A Review of Scale Development Practices in the Study of Organizations

Timothy R. Hinkin
Cornell University

Questionnaires are the most commonly used method of data collection in field research (Stone, 1978). Problems with the reliability and validity of measures used on questionnaires has often led to difficulties in interpreting the results of field research (Cook, Hepworth, Wall & Warr, 1981; Schriesheim, Powers, Scandura, Gardiner & Lankau, 1993). This article reviews scale development procedures for 277 measures used in 75 articles published in leading academic journals from 1989 to 1994. It points out some of the problems encountered and provides examples of what could be considered “best practices” in scale development and reporting. Based on the review, recommendations are made to improve the scale development process.

Questionnaires are the most commonly used method of data collection in field research (Stone, 1978). Over the past several decades hundreds of scales have been developed to assess various attitudes, perceptions, or opinions of organizational members in order to examine a priori hypothesized relationships with other constructs or behaviors. As Schwab (1980) points out, measures are often used before adequate data exist regarding their reliability and validity. Many researchers have drawn seemingly significant conclusions from the application of new measures, only to have subsequent studies contradict their findings (Cook et al., 1981). Often scholars are left with the uncomfortable and somewhat embarrassing realization that results are inconclusive and that very little may actually be known about a particular topic. Although there may be a number of substantive reasons why different researchers arrive at varying conclusions, perhaps the greatest difficulty in conducting survey research is assuring the accuracy of measurement of the constructs under examination (Barrett, 1972). For example, recent studies of power and influence (Schriesheim & Hinkin, 1990; Schriesheim, Hinkin & Podsakoff, 1991) and organizational commitment (Meyer, Allen & Gellatly, 1990; Reichers, 1985) have found that measurement problems led to difficulties in interpreting results in both of these areas of research. Korman (1974, p. 194) states that “the point is not that
adequate measurement is 'nice.' It is necessary, crucial, etc. Without it we have nothing.” Even with advanced techniques such as meta-analysis, strong conclusions often cannot be drawn from a body of research due to problems with measurement (Schmidt, Hunter, Pearlman & Hirsch, 1985).

Developing sound scales is a difficult and time-consuming process (Schmitt & Klimoski, 1991). The success in observing true covariance between the variables of interest is dependent on the ability to accurately and reliably operationalize the unobservable construct. Several criteria have been proposed for assessing the psychometric soundness of behavioral measures. The American Psychological Association (1985) states that measures should demonstrate content validity, criterion-related validity, construct validity, and internal consistency. Content validity refers to the adequacy with which a measure assesses the domain of interest. Criterion-related validity pertains to the relationship between a measure and another independent measure. Construct validity is concerned with the relationship of the measure to the underlying attributes it is attempting to assess. Internal consistency refers to the homogeneity of the items in the measure or the extent to which item responses correlate with the total test score. There are specific practices that can be utilized to establish evidence of validity and reliability of new measures.

This article provides a review of scale development procedures from recently published academic articles and describes the stages necessary for the development of scales in accordance with established psychometric principles. It also integrates findings from recent studies that are germane to the topic of scale development. This review is aimed at two audiences, those who conduct research and those who evaluate it for possible publication. It focuses on both common scale development and reporting practices, presents problems that seem to exist in these practices, and discusses what might be considered “best practices” to assure that new measures satisfy the APA criteria for validity and reliability.

Review of the Literature

A literature search was undertaken to identify a sample of studies published from 1989 through 1993 whose primary purpose was the development of new measures or that utilized a new measure or measures as the focal variables of interest in the study. Only field studies were included in the sample. Six journals were targeted that the author felt would be representative of field research in the area of organizational behavior, resulting in 75 articles that fulfilled the aforementioned criteria. To identify potential articles for inclusion, each issue of the journals chosen for the study was examined. First, the abstracts were read, looking for key words such as “measures were developed” or “scales created for this study.” The Methods section of potential articles was also scanned, focusing on the “Measures” subheading to identify articles appropriate for the current study. The sample included (number of studies in parentheses) Journal of Applied Psychology (25), Organizational Behavior and Human Decision Processes (5), Human Relations (10), Journal of Management (12),
The total number of measures examined was 277. Nineteen articles were published in 1989, 12 in 1990, 14 in 1991, 13 in 1992, and 17 in 1993. The resulting sample is clearly not exhaustive, but it is felt to be representative of published articles incorporating newly developed measures (See Appendix for articles included in the study).

