third piece presents ten very practical guidelines -- a list of "do's and don'ts" -- for selecting measurement instruments that will best meet one’s needs.

In “Refinements to the Lubben Social Network Scale: The LSNS-R,” James Lubben, Melanie Gironda, and Alex Lee describe in detail the process of successive changes and analyses thereof through which they modified the Lubben Social Network Scale to improve it in various ways. The process involved performing principal component analysis to determine which items in the instrument contribute to underlying factors being measured by the instrument, and thus which items best measure what they want the instrument to measure. This analytical process resulted in a "cleaner and meaner" version of the LSNS, the LSNS-R. In addition, by detailing the process of successive modifications and mathematical analyses that produced the LSNS-R, the authors outline a model for instrument revision that is broadly applicable.

In a piece titled “Self-Estimated Intelligence,” Adrian Furnham from London University reviews literature pertaining to estimation of intelligence by oneself and others, and then compares, contrasts, and offers explanations for the reported findings. The literature he reviewed shows that there is a gender difference in intelligence estimation -- females tend to underestimate their IQ and that of other females, while males, on the other hand, tend to overestimate their IQ and that of other males, and that this gender difference exists across cultures and nationalities, across socio-economic classes, and across age groups. He then explores possible explanations for the gender difference and suggests possible life consequences of over- and under-estimating one’s abilities.

Our regular contributor, Fred Bryant, presents a set of guidelines for choosing measurement tools, “Ten Commandments for Selecting Self-Report Instruments.” Some of the guidelines may seem to be common-sense. Many readers will be familiar with their basic content. Those experienced in instrument selection may see mistakes they’ve made in the past (and vowed never to make again). In any case, all of our readers, from neophytes to seasoned instrument users, will find value in the piece -- if not something they hadn’t known previously, then, certainly, practical reminders of the numerous subtleties, cautions, and pitfalls that should be attended to while selecting a measurement instrument.

Address comments and suggestions to The Editor, The Behavioral Measurement Letter, Behavioral Measurement Database Services, PO Box 110287, Pittsburgh, PA 15232-0787. If warranted and as space permits, your communication may appear as a letter to the
Since the measuring device has been constructed by the observer... we have to remember that what we observe is not nature in itself but nature exposed to our method of questioning.

Werner Karl Heisenberg
Refinements to the Lubben Social Network Scale: The LSNS-R (continued)

Hurwicz & Berkanovic, 1993; Rubenstein, L. et al., 1994; Rubenstein, R., Lubben, & Mintzer, 1994; Dorfman, Walters, Burke, Hardin, & Karanik, 1995; Luggen & Rini, 1995; Lubben & Gironda, 1997; Okwumabua, Baker, Wong, & Pilgrim, 1997; Mor-Barak, 1997; Gironda, Lubben, & Atchison, 1998; Chou & Chi, 1999; Martire, Schulz, Mittelmark, & Newsom, 1999; Mistry, Rosansky, McGuire, McDermott, & Jarvik, 2001). Further, the LSNS has been employed in a variety of ways, including use as a control variable as well as an outcome variable in health and social science studies. It also has been used as a screening tool for health risk appraisals and as a “gold” standard by which to evaluate other social network assessment instruments.

The LSNS was developed as an adaptation of the Berkman-Syme Social Network Index (BSSNI; Berkman & Syme, 1979). Whereas the LSNS was developed specifically for use among elderly populations, the BSSNI was initially developed for a study of an adult population that purposefully excluded older persons. The LSNS is based on items borrowed from questionnaires used in the original epidemiological study for which the BSSNI was constructed. However, the LSNS excluded BSSNI items dealing with secondary social relationships (viz., group and church membership) because these organizational participation items showed limited variance when used with older populations, especially those having large numbers of frail elderly persons (Lubben, 1988). In contrast, the LSNS elaborated on an array of items dealing with the nature of relationships with family and friends, in view of the growing body of empirical data suggesting that the structure and functions of kinship and friendship networks are particularly salient to the health and well-being of older persons.

The LSNS total scale score is computed by summing ten equally weighted items that quantify structural and functional aspects of primary social relationships. Scores for each LSNS item range from zero to five, with lower scores indicating smaller networks. The scale has been found to have relatively good internal consistency among a widely diverse set of study populations (α = 0.70). Factor analyses on the LSNS suggest that it measures three different types of social networks: family networks, friendship networks, and interdependent relationships (Lubben, 1988; Lubben & Gironda, 1997).

Methodology

The main purpose of the work presented here is to address deficits that became apparent in the original LSNS as it was used with diverse populations over the past decade. But because the LSNS was found to have relatively stable reliability and validity across this wide array of settings, any proposed modifications were not to jeopardize the relatively strong psychometric properties of the original LSNS.

Reliability is a fundamental issue in psychological measurement (Nunnally, 1978). One important type of measurement reliability is internal consistency, i.e., the extent to which items within a scale relate to the latent variable being measured (DeVellis, 1991; Streiner & Norman, 1995). Cronbach’s (1951) coefficient alpha was chosen to examine the internal consistency of the LSNS and modifications designed to improve upon the original version. The acceptable range of coefficient alpha values employed here was 0.70 to 0.90 (Nunnally; DeVellis) because assessment instruments with reliability scores higher than 0.90 are likely to suffer from excessive redundancy, whereas those with alpha less than 0.70 are likely to be unreliable (Streiner & Norman). A further test of item homogeneity used was the item-total test score correlation (DeVellis; Streiner & Norman). Here acceptable values of the item-total score correlation were 0.20 and greater (Streiner & Norman).

Principal component analysis looks for underlying (latent) components that account for most of the variance of a scale (Stevens, 1992). Principal component analysis with varimax rotation was used here to explore the component structure of various versions of the LSNS to see
Refinements to the Lubben Social Network Scale: The LSNS-R (continued)

if the modified versions conformed in actuality to the hypothesized structure. Although more sophisticated methods exist to examine factor or latent variable structures, such as maximum likelihood factor analysis and confirmatory factor analysis, many scholars contend that principal component analysis is both adequate and yet more practical than more sophisticated techniques, for principal component analysis is mathematically easier to manage, easier to interpret, and yields results similar to those from maximum likelihood factor analysis (Nunnally, 1978; Stevens). The size of the sample used in the analyses discussed below is adequate to conduct principal component analysis according to general sample size guidelines (Stevens; Guadagnoli & Velicer, 1988).

Four Objectives Used in Refining the LSNS

The work to refine the LSNS has four principal objectives. One was to distinguish between and better specify the nature of family and friendship social networks. A second was to replace, where feasible, items in the original LSNS that have small statistical variance. A third objective was to disaggregate “double-barreled” questions. The fourth was to produce a parsimonious instrument to encourage and facilitate its use in research and practice settings where time constraints or other issues preclude using longer social support network instruments.

With regard to the first two objectives above, it should be noted that social and behavioral measures are purposely designed to discriminate among groups for a certain construct, and so lack of variation within a given item limits a scale’s ability to identify and discriminate among variations (DeVellis, 1991; Streiner & Norman, 1995). Thus, eliminating items with limited statistical variance generally increases a scale’s overall sensitivity and specificity, and that in turn improves its effectiveness in measuring constructs of interest (McDowell & Newell, 1987; DeVellis).

Double-barreled items are those in which two different questions are contained in one item. Such items often confuse respondents because they are not sure as to which aspect of the double-barreled question they should respond (DeVellis, 1991; Streiner & Norman, 1995). Disaggregating double-barreled questions as per the third objective not only helps respondents in answering, it allows researchers to determine the extent to which each part of the original question helps to define a particular construct.

Plan and Procedures

Production of the LSNS-R progressed along a series of four analytical steps to address the objectives stated above. In each step, alpha reliability coefficients and the results of a series of principal component factor analyses were examined to determine whether and the extent to which items corresponded to the latent structural components of family networks and friendship networks.

The four steps are summarized in Figure 1. In the first step, reliability statistics were obtained for the original LSNS administered to the sample described above. These values then served as reference points for comparison with values obtained for subsequent modifications.