In beginning the review of this collection of articles, it was necessary to determine the practices that would be compared and the criteria that would be used for comparison. Schwab (1980) suggests that the development of measures falls into three basic stages. Stage 1 is item development, or the generation of individual items. Stage 2 is scale development, or the manner in which items are combined to form scales. Stage 3 is scale evaluation, or the psychometric examination of the new measure. The following review will be presented in the order of these stages and further broken down into steps in the scale development process.

**STAGE 1: Item Generation**

In item generation, the primary concern is content validity, which may be viewed as the minimum psychometric requirement for measurement adequacy and is the first step in construct validation of a new measure (Schriesheim et al., 1993). Content validity must be built into the measure through the development of items. As such, any measure must adequately capture the specific domain of interest yet contain no extraneous content. There seems to be no generally accepted quantitative index of content validity of psychological measures, and judgement must be exercised in validating a measure (Stone, 1978). There were two basic approaches to item development used in this sample. The first is deductive, sometimes called “logical partitioning,” or “classification from above.” The second method is inductive, known also as “grouping,” or “classification from below” (Hunt, 1991). In the present sample, 62 (83%) of the studies were deductive, eight (11%) were inductive, while five (6%) used a combination of both techniques.

Deductive scale development utilizes a classification schema or typology prior to data collection. This approach requires an understanding of the phenomenon to be investigated and a thorough review of the literature to develop the theoretical definition of the construct under examination. The definition is then used as a guide for the development of items (Schwab, 1980). This approach was used in the studies examined in two primary ways. First, researchers derived items designed to tap a previously defined theoretical universe. The second method was for the researchers to again develop conceptual definitions grounded in theory, but to then utilize a sample of respondents who were subject matter experts to provide critical incidents that are subsequently used to develop items.

Conversely, the inductive approach is so labeled because there is often little theory involved at the outset as one attempts to identify constructs and generate measures from individual responses. Researchers usually develop scales
inductively by asking a sample of respondents to provide descriptions of their feelings about their organizations or to describe some aspect of behavior. An example might be, "Describe how your manager communicates with you." Responses are then classified into a number of categories by content analysis based on key words or themes. Both deductive and inductively generated items may then be subjected to a sorting process that will serve as a pretest, permitting the deletion of items that are deemed to be conceptually inconsistent. In the current sample 13 (17%) of the studies reported the use of a content analysis of items, although the procedures varied substantially.

In the articles reviewed, it was frequently not reported exactly how items were generated or derived, if they were theoretically based, or if they had been pretested to assess content validity in any way. Double-barrel questions tapping more than one behavior or attitude were sometimes used (e.g., "I generally have sufficient information to make correct decisions and perform my job," Pierce, Gardner, Dunham & Cummings, 1993, p. 278). Often, it was merely stated that measures were, "developed expressly for this study" (e.g., Greenhaus, Parasuraman & Wormley, 1990, p. 73). In several cases it was stated that, "items that appear to capture..." a content domain did not when subjected to subsequent analysis (e.g., Ettlie and Reza, 1992, p. 811). Even with a well thought-out item development procedure, several authors found through subsequent sorting or factor analytical techniques that items were not perceived by respondents to tap the predicted construct (e.g., Pearce & Gregerson, 1991). Those items that did not load as predicted in subsequent factor analysis were usually deleted from the measure. There are many examples of concise and succinct descriptions of how items were derived (e.g., Giles & Mossholder, 1990; Yukl & Falbe, 1990). As an example of "best practices" reporting Giles and Mossholder clearly cite the theoretical literature on which the new measures are based and describe the manner in which the items were developed and the sample used for item development. In many articles this information was lacking, and it was not clear whether there was little justification for the items chosen or if the methodology employed was simply not adequately presented.

There were also many very good descriptions of item development "best practices" with respect to domain sampling. Adopting the deductive approach, MacKenzie, Podsakoff and Fetter (1991) clearly describe how the authors developed items to tap five organizational citizenship construct domains specified by Organ (1988). They first generated theoretically derived items and then subjected them to a content validity assessment by ten faculty members and doctoral students who were asked to classify each randomly ordered item to one of six categories, the five dimensions plus an "other" category. Those items that were assigned to the proper a priori category more than 80% of the time were retained for use in the questionnaire and are presented in the article. Butler (1991) utilized an inductive approach for the generation of items to assess conditions of trust. He clearly presents his use of semi-structured interviews of managers who described a trusted and mistrusted individual and also described critical incidents that led to trust or to distrust. The author then isolated 280 clauses concerning trust and 174 concerning mistrust. These clauses
were then classified independently by graduate students into 10 categories that were inferred to be conditions of trust. Interrater consistency was reported to be in excess of .78. These 10 conditions were then defined and four items were written to correspond to each of the definitions based on the clauses. Items were not presented in the article, however.

To summarize, the generation of items may be the most important part of developing sound measures. There seem to be two primary concerns with respect to item generation. First, and most important, it appears possible that some of the measures used in the studies reviewed may lack content validity. Second, the manner in which researchers report the item generation process may do a disservice, due to the omission of important information regarding the origin of measures. It would seem that a necessary prerequisite for new measures would be establishing a clear link between items and their theoretical domain. This can be accomplished by beginning with a strong theoretical framework and employing a rigorous sorting process that matches items to construct definitions. This process should be succinctly and clearly reported. The inductive approach may be more susceptible to problems at this stage and particular care must be taken to assure a consistent perspective within a measure. For example, even with the careful process undertaken by Butler (1991), managerial behaviors may be mixed with situational conditions in the same scale. Items that comprise a new measure should always be presented for examination. Because sorting is a cognitive task that requires intellectual ability rather than work experience, the use of students at this stage of scale development is appropriate (Schriesheim & Hinkin, 1990). As pointed out by Schriesheim et al. (1993), content adequacy should be assessed immediately after items have been developed as this provides the opportunity for the researcher to refine and/or replace items before large investments have been made in questionnaire preparation and administration.

STAGE 2: Scale Development

STEP 1—Design of the Developmental Study

At this stage of the process the researcher has identified a potential set of items for the construct or constructs under consideration. The next step is the administration of these items to examine how well they confirmed expectations about the structure of the measure. This process includes an assessment of the psychometric properties of the scale which will be followed by an examination of its relationship with other variables of interest.

There has been considerable discussion regarding several important issues in measurement that impact scale development. The first deals with the sample chosen, which should be representative of the population that the researcher will be studying in the future and to which results will be generalized. A clear description of the sample, the sampling technique, response rates, and the questionnaire administration process was provided in virtually every study. The samples used in the studies included business or industry (50, 67%), education
(12, 16%), government or military (7, 9%), and healthcare (6, 8%). The majority of studies were conducted within a single organization and the rationale for why these samples were selected was often not made clear.

The next issue of concern is the use of negatively worded (reverse-scored) items. Reverse-scored items have been employed primarily to attenuate response pattern bias (Idaszak & Drasgow, 1987). In recent years, however, their use has come under close scrutiny by a number of researchers. Reverse-scoring of items has been shown to reduce the validity of questionnaire responses (Schriesheim & Hill, 1981) and may introduce systematic error to a scale (Jackson, Wall, Martin & Davids, 1993). Researchers have shown that they may result in an artifactual response factor consisting of all negatively-worded items (Harvey, Billings & Nilan, 1985; Schmitt & Stultz, 1985). Reverse-scored items were reported to be used in 31 (41%) of the studies, although in 13 (17%) of the studies it was not possible to determine if reverse scoring was used because it was not mentioned or items were not presented. An examination of those studies that used negatively worded items did not reveal any discernible pattern of problems in subsequent analyses, however, item loadings for reverse-scored items were often lower than positively worded items that loaded on the same factor.