In the second analytical step, two items from the original LSNS scale – L9 (“Helps others with various tasks”) and L10 (“Living arrangements”) – were dropped because they demonstrated limited response variation among a number of sample groups including the present one. Furthermore, neither of these items helps to distinguish between family networks and friendship networks better than the other items in the original LSNS.

The L9 item was originally included in the LSNS in part because social exchange theory suggests that a reciprocal social relationship is stronger than one that is unidirectional (Jung, 1990; Burgess & Huston, 1979). Thus, rather than only capturing what others do for the older person being assessed, it is desirable to include items that also assess what the older person does for other people, i.e., items should be included to assess reciprocity of social support within kinship and friendship networks. Moreover, in past studies L9 generally demonstrated insufficient item variance and thus was a good candidate for elimination or replacement.
Refinements to the Lubben Social Network Scale:
The LSNS-R (continued)

The L10 item on living arrangements also had not worked out well over time. When the original LSNS was constructed, both living arrangements and marital status were common items included in measures of social support networks. It therefore seemed entirely appropriate to include an item merging these two related constructs. However, the L10 item has been the worst performing item on the LSNS across different settings. Part of the problem has been scoring it, which is constrained by the limited number of response options available as well as by disagreements among scorers in assigning ordinal weights to specific response options. Perhaps most important, marital status and living arrangements are generally not malleable nor appropriate for intervention, so items concerning them should not be included in any case.

In the third analytical step, two “double-barreled” questions were disaggregated. The two items, L3 (relatives) and L4 (friends) ask, “How many (relatives) (friends) do you feel close to? That is, how many of them do you feel at ease with, can talk to about private matters, or call on for help?” In this step, L3 and L4 were each recast as two distinct questions. One asks, “How many (family members) (friends) do you feel at ease with such that you can talk to them about private matters?” whereas the other asks, “How many (family members) (friends) do you feel close to such that you can call on them for help?” The first of these substitute questions examines somewhat intangible or expressed support, whereas the other taps into more tangible support, such as help with running an errand. Both types of support have been suggested as important aspects of social support networks (Litwak, 1985; Sauer & Coward, 1985).

In the fourth step, items that identify both the targets and sources of respondents’ confidant relationships were constructed and tested. Here the two confidant relationship items in the original LSNS (L7 and L8) were recast to distinguish between confidant relationships with family members and those with friends. These changes recognize that confidant relationships with family members may serve different functions than confidant relationships with friends (Keith, Hill, Goudy, & Power, 1984). The final result of the four-step process is a revised version of the LSNS, the LSNS-R.

Plan of Analysis

Step 1: Analyze original LSNS
Original LSNS Items:
- L1 Family: Number seen or heard from per month
- L2 Family: Frequency of contact with family member most in contact
- L3 Family: Number feel close to, talk about private matters, call on for help
- L4 Friends: Number feel close to, talk about private matters, call on for help
- L5 Friends: Number seen or heard from per month
- L6 Friends: Frequency of contact with friend most in contact
- L7 Confidant: Has someone to talk to when have important decision to make
- L8 Confidant: Others talk to respondent when they have important decision to make
- L9 Helps others
- L10 Living arrangements

Step 2: Eliminate items with limited variation
Items eliminated: L9 and L10

Step 3: Uncouple double-barreled questions
Items modified: L3 and L4 each split into two separate questions
- L3A Family: Number feel at ease with whom you can talk about private matters
- L3B Family: Number feel close to whom you can call on for help
- L4A Friends: Number feel at ease with whom you can talk about private matters
- L4B Friends: Number feel close to whom you can call on for help

Step 4: Distinguish between source and target confidant relationships with family and friends
Items modified: L7 and L8 each split into separate questions for family and friends
- L7A Family: Respondent functions as confidant to other family members
- L7B Friends: Respondent functions as confidant to friends
- L8A Family: Respondent has family confidant
- L8B Friends: Respondent has friend who is a confidant.

Data Source

The data are from a survey of older white, non-Hispanic Americans in Los Angeles County, California done between June and November 1993. A self-weighting, multistage probability
sample was selected from 861 census tracts in the area in which the white, non-Hispanic population exceeded any other single racial or ethnic group. This sampling strategy insured a high level of homogeneity in the sample.

The first three sampling stages were: (a) random selection of tracts, (b) random selection of blocks, and (c) random selection of households within selected blocks. Households were then contacted by telephone to determine the age and ethnicity of household members. All white, non-Hispanic persons aged 65 or over in each household were thus identified and then potential participants were randomly selected from this pool. Of the 265 older persons thus selected, 76 percent agreed to be interviewed, resulting in a final sample of 201. The sample included 130 women (65%) and 71 men (35%) and had a mean age of 75.3. Additional details on the sample are reported elsewhere (Villa, Wallace, Moon, & Lubben, 1997; Moon, Lubben, & Villa, 1998; Pourat, Lubben, Wallace, & Moon, 1999; Pourat, Lubben, Yu, & Wallace, 2000).

Results

Cronbach Alpha Values for the Products of Each Step

Table 1 presents internal consistency values for the products of the four analytical steps. The Cronbach alpha value for the original 10-item LSNS scale administered to the present sample ($\alpha = 0.66$) is slightly lower than those previously reported (Lubben, 1988; Lubben & Gironda, 1997) and below the desired standard for internal consistency. The Cronbach alpha values increased in each subsequent step in the analysis, indicating that each successive modification contributed to improving the final product’s internal consistency. Although the variant produced in Step 2 has two items less than the original LSNS, the alpha value is slightly higher than that for the original LSNS, suggesting that dropping items L9 and L10 was appropriate.

Further, the greatly improved alpha value obtained for the variant produced in Step 3 indicates that disaggregation of “double-barreled” questions (items L3 and L4) was quite beneficial. In Step 4, the product resulting from including items that distinguish between source and target confidant relationships with family members and those with friends had further increased internal consistency.

Table 1

<table>
<thead>
<tr>
<th>Step</th>
<th>Cronbach’s Alpha Value by Product of Each Step of Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>.66</td>
</tr>
<tr>
<td>2.</td>
<td>.67</td>
</tr>
<tr>
<td>3.</td>
<td>.73</td>
</tr>
<tr>
<td>4.</td>
<td>.78</td>
</tr>
</tbody>
</table>

Factor Analysis

Principal component factor analyses were performed for each step of the revision to explore for latent factors and to determine whether the final modified version has latent structural components corresponding to both kinship and friendship networks. The number of factors found in each step of the analysis was determined by considering factors with eigenvalues over one (Kaiser, 1960) and by identifying the elbow in the screenplot tests (Cattell, 1966). The factor loadings were subjected to varimax rotation.

Table 2 shows the rotated factor matrix for the original LSNS administered to the present sample. Although previous studies have reported three factors (Lubben, 1988; Lubben & Gironda, 1997), the rotated factor structure for the LSNS here showed a two-factor solution, with one factor consisting largely of family-related items and the other consisting primarily of items concerning friendships. However, the former, i.e., the family factor, also incorporates items concerning confidant relationships and the “helps others” items. Both confidant items clearly load onto the family factor in this step, but the “living arrangements” item cross-loads on both the family and friend factors.
Refinements to the Lubben Social Network Scale: The LSNS-R (continued)

Table 2

**Step 1: Original 10-Item LSNS Factor Matrix**

<table>
<thead>
<tr>
<th>Item</th>
<th>Family Factor</th>
<th>Friend Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3</td>
<td>.7298</td>
<td>.1242</td>
</tr>
<tr>
<td>L8</td>
<td>.6743</td>
<td>.0377</td>
</tr>
<tr>
<td>L1</td>
<td>.6633</td>
<td>-.0639</td>
</tr>
<tr>
<td>L2</td>
<td>.6629</td>
<td>-.0761</td>
</tr>
<tr>
<td>L9</td>
<td>.5702</td>
<td>-.0269</td>
</tr>
<tr>
<td>L7</td>
<td>.5391</td>
<td>.2047</td>
</tr>
<tr>
<td>L4</td>
<td>.2606</td>
<td>.7807</td>
</tr>
<tr>
<td>L5</td>
<td>.2072</td>
<td>.7520</td>
</tr>
<tr>
<td>L6</td>
<td>-.1082</td>
<td>.5225</td>
</tr>
<tr>
<td>L10</td>
<td>-.3490</td>
<td>.4461</td>
</tr>
</tbody>
</table>