The third issue in scale construction is the number of items in a measure. Both adequate domain sampling and parsimony are important to obtain content and construct validity (Cronbach and Meehl, 1955). Total scale information is a function of the number of items in a scale, and scale lengths could affect responses (Roznowski, 1989). Keeping a measure short is an effective means of minimizing response biases (Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1990) but scales with too few items may lack content and construct validity, internal consistency and test-retest reliability (Kenny, 1979; Nunnally, 1976), with single-item scales particularly prone to these problems (Hinkin & Schriesheim, 1989). Scales with too many items can create problems with respondent fatigue or response biases (Anastasi, 1976). Additional items also demand more time in both the development and administration of a measure (Carmines & Zeller, 1979). Adequate internal consistency reliabilities can be obtained with as few as three items (Cook et al., 1981) and adding items indefinitely makes progressively less impact on scale reliability (Carmines & Zeller, 1979). In the current study, measures of a single construct varied in length from a single item to 46 items. Six studies reported the use of single-item measures while 12 studies used 2-item measures. Some very long scales had acceptable reliabilities but appeared to tap more than one conceptual dimension. In many cases poorly conceptualized items had to be deleted due to low factor loadings resulting in shortened scales, while in other cases there was redundancy in a long measure. Table 1 presents the frequency of use of scales for the 277 measures examined in the current study.

With respect to the fourth issue, scaling of items, it is important that the scale used generates sufficient variance among respondents for subsequent statistical analysis. Likert-type scales were used in all but two of the studies, with response options ranging from 3 points to 10 points. Coefficient alpha
Table 1.* Frequency of Use of Scale By Number of Items

<table>
<thead>
<tr>
<th>Number of Items in Scale</th>
<th>Number of Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Greater Than 10</td>
<td>24</td>
</tr>
<tr>
<td>Unspecified</td>
<td>2</td>
</tr>
</tbody>
</table>

*Note: a. Total Studies = 75
Total Scales = 277

reliability with Likert-type scales has been shown to increase up to the use of five points, but then it levels off (Lissitz & Green, 1975). Thirty-seven (49%) studies reported use of a 5-point response, while 30 (40%) used a 7-point response. Four studies used a “Yes, Uncertain, No” format, one used a 4-point bipolar scale, and one used a 7-point semantic differential. Eight of the studies used two different types of scaling.

The fifth issue is that of the sample size needed to appropriately conduct tests of statistical significance. The results of many multivariate techniques can be sample specific and increases in sample size may ameliorate this problem (Schwab, 1980). Simply put, if powerful statistical tests and confidence in results are desired, the larger the sample the better, but obtaining large samples can be very costly (Stone, 1978). As sample size increases, the likelihood of attaining statistical significance increases, and it is important to note the difference between statistical and practical significance (Cohen, 1969).

Both exploratory and confirmatory factor analysis, discussed below, have been shown to be particularly susceptible to sample size effects. Stable estimates of the standard errors provided by large samples result in enhanced confidence that observed factor loadings accurately reflect true population values. Recommendations for item-to-response ratios range from 1:4 (Rummel, 1970) to at least 1:10 (Schwab, 1980) for each set of scales to be factor analyzed. Recent research, however, has found that in most cases, a sample size of 150 observations should be sufficient to obtain an accurate solution in exploratory factor analysis as long as item intercorrelations are reasonably strong (Guadagnoli & Velicer, 1988). For confirmatory factor analysis, a minimum sample size of 200 has been recommended (Hoelter, 1983). Only three studies with total samples of less than 100 respondents were reported, and the largest total sample reported was 9205. Although sample sizes were generally large enough to provide adequate statistical power, sample size may have impacted...
the results of several studies. For example, Viswanathan (1993) found different factor structures for samples of 90 and 93 when factor analyzing a 20-item measure.

An excellent example of "best practices" when dealing with these five issues is provided by Jackson et al. (1993). In developing scales to assess the degree to which workers have control over their jobs they selected multiple large samples (225, 165) from manufacturing companies that utilized computer technology and manufacturing equipment that varied in the degree to which it was manually controlled. They intentionally did not use any reverse-scored items, citing studies that have shown that they may introduce systematic error. They derived 22 items to assess five constructs, using a 5-point Likert scale.

To summarize, in designing a study to examine the psychometric properties of a new measure it should be made clear why a specific sample was chosen. Based on previous research and the studies included in this review, a sample of 150 would seem to be the minimum acceptable for scale development procedures. The use of negatively worded items will probably continue due to the belief that they reduce pattern response bias, but researchers should carefully examine factor loadings of individual items and their impact on internal consistency reliability. Scale length is also an important issue as both long and short measures have potential negative effects on results. Proper scale length may be the most effective way to minimize a variety of types of response biases, assure adequate domain sampling, and provide adequate internal consistency reliability. Based on the results of the current study, scales comprised of five or six items that utilize five or seven point Likert scales (89% in the current study) would be adequate for most measures.

**STEP 2—Scale Construction**

Factor analysis is the most commonly used analytic technique for data reduction and refining constructs (Ford, McCallum & Tait, 1986) and was frequently used in the reviewed studies. Fifty-three (71%) studies reported the use of some type of factor analytical technique to derive the scales. Several studies also reported using interitem correlations to determine scale composition. In two studies item response theory was used for item analysis. Roznowski (1989) examined item popularity and biserial and point-biserial correlations between item responses and scale scores as measures of item discrimination. In 22 studies, however, different criteria were used and these will be discussed below. Table 2 presents a summary of the methods used to aggregate items into scales.

Principal components analysis with orthogonal rotation was the most frequently reported factoring method (25, 33%). Retaining factors with Eigenvalues greater than one was the most commonly used criteria for retention of factors, although the use of scree tests based on a substantial decrease in Eigenvalues were occasionally reported. Factor analytical results were reported more frequently than the 20% noted previously by Podsakoff and Dalton (1987). This could be expected due to the type of article selected for this review, however, reporting problems noted by Ford et al. (1986) were also found as the total
Table 2. Item Aggregation Procedures

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Components Analysis</td>
<td>25</td>
</tr>
<tr>
<td>No Factor Analysis</td>
<td>22</td>
</tr>
<tr>
<td>Unspecified Factor Technique</td>
<td>12</td>
</tr>
<tr>
<td>Principal Axis Analysis</td>
<td>6</td>
</tr>
<tr>
<td>LISREL with Other Factoring Technique</td>
<td>5</td>
</tr>
<tr>
<td>LISREL</td>
<td>3</td>
</tr>
<tr>
<td>Common Factor Analysis</td>
<td>3</td>
</tr>
<tr>
<td>Item-Total Correlations</td>
<td>2</td>
</tr>
</tbody>
</table>

percentage of explained variance or the variance accounted for by each factor was reported in only 19 (25%) of the studies. In cases where total item explained variance was reported it ranged from 37% to 85.4%.

Several studies reported that items did not load as predicted but were retained in the measure, often resulting in low internal consistency reliabilities. For example, Arvey, Strickland, Drauden and Martin (1990, p. 700) stated that "...this resulting structure should be viewed as rationally constructed with the aid of empirical evidence." Coefficient alphas for five of nine measures developed in this study were less than .60. In several cases items with poor loadings had to be dropped, impacting the validity and reliability of the measure. In eliminating an inappropriately loading item, Johnston and Snizek (1991, p. 1264) stated that "While this reduces the reliability of the scale (the alpha drops from .66 to .59), it does strengthen the scale's construct validity." In many cases the criteria for retaining factors was not made clear as items and actual factor loadings were not presented. In several situations factor analysis forced the deletion of items, resulting in single, two, or three-item measures. Although mentioned very infrequently, there was some consistency in the method used and reported to determine appropriate item loadings, with .40 being the most commonly mentioned criterion, but items were retained with as little as a .30 loading on a specified factor. In several studies, measures administered to two independent samples resulted in very different factor structures (e.g., Kumar & Beyerlein, 1991; Viswanathan, 1993). There was often little attempt to attain parsimony, with as many as 46 items used to measure a single construct and 26 measures were comprised of more than ten items. Several studies reported sample-to-variable ratios lower than 3:1.