Table 3

**Step 2: Factor Matrix after Eliminating Items L9 and L10**

<table>
<thead>
<tr>
<th>Item</th>
<th>Family Factor</th>
<th>Friend Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3</td>
<td>.7722</td>
<td>.1255</td>
</tr>
<tr>
<td>L2</td>
<td>.7393</td>
<td>-.1103</td>
</tr>
<tr>
<td>L1</td>
<td>.6807</td>
<td>-.0218</td>
</tr>
<tr>
<td>L8</td>
<td>.6435</td>
<td>.0850</td>
</tr>
<tr>
<td>L7</td>
<td>.5622</td>
<td>.2040</td>
</tr>
<tr>
<td>L4</td>
<td>.2105</td>
<td>.8162</td>
</tr>
<tr>
<td>L5</td>
<td>.1514</td>
<td>.7810</td>
</tr>
<tr>
<td>L6</td>
<td>-.1230</td>
<td>.5558</td>
</tr>
</tbody>
</table>

Table 4

**Step 3: Factor Matrix after Disaggregating Items L3 and L4**

<table>
<thead>
<tr>
<th>Item</th>
<th>Family Factor</th>
<th>Friend Factor</th>
<th>Confidant Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>.8133</td>
<td>.0029</td>
<td>.0604</td>
</tr>
<tr>
<td>L3A</td>
<td>.8020</td>
<td>.0959</td>
<td>.2442</td>
</tr>
<tr>
<td>L3B</td>
<td>.7775</td>
<td>.1226</td>
<td>.2543</td>
</tr>
<tr>
<td>L4B</td>
<td>.0522</td>
<td>.7984</td>
<td>.1707</td>
</tr>
<tr>
<td>L5</td>
<td>.1873</td>
<td>.7653</td>
<td>-.0319</td>
</tr>
<tr>
<td>L4A</td>
<td>.1632</td>
<td>.7504</td>
<td>.0523</td>
</tr>
<tr>
<td>L6</td>
<td>-.1817</td>
<td>.4877</td>
<td>.0685</td>
</tr>
<tr>
<td>L7</td>
<td>.0425</td>
<td>.1524</td>
<td>.8276</td>
</tr>
<tr>
<td>L8</td>
<td>.2365</td>
<td>.1647</td>
<td>.6689</td>
</tr>
<tr>
<td>L2</td>
<td>.4253</td>
<td>-.1524</td>
<td>.6166</td>
</tr>
</tbody>
</table>

In the second step, items L9 and L10 were eliminated as planned due to their general poor performance in previously discussed studies. Similar problems are demonstrated in the current study by the heavy cross-loading found for L10 in Step 1. As in Step 1, a two-factor structure was found (Table 3) and the "confidant" items were found to clearly load on the family factor. No heavy cross-loading was found for the scale variant produced in this step.

Table 4 shows the factor structure found in Step 3. In this step, the double-barreled items L3 and L4 ("talk about private matters" and "call on..."
Refinements to the Lubben Social Network Scale: The LSNS-R (continued)

Table 5
Step 4: LSNS-R Factor Matrix

<table>
<thead>
<tr>
<th>Item</th>
<th>Item description</th>
<th>Family Factor</th>
<th>Friend Factor A</th>
<th>Friend Factor B</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3B</td>
<td>Family: call on for help</td>
<td>.7600</td>
<td>-.0352</td>
<td>.2327</td>
</tr>
<tr>
<td>L8A</td>
<td>Family: has confidant</td>
<td>.7402</td>
<td>.0848</td>
<td>-.0024</td>
</tr>
<tr>
<td>L7A</td>
<td>Family: is confidant</td>
<td>.7358</td>
<td>.2490</td>
<td>-.0384</td>
</tr>
<tr>
<td>L3A</td>
<td>Family: discuss private matters</td>
<td>.7345</td>
<td>-.0292</td>
<td>.2320</td>
</tr>
<tr>
<td>L2</td>
<td>Family: frequency of contact with family member most in contact</td>
<td>.7134</td>
<td>.0514</td>
<td>-.1438</td>
</tr>
<tr>
<td>L1</td>
<td>Family: number in contact</td>
<td>.6712</td>
<td>-.1576</td>
<td>.1565</td>
</tr>
<tr>
<td>L8B</td>
<td>Friends: has confidant</td>
<td>.0487</td>
<td>.8800</td>
<td>.1279</td>
</tr>
<tr>
<td>L7B</td>
<td>Friends: is confidant</td>
<td>.1907</td>
<td>.8488</td>
<td>.1140</td>
</tr>
<tr>
<td>L6</td>
<td>Friends: frequency of contact with friend most in contact</td>
<td>-.1428</td>
<td>.5493</td>
<td>.1856</td>
</tr>
<tr>
<td>L5</td>
<td>Friends: number in contact</td>
<td>.0915</td>
<td>.0594</td>
<td>.8467</td>
</tr>
<tr>
<td>L4B</td>
<td>Friends: call on for help</td>
<td>.0839</td>
<td>.2711</td>
<td>.7663</td>
</tr>
<tr>
<td>L4A</td>
<td>Friends: discuss private matters</td>
<td>.0922</td>
<td>.4920</td>
<td>.6028</td>
</tr>
</tbody>
</table>

For help") were disaggregated. Principal component factor analysis performed here indicated a three-factor solution, with items loading on a family factor, a friendship factor, and a confidant factor. Generally the factors are clean (i.e., for each item, there is predominant loading on one factor and little loading on the others). However, the family "frequency of contact" item (L2) cross-loads on both the family and confidant factors.

Step 4 involved distinguishing family and friends as both possible sources and possible targets of confidant relationships, and resulted in the LSNS-R. Principal component factor analysis in this step (Table 5) revealed a single, clean family factor and two friendship factors. The friendship confidant items (L7B, L8B) and the frequency of contact with a friend item (L2) constitute one of the friendship factors, while the remaining friendship items make up a second friendship factor. The item on being able to talk to a friend about private matters (L4A) loads on both friendship factors.

Item-Total Scale Correlations

Item-total scale correlational analysis yielded coefficients ranging from 0.27 to 0.75, indicating that LSNS-R items are sufficiently homogeneous and without excessive redundancy. All internal reliability coefficients fell within the acceptable range suggested by Steiner and Norman (1995). The correlation coefficient between the original LSNS and LSNS-R was 0.68.

Conclusion

As gerontologists and geriatricians begin to identify means to increase active life expectancy rather than mere life expectancy, it is likely that older persons' social support networks will be shown to be necessary to healthy aging. This means that there is increasing need for a variety of reliable and valid social support scales for use in research and practice settings. The work discussed here should be viewed as part of an ongoing pursuit of such well-constructed social integration scales, for improved measures of social support networks are essential to understanding better the reported link between social integration and health. Such improved knowledge thus will enhance future gerontological research, geriatric care, and the quality of life of the elderly.

From applied research and clinical perspectives, there is growing pressure to develop short and efficient scales. Some elderly populations are unable to complete long questionnaires, and time constraints in most clinical practice settings necessitate use of efficient and effective screening tools. Shorter scales require less time and energy of both the administrator and respondent. Thus, parsimonious and effective screening tools are needed that are acceptable to elders, researchers, and health care providers as well.
Refinements to the Lubben Social Network Scale: The LSNS-R (continued)

For more basic researchers, somewhat longer research instruments are desirable, because having a larger number of items as well as better clarity of concepts generally contributes to a scale’s reliability and facilitate the analysis and appraisal of subtle differences. But rather than attempting to design a single social support network scale for use with all elderly populations and for all research purposes, it seems far more practical to design measurement instruments for specific populations along with clear indications of how they should be used (Mitchell & Trickett, 1980). Use of such well-targeted social support network assessment instruments will yield more valid research results than use of less well-targeted measurement tools.