Many studies did present clear and thorough descriptions of factor analytical techniques and results as advocated by Ford et al. (1986). For example, Snell and Dean (1992) presented items and factor loadings and reported factoring and rotational method (principal components with varimax rotation), criteria for determining the number of factors to retain (Eigenvalues) and for satisfactory item loadings (magnitude of loadings and cross-loadings), and the percentage of variance accounted for (by factor and total). Rationale for the retention and deletion of items was clearly linked both theoretically and empirically.
Table 3. Frequency of Use of Indices for Assessment of Fit

<table>
<thead>
<tr>
<th>Assessment of Fit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significance of Chi Square</td>
<td>4</td>
</tr>
<tr>
<td>Root Mean Square Residuals</td>
<td>4</td>
</tr>
<tr>
<td>Significance of Item Loadings</td>
<td>3</td>
</tr>
<tr>
<td>Adjusted Goodness-of-Fit Index</td>
<td>3</td>
</tr>
<tr>
<td>Tucker-Lewis Index</td>
<td>3</td>
</tr>
<tr>
<td>Difference in Chi Square Between Models</td>
<td>2</td>
</tr>
<tr>
<td>Goodness-of-Fit Index</td>
<td>2</td>
</tr>
<tr>
<td>Respecification Using Modification Indices</td>
<td>2</td>
</tr>
<tr>
<td>Bentler-Bonett Index</td>
<td>1</td>
</tr>
<tr>
<td>Comparative Fit Index</td>
<td>1</td>
</tr>
<tr>
<td>Relative Noncentrality Index</td>
<td>1</td>
</tr>
<tr>
<td>Parsimonious Normed Fit Index</td>
<td>1</td>
</tr>
<tr>
<td>Non-Normed Fit Index</td>
<td>1</td>
</tr>
<tr>
<td>RHO</td>
<td>1</td>
</tr>
<tr>
<td>Ratio of Chi Square to Degrees of Freedom</td>
<td>1</td>
</tr>
</tbody>
</table>

Confirmatory factor analysis was reported for assessing the measurement model in eight (11%) of the studies, 5 of which were conducted in combination with exploratory factor analysis. In each of these, LISREL was used to assess the quality of the factor structure by statistically testing the significance of the overall model and of item loadings on factors. The purpose of the analysis is to assess the goodness-of-fit of rival models: a null model where all items load on separate factors, a single common factor model, and a multi-trait model with the number of factors equal to the number of constructs in the new measure (Jöreskog & Sörbom, 1989).

Recently, there has been much discussion about assessing the extent to which a model fits the data and 30 goodness-of-fit indices are now available for use (MacKenzie et al., 1991). In the eight studies in the current sample, 15 different means of assessing degree of fit were used. Table 3 presents the various indices and the frequency of their use.

There seems to be little consensus on what are the appropriate indices. Significance of Chi-square was reported most frequently; the smaller the Chi-square, the better the fit of the model. It has been suggested that a Chi-square two or three times as large as the degrees of freedom is acceptable (Carmines & McIver, 1981), but the fit is considered better the closer the Chi-square value is to the degrees of freedom for a model (Thacker, Fields & Tetrick, 1989). In the present sample, it was suggested that a ratio of 5 to 1 was "a useful rule of thumb" (Jackson et al., 1993, p. 755). As there is no statistical test of fit, evaluation of fit indices is somewhat subjective but Meyer, Allen and Smith (1993, p. 543) suggested that "higher values indicate a better fit to the data." Fit indices above .85 were reported as generally acceptable in the current sample. Several researchers expressed concern about the effects of sample size and encouraged the use of relative fit indices such as the comparative fit index (e.g.,
MacKenzie et al., 1991). With respect to reporting, all of the researchers presented items, item loadings, and fit indices clearly. Sweeney and McFarland (1993) provide an example of "best practices" by describing procedures and presenting results. They reported the proposed measurement model and item loadings, results from competing models, Chi square statistic, degrees of freedom, adjusted goodness-of-fit indices, Tucker-Lewis index, Bentler-Bonett index, and root mean square residuals.

The issues of both content and construct validity seem very important for those 22 (29%) scales not reported to have been subjected to a factor analysis. Three studies reported the use of item-total correlations to form scales. The most common practice was to merely provide internal consistency reliabilities for these measures. If alphas were of an acceptable level, typically greater than .70, it was inferred that they were adequate for use. Items that reduced alpha levels were usually eliminated, often resulting in two or three-item measures. In many cases items were not presented. At the other extreme, very long measures that were purported to measure a single construct were not subjected to any structural examination. For example Smith and Tisak (1993, p. 295) measured role disagreement using a 46-item measure assessing "various behaviors, knowledge, skills, traits, and abilities..." with a coefficient alpha in excess of .90. Many other measures appeared to be multidimensional or were strongly correlated with other scales purported to be measuring independent constructs such as Motowildo, Dunnette and Carter (1990) who report the use of two scales with coefficient alphas of .68 and .88 that were correlated at .85.

To summarize, the primary purposes of either exploratory or confirmatory factor analysis in scale construction are to examine the stability of the factor structure and provide information that will facilitate the refinement of a new measure. Excellent examples of both the procedures used and reporting practices of these types of analyses have been discussed (Snell & Dean, 1992; Sweeney & McFarlin, 1993). Because of the objective of the task of scale development, it is recommended that a confirmatory approach be utilized. Exploratory techniques allow the elimination of obviously poorly loading items, but the advantage of the confirmatory (LISREL, or similar approaches) analysis is that it allows the researcher more precision in evaluating the measurement model. This technique was utilized in only a small percentage of studies in this sample. This does seem somewhat odd, given the nature of the research under examination. This is a relatively new technique, however, and five of the eight studies reporting the use of LISREL were published in 1993. All of these studies, however, provided strong evidence of the stability of the measures. Although there is some disagreement about appropriate fit indices, there are useful heuristics that, when taken in aggregation can provide a relatively clear picture of the factorial stability of a new measure.

It is at this stage of scale construction, however, that poor item development practices create further problems. Scales should not be derived post hoc, based only on the results of factor analysis. Simply because items load on the same factor does not mean that they necessarily measure the same theoretical construct (Nunnally, 1978). Similarly, simply using internal consistency
reliabilities for scale construction is not adequate. In several cases, an examination of items within individual scales by the author revealed that they were either multidimensional and tapping more than one construct, or were examining more than one perspective, for example mixing behaviors with affective responses. Many of these scales demonstrated internal consistency reliabilities lower than the .70 recommended by Nunnally (1978, discussed further below). In several cases scales were developed to measure constructs where previously validated measures of the same construct already exist (e.g., commitment).

**STEP 3—Reliability Assessment**

The assessment of reliability could be considered part of the testing stage of the newly developed measure. As previously mentioned, however, several researchers deleted items to increase coefficient alpha in the construction of their measure, so the discussion of reliability is being included in the scale development stage. There are two basic concerns with respect to reliability, consistency of items within a measure and stability of the measure over time. Although reliability may be calculated in a number of ways, the most commonly accepted measure is internal consistency reliability using Cronbach's Alpha (Price & Mueller, 1986). Assessing the stability of a measure with a method such as test-retest reliability is appropriate only in those situations where the attribute being measured is not expected to change over time (Stone, 1978). Podsakoff and Dalton (1987) reported that two-thirds of articles in the top-tier organizational behavior journals published in 1985 reported reliability coefficients. Almost all of the studies in the current sample report coefficient alpha reliabilities (73, 97%). Seven studies reported using test-retest reliability, three used intrerrater reliability, one used split-half reliability, one the Spearman-Brown prophecy formula, and two studies did not report any form of reliability. Thirty-two (12%) of the measures reported internal consistency reliabilities of less than .70 (minimum recommended by Nunnally, 1978) and they ranged to as low as .55. The majority of the low reliabilities were reported for scales of 5 items (8 of 38, 21%), 4 items (8 of 55, 15%), 3 items (9 of 46, 20%), and 2 items (3 of 28, 11%). With respect to the 2-item measures, it was seldom made clear if coefficient alpha or correlation coefficients were being reported.