The original LSNS was designed specifically for an elderly population. Although it has proven adaptable to a variety of settings, some deficiencies have been noted over the past ten years of use. The LSNS-R addresses these problems, resulting in an improved measure of social support networks. The refinements are theory driven and involved reworking items in the original LSNS so that the revised scale can better measure the distinct aspects of kinship and friendship networks (Lubben & Gironda, in press). An abbreviated version of the LSNS-R has been developed and is reported elsewhere (Lubben & Gironda, 2000). This six-item scale (the LSNS-6) can be especially suitable in practice settings as a screening tool for social isolation or for more general use in those research settings where longer social support network scales cannot be accommodated. For those social and behavioral researchers desiring more extensive inquiry into the nature of social relationships of the elderly, an 18-item version of the LSNS has also been developed (Pourat, et al., 1999; Pourat, et al., 2000). The major advantage of the LSNS-18 over the LSNS-R is that the former distinguishes friendship ties with neighbors from those with friends who do not live in close proximity to the respondent. Such distinctions are desirable for exploring a growing number of social and behavioral research questions regarding the functioning of social support networks.

In summary, development and validation of social support network instruments are cumulative and ongoing processes. They require testing and retesting with diverse populations, in various research and practice settings, and using both psychometric and practical standards to assess their actual utility. In addition, future analyses of these scales should include assessment of their sensitivity to various differences within and between groups, for example, cultural and socio-demographic differences, or differences in levels of health and functional status that might affect response patterns.

The four-step process described above continues in the tradition of instrumentation refinement, tailoring new measurement tools to address special needs as well as offering a revised version of the LSNS incorporating modifications that greatly enhance its psychometric properties. This work also offers a paradigm that could be employed to evaluate and refine other social support network measurement tools.

References


Refinements to the Lubben Social Network Scale: The LSNS-R (continued)


Refinements to the Lubben Social Network Scale: The LSNS-R (continued)


James Lubben, DSW, MPH is Professor of Social Welfare and Urban Planning at the University of California, Los Angeles (UCLA). Both his DSW and MPH are from the University of California, Berkeley. He is Principal Investigator for the Hartford Doctoral Fellows Program in Geriatric Social Work, a program administered by the Gerontological Society of America, and a consultant to the World Health Organization-Kobe Centre on health and welfare systems development for aging societies. Dr. Lubben’s research examines social behavioral determinants of vitality in old age, with a particular focus on the roles of social support networks.

Melanie Gironda, PhD, MSW is a Lecturer in the Department of Social Welfare at UCLA where she teaches courses on social gerontology and research on aging. She received both her PhD and MSW from UCLA. Dr. Gironda is Deputy Program Director of the Hartford Doctoral Fellows Program in Geriatric Social Work. Her research examines loneliness in various populations of the elderly, with a special focus on the nature of social support networks of older adults without children.

Alex E. Y. Lee, PhD, MSW is Assistant Professor in the Department of Social Work and Psychology at the National University of Singapore where he teaches courses on social work and gerontology. He received both his PhD and MSW from UCLA. His current research concerns social service delivery systems for the elderly and family-based gerontological counseling for Asian elders.

In small proportions, we just beauties see, And in short measures life may perfect be.

Ben Jonson
Self-Estimated Intelligence

Adrian Furnham

Do beliefs about one’s ability affect scores on an ability test? Is it important to believe that you are bright to do well on an ability/intelligence test? What do people think about their own intelligence? How do they estimate their own intelligence and that of their relatives? What determines their self-estimates?

Studies of self-estimated intelligence date back nearly a quarter of a century (Hogan, 1978). More recently Beloff (1992) provoked a great deal of interest in the effects of gender differences on self-estimated intelligence. Among her Scottish undergraduates, she found a six-point difference, with males estimating their score significantly higher than females. Additional work led her to conclude that “The younger women see themselves as intellectually inferior compared to young men....Women see equality with their mothers, men with their fathers and men superior to their mothers. Mothers therefore come out as inferior to fathers. This pattern has been consistent each year.” Beloff argued that the modesty training girls receive in socialization accounts for these data.

Beloff’s research stimulated others to replicate these gender-related differences in self-estimated intelligence (Bennett, 1996, 1997, 2000; Byrd & Stacey, 1993; Furnham & Rawles, 1995, 1999). These studies were in four related research areas: (a) studies to replicate and explain these differences in various countries and relate national/cultural variables to them; (b) research to see if these gender-related differences can be replicated not only for overall (general) intelligence, but also for more specific and multiple intelligences; (c) work to examine whether these differences occur in estimations of the intelligence of others, notably male and female members of one’s immediate family.

Table 1
Results of Studies Where Participants Made an Overall IQ (g) Rating on Themselves and Others.

<table>
<thead>
<tr>
<th>Study</th>
<th>Women</th>
<th>Men</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beloff (1992) – Scotland (N = 502)</td>
<td>Self 120.5</td>
<td>Father 127.7</td>
<td>7.2</td>
</tr>
<tr>
<td></td>
<td>Mother 119.9</td>
<td></td>
<td>-0.8</td>
</tr>
<tr>
<td>Byrd &amp; Stacey (1993) – New Zealand (N = 105)</td>
<td>Self 121.9</td>
<td>Mother 114.5</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td>Father 127.9</td>
<td></td>
<td>-5.6</td>
</tr>
<tr>
<td>Reilly &amp; Mulhern (1995) – Ireland (N = 80)</td>
<td>Self 105.3</td>
<td>Father 114.1</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Measured 106.9</td>
<td></td>
<td>-0.8</td>
</tr>
<tr>
<td></td>
<td>Father 114.18</td>
<td></td>
<td>3.9</td>
</tr>
<tr>
<td>Furnham &amp; Rawles (1996) – England (N = 140)</td>
<td>Self 116.64</td>
<td>Male child 1st child 103.84</td>
<td>4.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Female child 1st child 102.57</td>
<td>-0.21</td>
</tr>
<tr>
<td>Furnham &amp; Gasson (1998) – England (N = 112)</td>
<td>Self 103.84</td>
<td>Male child 1st son 116.69</td>
<td>5.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Female child 1st son 101.82</td>
<td>-1.77</td>
</tr>
<tr>
<td>Furnham &amp; Reeves (1998) – England (N = 84)</td>
<td>Self 104.84</td>
<td></td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male child 1st son 116.69</td>
<td>5.31</td>
</tr>
</tbody>
</table>
Self-Estimated Intelligence (continued)

(grandparents, parents, siblings, children); (d) research to examine the relationship between estimated intelligence and “actual” intelligence scores. Furnham (2001) reported on various such studies, all but one of which showed significant gender differences in self-rated overall IQ ranging from 3.9 to 8.6 points.

Beginning with Beloff (1992), a number of these studies looked at estimates of relations’ IQs (Furnham, Fong, & Martin, 1999). These studies showed that people believe there are clear generational effects in IQ – they believe that they are a little brighter than their mothers and certainly much brighter than their grandparents, while parents tend to believe their children are brighter than they are. In general, a half standard deviation (6-8 points) difference was found in estimated IQ between generations. Gender differences were also found for estimated IQ of relatives (Flynn, 1987). Thus people believe their grandfathers are brighter than their grandmothers, their fathers brighter than their mothers, their brothers brighter than their sisters, and their sons brighter than their daughters. Interestingly for parents estimating their children, the results were stronger for first-borns compared to those born later, indicating the possible working of the principle of primogeniture.

These results have been shown to be cross-culturally robust as the gender difference effect has been demonstrated in Asia (Japan, Hong Kong), Africa (Uganda, South Africa), Europe (Belgium, Britain, but not Slovakia) and America (Furnham, et al., 2002) as well as New Zealand (Furnham & Ward, 2001) and Iran (Furnham, Shahidi, & Baluch, 2002). Thus there seems to be a robust and cross-culturally valid finding that there is a clear, consistent gender difference in self-rating of overall intelligence, with males rating themselves and their male relations higher than females rate themselves and their female relations.

It is also interesting to note that it is very rare for people to rate their scores, or indeed those of relations, as below average (<100 IQ points). For example, it was found that students tended to believe their IQ to be about one-and-a-half standard deviations above the mean (Furnham, Clark, & Bailey, 1999) – around 120, whereas nonstudent British adults believed their IQ is to be around a half standard deviation above the norm (Furnham, 2000).

Multiple Intelligence

Over the past decade there have been many attempts to redefine intelligence and define different types of intelligence. Thus we have emotional intelligence as well as practical intelligence, for example, but perhaps the idea that has appealed most to lay people (not academics, however) is Gardner’s (1999) multiple intelligence theory. Gardner (1983) initially argued that there were seven types of intelligence:

- Verbal or linguistic intelligence (the ability to use words),
- Logical or mathematical intelligence (the ability to reason logically, to solve numerical problems),
- Spatial intelligence (the ability to find one's way around one’s environment and form mental images),
- Musical intelligence (the ability to perceive and produce pitch and rhythm),
- Body-kinetic intelligence (the ability to carry out motor movements, for example, in performing surgery, ballet dancing, or playing sports),
- Interpersonal intelligence (the ability to understand people),
- Intrapersonal intelligence (the ability to understand oneself and develop a sense of one’s identity).

He suggested that the first two are those types valued at school, the next three are valuable in the arts, and the latter two constitute a sort of “emotional intelligence.”

At least eight studies have examined gender differences in estimates of Gardner’s (1983, 1999) seven multiple intelligences. The results have been consistent and give an important clue into gender differences in overall performance. Most people rate their own interpersonal and intrapersonal intelligence as very high, about 1 SD (standard deviation) above the mean, and their musical and body kinetic intelligence as strictly average, around 100. Of the remaining three types of intelligence in the Gardner model
Self-Estimated Intelligence (continued)

- verbal, mathematical and spatial, gender differences were found only in mathematical and spatial intelligence, particularly in spatial intelligence where females typically rate themselves six to 10 points lower than do males. No gender differences were found in self-estimations of verbal intelligence (Furnham, 2001).

Some studies have asked participants to estimate their overall (g, general intelligence) score first and then their scores on the seven specific multiple intelligences. This has made it possible to regress simultaneously the seven multiple intelligence estimates on the overall intelligence estimate. These regression studies show that people believe that mathematical (logical), then spatial, and then verbal intelligence are the primary contributors to overall IQ. This, in turn, allowed for testing the hypothesis that lay conceptions of intelligence are male normative in the sense that those abilities that men tend to be better at are those that most people consider to be the essence of intelligence. These studies showed that lay people tend to confuse mathematical/spatial and overall intelligence, thereby explaining the consistent gender differences in estimates of overall score. This means that quite possibly the often-observed and debated gender difference in spatial IQ accounts for the difference in overall IQ.

Correlations Between Self-Estimated and Test-Measured IQ

Are self-estimates accurate? In other words, is the correlation between (valid) IQ test scores and self-estimated scores high, low, or “on the mark”? This is an important question because some psychologists have suggested that if these scores correlate highly, self-estimates may serve as useful measures at a fraction of time, effort, and costs of administering and scoring IQ tests.

Various studies have found that correlations of actual and self-estimated IQ are around $r = 0.30$ and that therefore self-estimates cannot serve as proxy measures of IQ (Paulus, Lysy, & Yik, 1998). One study, however, examined the effect of outliers and concluded that, if these outliers are removed, self-estimated and actual IQ correlate highly ($r > 0.90$).

Reilly and Mulhern (1995) note that IQ-estimates research should not be based on the “assumption that gender differences at group level represent a generalized tendency on the part of either gender to either overconfidence or lack of confidence with regard to their own intelligence” (Reilly & Mulhern). Studies do show, however, a tendency for males to overestimate and females to underestimate their score [reference(s)] but this is related, in part, to the IQ test used.

Some researchers have tried to understand and then increase the correlation between self-estimates and test scores by using more tests on bigger populations, yet the correlations remained the same, as noted above, around $r = 0.30$ (Borkenau & Liebler, 1993). Perhaps the explanation should be sought in motivational factors that may be involved in the self-estimation of intelligence and that may lead to serious distortions in estimated scores. Thus a close examination of the conditions and instructions under which participants self-estimate their intelligence may provide a clue as to how they make their self-estimates. For example, if social norms and conventions in part dictate how people respond, then anonymity in responding may reduce the distortions in self-estimates.

From Whence the Differences?

Gender differences in self-estimated IQ need to be explained because the issue is so frequently discussed, there has been an academic consensus for many years to the effect that gender differences are minimal, and since before the Second World War, test constructors have been careful to produce IQ tests that minimize gender differences.

Essentially there are three positions on the gender differences of estimates issue (Furnham, 2001). The “feminist” position was clearly articulated by Beloff (1992) in her first study, where she suggested that these differences are erroneous and simply due to attribution errors. One explanation she offers is a gender difference in modesty and humility. “Modesty-training is given to girls. Modesty and humility are likely to be connected to the over estimates of women and for women” (Beloff). Another explanation
she proposes is that because IQ is correlated with occupational grade and men tend to occupy certain higher status occupations more than women (for political rather than ability reasons in her view), females tend to believe that they are less intelligent than males. There seem to be no direct data to test Beloff’s assumptions, and they look a little outdated, particularly in the case of America where students still show the same gender difference pattern despite their more equal socialization and occupational choice. However, unusual data from Slovakia may be evidence for this position. Slovakian females awarded themselves higher overall (g) and verbal scores than did equivalent Belgian and British student groups (Furnham, Rakow, Samany-Schuller, & De Fruyt, 1999). The authors offered the following possible explanation for this unique group of confident females: It may well be that under the pressure of socialist governments of Eastern Europe, the role of females in society was somewhat different from that of women in capitalist Western Europe – the former took a more active role in the economy and were socialized differently in school. In fact, Slovakian society attached high prestige to education and the government made a consistent effort to improve the position of women in society by encouraging them to obtain educational qualifications, facilitating their employment in traditionally male-dominated occupations, and even mandating a given percentage of women in its parliament. Another explanation they present is that among the various nationalities studied, the Slovakian women had the most experience in taking intelligence tests and were therefore presumably more likely to recognize that gender differences in intelligence are actually very small (Furnham, et al.).

Consequences of Gender Differences in Ability Estimates

In a series of studies, Beyer (1990) demonstrated gender differences in self-expectations, self-evaluations, and performance on ability-related tasks. Her results support the “male hubris, female humility” thesis of Beloff. Further, she argued that gender differences in self-evaluations affect expectancies of success and failure, and, ultimately, performance on ability-related tasks. Thus the importance of studies of self-estimated intelligence may lie not only in exploring lay theories of intelligence, but also in understanding the self-fulfilling nature of self-evaluation of ability. “Because of the serious implications of under-estimations for self-confidence and psychological health more attention should be devoted to the investigation of gender differences in the accuracy of self-evaluations. Such research will not only elucidate the underlying processes of self-evaluation biases and therefore be of theoretical interest but will also be of practical value by suggesting ways of eliminating women’s under-estimation of their performance” (Beyer).

Beyer (1998) reviewed studies concluding that individuals make poor self-evaluators. In one such study, the correlation between medical students’ self-rated knowledge and exam grades was found to be almost exactly zero (r = -0.01). In another study Beyer reviewed, the correlation between self-perceptions of one’s own physical attractiveness and judges’ ratings of their attractiveness was 0.22. Another study she considered found a correlation of 0.32 between how one judged ones own performance on a test of managerial-skill and how independent experts rated it. “Interestingly, outside evaluators seem to be better assessors of a target’s performance that the target her/himself” (Beyer). Thus it seems that while self-estimates of intelligence may not be useful as proxy IQ tests, various studies (Paulus et al., 1998) have shown that they could be very useful in explaining some of the variability in actual test results through the self-expectation effect that Beyer (1999) discussed.

The work of Dweck and colleagues is particularly salient here (Dweck & Bempechat, 1983; Mueller & Dweck, 1998). Dweck argued that lay persons tend either to believe that intelligence is malleable (incremental lay theorists) or else it is fixed (entity lay theorists). These beliefs logically relate to goal setting, motivation, and attribution for success and failure. Further, they relate to one’s perceptions of others. While these beliefs are not necessarily related to actual ability, they can have powerful, even paradoxical, behavioral impacts. Thus Mueller and Dweck pointed out if a young
person believes him/herself to be intelligent, entity theory predicts that they are likely to be less motivated to work hard to achieve their goals.

The interaction between beliefs about one’s level of intelligence on the one hand and the changeability of intelligence on the other is potentially very important. Thus entity lay theorists who believe themselves to be below average on overall or specific intelligences may shun intellectual tasks, set themselves low goals, and become depressed and ineffectual. In contrast, entity lay theorists who rate themselves way above average may be complacent and lazy, believing their “natural wit” sufficient to let them succeed at most academic and work assignments. However, lay incrementalists who are low self-raters may, if so moved, be prepared to undertake tasks or training that they believe will increase their intelligence. High self-rating incrementalists, in contrast, may also believe that they need to work hard to maintain their levels of intelligence.

Conclusion

The measurement of intelligence has always been a controversial topic. So, too, have the topics of lay people’s beliefs about their own intelligence and the extent to which it can be changed. Research relating to these topics has yielded evidence of a “self-Pygmalion effect” in that beliefs about the nature of intelligence and one’s intelligence level can profoundly influence not only how one performs on a test, but one’s personal motivations in academic and work settings throughout life (Spitz, 1999).

References


Self-Estimated Intelligence (continued)


Dr. Adrian Furnham is Professor of Psychology at London University (UCL) where he has taught for 20 years. He holds doctoral degrees from both Oxford University and UCL. He is the author of 36 books and 450 peer-reviewed papers. Dr. Furnham is currently working on a book about managerial incompetence.

Every tool carries with it the spirit by which it has been created.

Werner Karl Heisenberg

Dr. Furnham is currently working on a book about managerial incompetence.

HaPI Advisory Board

Aaron T. Beck, MD
University of Pennsylvania School of Medicine

Timothy C. Brock, PhD
Ohio State University, Psychology

William C. Byham, PhD
Development Dimensions International

Donald Egolf, PhD
University of Pittsburgh, Communication

Sandra J. Frawley, PhD
Yale University School of Medicine, Medical Informatics

David F. Gillespie, PhD
Washington University
George Warren Brown School of Social Work

Robert C. Like, MD, MS
University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School

Joseph D. Matarazzo, PhD
Oregon Health Sciences University

Vickie M. Mays, PhD
University of California at Los Angeles, Psychology

Michael S. Pallak, PhD
Behavioral Health Foundation

Kay Pool, President
Pool, Heller & Milne, Inc.

Ora Lea Strickland, PhD, RN, FAAN
Emory University Woodruff School of Nursing

Michael Zaltman, PhD
Harvard University Graduate School of Business Administration

Stephen J. Zyzanski, PhD
Case Western Reserve University School of Medicine

Vol. 7, No. 2, Winter 2002
Ten Commandments for Selecting Self-Report Instruments

Fred B. Bryant

Selecting appropriate measurement instruments is among the tasks researchers most frequently face. Yet, surprisingly little has been written about how best to go about the process of instrument selection. Given the prevalence of self-report methods of measurement in the social sciences, the task of selecting an instrument most often involves choosing from among a set of seemingly relevant questionnaires, surveys, inventories, checklists, and scales. For example, a researcher who wishes to measure depression in college students might locate dozens of potentially useful self-report instruments designed to assess this construct. Indeed, the September 2001 release of the Health and Psychosocial Instruments (HaPI) database contains 105 primary records of self-report instruments with the term “depression” in the title, excluding measures developed for use with children or the elderly, those written in foreign languages, and those assessing attitudes toward, knowledge of, or reactions to depression rather than depression per se. The seemingly appropriate instruments thus identified include the Beck Depression Inventory (BDI; Beck, 1987), Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960), Center for Epidemiologic Studies Depression Scale (CES-D; National Institute of Mental Health, 1977), and Self-Rating Depression Scale (SDS; Zung, 1965), to name just a few. How should the researcher decide which to use?

One strategy for selecting instruments is to employ only those most commonly used in published studies. Not only is this strategy simple and straightforward, but some researchers follow it in the hope of increasing the likelihood that their research will be published. However, it limits conceptual definitions to those created within the theoretical frameworks of commonly used instruments. Over time, this effectively constricts the generalizability of research on these constructs. Further, all measurement instruments tap irrelevancies that have nothing to do with the constructs they are intended to assess. Using only a single measure of a particular research construct makes it impossible to know the degree to which the irrelevancies in the measure affect the obtained results. Moreover, diversity in operationalization helps us better understand not only what we’re measuring, but also what we’re not measuring. Thus, in the long run, employing only the most commonly used instruments limits and weakens the body of scientific knowledge.

Although it is difficult to devise a universally applicable set of rules for selecting measurement instruments, it is possible to suggest some general guidelines that researchers can use in choosing appropriate self-report measures. What follows, then, is a set of precepts and principles for selecting instruments for research purposes, along with concrete examples illustrating each. Note that the order of presentation is not intended to reflect the guidelines’ relative importance in the process of instrument selection. In fact, each principle is essential in selecting the right measurement tool for the job.

1. Before choosing an instrument, define the construct you want to measure as precisely as feasible. Ideally, researchers should not rely merely on a label or descriptive term to represent the construct they wish to assess, but should instead define the construct of interest clearly and precisely in theory-relevant terms at the outset. Being unable to specify beforehand what it is you want to measure makes it hard to know whether or not a particular instrument is an appropriate measure of the target construct (Bryant, 2000). Potentially useful strategies for defining research constructs are to draw on published literature reviews that synthesize available theories concerning a particular construct, or to review the published literature on one’s own in search of alternative theoretical definitions. If you can find alternative conceptual definitions of the target construct, then you can choose from among them a particular conception that resonates with your own thinking. The process of explicitly conceptualizing the construct that you wish to measure is known as “preoperational explication” (Cook & Campbell, 1979).

Imagine, for example, that a clinical researcher wants to use a self-report measure of shyness. A
Ten Commandments for Selecting Self-Report Instruments (continued)

wide variety of potentially relevant measures can be found in the Health and Psychosocial Instruments database, depending on how the researcher conceptually defines shyness. Is shyness (for which there are 30 primary records in the database) synonymous with introversion (30 primary records), timidity (17 primary records), emotional insecurity (2 primary records), social anxiety (63 primary records), social fear (1 primary record), social phobia (29 primary records), social avoidance (4 primary records), or social isolation (76 primary records)? The clearer and more precise the initial conceptual definition, the easier it will be to find appropriate measurement tools. An added benefit of precisely specifying target constructs at the outset is that it helps to focus the research.

Although precise preoperational explication is the ideal when selecting measures, in actual practice it is often difficult beforehand to specify a clear conceptual definition of the target construct. Many times the published literature does not provide explicit alternatives and this, then, forces researchers to explicate constructs on their own -- the equivalent of trying to define an unknown word without having a dictionary. In actual practice, researchers often begin by selecting a particular instrument that appears useful, thus adopting by default the conceptual definition of the target construct that underlies the chosen instrument. Truly, therefore, an available instrument often dictates one's conceptual definition post hoc.

2. Choose the instrument that is designed to tap an underlying construct whose conceptual definition most closely matches your own. Carefully consider the theoretical framework on which the originators based their instrument. Select an instrument that stems from a theory that defines the construct the same way you do, or at least in a way that does not contradict your conceptual orientation, for the theoretical underpinnings of the instrument should be compatible with your own conceptual framework.

Consider a sociologist and a psychologist, each of whom wants to measure guilt. The most appropriate self-report instrument in each case will be the one whose underlying conceptual definition most closely corresponds to that of the researcher. The sociologist, on the one hand, might be studying people's reactions to homeless adults from a macro-level, sociocultural perspective. If so, then she might begin by defining guilt to be a prosocial emotion experienced when one perceives oneself as being better off than another person who is disadvantaged. Consistent with this conceptualization is Montada and Schneider's (1989) three-item measure of "existential guilt," conceived as prosocial feelings in reaction to the socially disadvantaged. The psychologist, on the other hand, might be studying personality from a micro-level, individual perspective. If so, then she might begin by defining guilt to be a dispositional feeling of regret or culpability in reaction to perceived personal or moral transgressions. Consistent with this conceptual definition is Kugler and Jones' (1992) 98-item Guilt Inventory, which includes a separate subscale specifically tapping personal guilt as an unpleasant emotional trait. Clearly, instruments must have underlying conceptual definitions that match your own conceptual framework (Brockway & Bryant, 1998).

3. Never decide to use a particular instrument based solely on its title. Just because the name of an instrument includes the target construct does not guarantee that it either defines this construct the same way you do or even measures the construct at all. Don't let the title lead you to select an inappropriate instrument.

As a case in point, consider a developmental psychologist who wants to measure parents' psychological involvement in their families. Based on its promising title, the researcher decides to use the Parent/Family Involvement Index (PFII; Cone, DeLawyer, & Wolfe, 1985). After obtaining the instrument and inspecting its constituent items, the researcher realizes to his chagrin that the PFII requires a knowledgeable informant (e.g., a teacher or teacher's aide) to indicate whether or not the particular parent of a school-aged child has engaged in each of 63 specific educational activities. Based on an underlying conceptualization of family involvement in psychological terms, a more appropriate instrument would be a measure
developed by Yogev and Brett (1985) that assesses parental involvement in terms of the degree to which parents identify psychologically with and are committed to family roles. This example shows clearly that you can’t judge an instrument by its title any more than you can judge a book by its cover (Brockway & Bryant, 1997).

4. **Choose an instrument for which there is evidence of reliability and validity.** Reliability is measurement that is accurate and consistent. Good reliability in measurement strengthens observed statistical relationships -- the more reliable the instrument, the smaller will be the error in measurement, and the closer observed results will be to actual results. For example, imagine a medical researcher who wants to determine whether an experimental antipyretic agent reduces fever more rapidly than available antipyretics, but who is using an unreliable thermometer that gives different readings over time, even when body temperature is stable. These inconsistencies in measurement make it nearly impossible to assess temperature accurately, and greatly decrease the likelihood of finding any experimental effects.

Data supporting the validity of an instrument increase one’s confidence that it really measures what it is designed to measure. For example, the medical researcher referred to above can be more confident that the thermometer actually measures temperature if its readings correlate highly with those of infrared-telemetry or body-scanning devices. Although instrument developers sometimes report reliability and validity data, such empirical evidence is often available only in published studies that have used the given measure. As a rule, avoid judging the validity of an instrument by the content of its constituent items. What an instrument appears to measure “on its face” (i.e., face validity) is not necessarily what it actually measures. As in the case of an instrument’s title, what you see is not necessarily what you get.

Judging the quality of research evidence concerning measurement reliability and validity can be difficult and confusing. There are various types of reliability (e.g., internal consistency, split-half, parallel-forms, interrater, test-retest) and of reliability coefficients (e.g., Cronbach’s alpha, coefficient kappa, intraclass correlation, KR-20). Similarly, there are numerous types of validity (e.g., construct, concurrent, criterion-referenced, convergent, discriminant). Thus a host of specialized statistical tools has been developed to quantify both reliability (Strube, 2000) and validity (Bryant, 2000). Given the numerous types of reliability and reliability coefficients, validity, and tools used to assess reliability and validity, instrument selection requires at least a basic understanding of psychometrics and of principles of scale validation in order to make informed judgments of instrument quality. When there is no evidence concerning an instrument’s reliability or validity, measurement becomes a “shot in the dark.”

5. **Given a choice among alternatives, select an instrument whose breadth of coverage matches your own conceptual scope.** If you define your target construct as having a wide range of requisite constituent components, then choose an instrument whose items tap a broad spectrum of content relating to those components. On the other hand, if you define your target construct in a way that specifies a narrower set of conceptual components, then choose an instrument that has a more restrictive and specific content.

Breadth of coverage varies widely across available instruments. For example, to measure coronary-prone Type A behavior, alternatives include: the Student Jenkins Activity Survey (Glass, 1977) that taps the Type A components of hard-driving competitiveness and speed-impatience; the Type A Self-Rating Inventory (Blumenthal, Herman, O'Toole, Haney, Williams, & Barefoot, 1985) that taps hard-drivingness and extraversion; the Type A/Type B Questionnaire (Eysenck & Fulker, 1983) that taps tension, ambition, and activity; the Time Urgency and Perpetual Activation Scale (Wright, 1988) that taps activity and time urgency; and the Self-Assessment Questionnaire (Thorensen, Friedman, Powell, Gill, & Ulmer, 1985) that taps hostility, time urgency, and competitiveness. Clearly, your choice of
Ten Commandments for Selecting Self-Report Instruments (continued)

instrument depends on the specific components of Type A behavior that you want to investigate.

The choice between general versus specific measures of a given construct may also depend on your particular research question. Consider the process of coping, for example. Numerous instruments exist for measuring people's general style of coping in response to stress. However, if you want to study coping in relation to a specific problem or stressor, a host of other measures have been developed to assess individual differences in coping with such specific concerns as arthritis, asthma, cancer, chronic pain, diabetes, heart disease, hypertension, stroke, multiple sclerosis, spinal cord injury, HIV/AIDS, rape, sexual abuse, sexual harassment, pregnancy, miscarriage, childbirth, economic pressure, job stress, unemployment, depression, bereavement, natural disasters, prison confinement, test anxiety, and war trauma, to name just a few. Compared to a broad-band measure of general coping, a narrow-band measure of coping that is specific to a particular stressor would be expected to show stronger relationships with reactions to that specific stressor.

6. Select an instrument that provides a level of measurement appropriate to your research goals. Some instruments are based on a theoretical model in which the underlying construct is assumed to be "unitary." Such instruments provide only a general, global "total score" that summarizes the overall level of responses. Other instruments are based on a theoretical framework in which the latent construct is considered to be "multidimensional." Such instruments provide multiple "subscale scores," each of which taps a separate dimension of the underlying construct. Thus, if you want to gather global summary information about a target construct, then use a unitary instrument. If you want to gather information about multiple facets of a target construct, then use a multidimensional instrument.

Imagine two nursing researchers, each of whom wishes to measure patients' life satisfaction. One seeks a global summary of patients' overall life satisfaction, whereas the other wants to compare levels of satisfaction across important aspects of patients' lives. The first researcher could use the five-item Satisfaction with Life Scale (Diener, Emmons, Larsen, & Griffin, 1985) to obtain a global total score. The second researcher could use the 66-item Quality of Life Index (Ferrans & Powell, 1985) to obtain individual scores for four separate satisfaction subscales: Health/Functioning, Socioeconomic, Psychological/Spiritual, and Family.

7. Choose an instrument with a time frame and response format that meet your needs. Don't use a "trait" measure (i.e., an instrument that defines the underlying construct as a stable disposition) to assess a transient, situational "state" that you expect to change over time. And don't modify the labels on a response scale (e.g., from "rarely" to "never") or the time frame of measurement (e.g., from "in general" or "on average" to "during the past day" or "in the past hour") unless you have no other recourse. Substantial changes in an instrument's response format or time frame can compromise its construct validity and therefore require revalidation.

In measuring hostility, for example, the choice of an appropriate instrument would depend on whether you conceive of hostility as a transitory, variable state or a stable, dispositional trait. An appropriate tool for measuring "state" hostility might be the 35-item Current Mood Scale (Anderson, Deuser, & DeNeve, 1995), which is designed to assess situational hostility, whereas an appropriate tool for measuring "trait" hostility might be the 50-item Cook-Medley Hostility Scale (Cook & Medley, 1954), which is based on the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1989) and conceptualizes hostility as a personality trait. Clearly, when selecting instruments you need to distinguish carefully between states and traits.

8. Match the reading level required to understand the items in the instruments you select to the age of the intended respondents. In studying children or adolescents, for example, avoid using an instrument designed for use with adults. When in doubt, use word-processing or
linguistic software to determine the reading ability level required to understand an instrument’s constituent items.

Imagine a researcher who wants to measure depression in children. Depending on the average age of the subjects, the researcher could choose from a variety of different self-report instruments specifically designed to assess depression in children of various ages, including those with a first-grade reading level (Children’s Depression Inventory; Kovacs, 1985), children age 7-13 (Negative Affect Self-Statement Questionnaire; Ronan, Kendall, & Rowe, 1994), children age 8-12 (Childhood Depression Assessment Tool; Brady, Nelms, Albright, & Murphy, 1984), or children age 8-13 (Depression Self-Rating Scale; Asarnow, Carlson, & Guthrie, 1987). In any case, the researcher studying childhood depression should avoid adopting an instrument designed to tap depression in adults.

9. Never use an instrument unless you know how you’ll score it and how you’ll analyze it.

This rule may seem self-evident, but it is sometimes violated unintentionally. No matter how interesting or important an instrument seems, it is useless unless you can convert responses to it into meaningful data. Sometimes the scoring rules are difficult to obtain or hard to follow, particularly when an instrument consists of multiple composite subscales, reverse-scored items, or item-specific scoring weights. This suggests that researchers should first make sure they know how to score an instrument before they administer it.

Consider the SF-12 (Ware, Kosinski, & Keller, 1996), a 12-item self-report instrument designed to measure functional health status. At first glance, it might appear simple enough to score this instrument by simply summing or averaging responses to its constituent items. But the test manual for the SF-12 (Ware, 1993) specifies a complex set of mathematical computations designed to weight and combine the 12 item scores to produce composite scores reflecting mental, physical, and total functioning. Users cannot score the SF-12 correctly unless they have access to the detailed scoring instructions contained in the test manual. Clearly, then, administering an instrument is one thing, but scoring it can be an entirely different matter.

10. Rather than choosing only one measure, when feasible use multiple measures of the construct you wish to assess.

A central tenet of classical measurement theory is that any single way of measuring a construct has unavoidable idiosyncrasies that are unique to the measure and have nothing to do with the underlying conceptual variable. By studying what multiple measures of the same construct have in common, researchers can converge or triangulate on the referent construct. Using multiple measures also allows an assessment of the generalizability of results across alternative operational or conceptual definitions, to probe the generality versus specificity of effects. And in the long run, using multiple instruments will advance our understanding of the targeted construct much further than simply using a single instrument.

Even when following the ten guidelines for instrument selection discussed above, you will still sometimes face difficult, highly subjective decisions in choosing appropriate measures. For example, which mood measure is more appropriate: one that uses a seven-point response scale, or one that uses a four-point response scale; one that includes a specific label for each individual point on its response scale, or one that has labels only at its endpoints; one that assesses the frequency with which respondents experience certain feelings, or one that taps the percentage of time respondents experience certain feelings? Given the subjectivity of such decisions, it makes all the more sense to use multiple instruments whenever possible so as to evaluate the generalizability of results across alternative operational definitions of the same underlying construct.

To recap, I have suggested “Ten Commandments” for selecting self-report instruments:

1. Before choosing an instrument, define the construct you want to measure as precisely as feasible.
Ten Commandments for Selecting Self-Report Instruments (continued)

2. Choose the instrument that is designed to tap an underlying construct whose conceptual definition most closely matches your own.

3. Never decide to use a particular instrument based solely on its title.

4. Choose an instrument for which there is evidence of reliability and validity.

5. Given a choice among alternatives, select an instrument whose breadth of coverage matches your own conceptual scope.

6. Select an instrument that provides a level of measurement appropriate to your research goals.

7. Choose an instrument with a time frame and response format that meet your needs.

8. Match the reading level required to understand the items in the instruments you select to the age of the intended respondents.

9. Never use an instrument unless you know how you’ll score it and how you’ll analyze it.

10. Rather than choosing only one measure, when feasible use multiple measures of the construct you wish to assess.

Following these guidelines will help you to select instruments wisely.

Addendum

Obtaining Instruments Once Identified

Adhering to these ten guiding principles can help you identify appropriate measurement instruments, but then you need to obtain them and permission to use them. Indeed, physical availability may ultimately dictate instrument choice. An instrument may be unavailable for a variety of reasons, including copyright restrictions, being out-of-print, or death of the primary author. Given such obstacles, rather than trying to contact the original developer to obtain copies of an instrument, it may be best to contact BMDS, Behavioral Measurement Database Services, creator of the HaPI database, to secure permission to use an instrument, and to obtain a hardcopy of it along with any scoring instructions. For a reasonable fee, BMDS will perform these services.

References


Ten Commandments for Selecting Self-Report Instruments (continued)


Ware, J.E., Jr. (1993). *SF-12 Health Survey (SF-12)*. Boston, MA: Medical Outcomes Trust.

Fred Bryant is Professor of Psychology at Loyola University, Chicago. He has roughly 90 professional publications in the areas of social psychology, personality psychology, measurement, and behavioral medicine. In addition, he has coedited 6 books, including *Methodological Issues in Applied Social Psychology* (1993; New York: Plenum). Dr. Bryant has extensive consulting experience in a wide variety of applied settings, including work as a research consultant for numerous marketing firms, medical schools, and public school systems; a methodological expert for the U.S. Government Accounting Office; and an expert witness in several federal court cases involving social science research evidence. He is currently on the Editorial Board of the journal *Basic and Applied Social Psychology*. His current research interests include happiness, the measurement of cognition and emotion, and structural equation modeling.

Remember me, I’m HaPI at BMDS!

Fred Bryant is Professor of Psychology at Loyola University, Chicago. He has roughly 90 professional publications in the areas of social psychology, personality psychology, measurement, and behavioral medicine. In addition, he has coedited 6 books, including *Methodological Issues in Applied Social Psychology* (1993; New York: Plenum). Dr. Bryant has extensive consulting experience in a wide variety of applied settings, including work as a research consultant for numerous marketing firms, medical schools, and public school systems; a methodological expert for the U.S. Government Accounting Office; and an expert witness in several federal court cases involving social science research evidence. He is currently on the Editorial Board of the journal *Basic and Applied Social Psychology*. His current research interests include happiness, the measurement of cognition and emotion, and structural equation modeling.

**HaPI**

Health and Psychosocial Instruments
Search for measurement instruments
in the Health and Psychosocial Instruments (HaPI) 
database with over 105,000 records of measurement instruments online or on CD-ROM!

Produced by BMDS • Behavioral Measurement Database Services

SUBSCRIBE TO HaPI!
Call Ovid Technologies at 800-950-2035 ext 6472 
for pricing and to order today!

Obtain copies of measurement instruments 
from BMDS Instrument Delivery!

Call BMDS at 412-687-6850, fax 412-687-5213, or 
e-mail bmdshapi@aol.com ($20/$30 processing/handling fee)

HaPI Thoughts

Good News!

Ovid Technologies now comprises both the Ovid and SilverPlatter platforms. 
Health and Psychosocial Instruments (HaPI) continues on the Ovid platform 
and can be found in the categories of Medicine & Allied Health, and Social Sciences.

Measurement Assistance?

Yes indeed! It’s just a phone call away and with a smile you get searching tips and suggestions for quick retrieval of the instruments you need. 
Call Evelyn Perloff, PhD, for Measurement Assistance from BMDS at 412-687-6850.
In This Issue:

- Introduction to This Issue
  *Al K. DeRoy* .................................................. 1

- Refinements to the Lubben Social Network Scale: The LSNS-R
  *James Lubben, Melanie Gironda, and Alex Lee* ............ 2

- Self-Estimated Intelligence
  *Adrian Furnham* ............................................... 12

- Ten Commandments for Selecting Self-Report Instruments
  *Fred B. Bryant* ............................................... 18

- HaPI Thoughts .................................................. 25