An examination of specific measures with low reliabilities revealed that these problems were largely attributable to the item generation and scale construction problems previously discussed. For example, Rapoport (1989) conducted a factor analysis of ten items in several samples that consistently had several items that loaded very poorly (<.40). The items appeared to be multidimensional and the resulting internal consistency reliability was less than .70. Similarly, Gaertner and Nollen (1989, p. 988) retained items with poor factor loadings “because of conceptual importance” to form a measure with an alpha of .65. The measures with low reliabilities and unsupportive findings were often provided with a caveat such as, “...the reliability of this scale was low with an alpha of .64 which may explain this [unpredicted] finding” (Oliver, 1990, p. 522). In some cases it was necessary to eliminate items to obtain an acceptable
reliability coefficient, resulting in two or three item measures which threaten the validity of the measure. As an example of the complex nature of scale development, one study reported a three-item scale with factor loadings of .79, .70, and .69 and an internal consistency reliability of only .55 (Parkes, 1990). As examples of "best practices," McAuley, Lombardo and Usher (1989) provide a very good example of assessing the reliability of a measure as they presented the items in the scale and used multiple methods including internal consistency, test-retest, and interrater reliability in the development and testing of a new measure. Smith and Tisak (1993) show that using the Spearman-Brown prophecy formula allows one to reduce the number of items in a long scale without negatively affecting the reliability.

To summarize, it would seem that progress is being made in the estimating and reporting of internal consistency reliability and it should be considered a necessary part of the scale development process. Almost 20 years ago Nunnally (1978) suggested that an alpha of .70 be the minimum acceptable standard for demonstrating internal consistency and there is little reason to believe that anything less than that is adequate today. It is troubling that such a large number of measures (12%) did not reach the .70 level. Many of these were scales comprised of just a few items and it might be recommended that measures would include at least five items. Seldom is the effort made to increase the reliability of an instrument by developing new items for administration to another sample if it is low. Reliability is a necessary pre-condition for validity (Nunnally, 1978). The problems with reliability again seem to reflect lack of attention at the item development stage of the research project. Alternatively, many measures have acceptable levels of internal consistency reliability yet may in fact lack content validity due to multidimensionality or inappropriate representation of the construct under examination. Finally, fewer than 10% of the studies reported the examination of the stability of a measure over time using test-retest. The lack of the use of multiple methods of reliability assessment should be an issue of concern, however, as suggested by Stone (1978), this method may be appropriate only in situations where changes in the construct under examination over time are not expected. For example, one study reported very high internal consistency reliabilities (> .80) but test-retest coefficients as low as .43 for the same measure (McCaulay et al., 1989). Although assessing both stability and internal consistency would be desirable, a recommended "best practices" alternative would be the administration of the measure to an additional sample as done by Hinkin & Schriesheim (1989).

**STAGE 3: Scale Evaluation**

The objective of the previous stages in the scale development process was to create measures that demonstrate validity and reliability. Factor analysis, internal consistency, and test-retest reliability provide evidence of construct validity, but it can also be examined in other ways (Cronbach & Meehl, 1955). Demonstrating the existence of a nomological network of relationships with other variables through criterion-related validity, assessing two groups who
would be expected to differ on the measure, and the demonstrating discriminant and convergent validity using a method such as the multitrait-multimethod matrix developed by Campbell and Fiske (MTMM, 1959) would provide further evidence of the construct validity of the new measure.

With respect to criterion-related validity, most of the studies in the sample focused on specific relationships that were theoretically justified in the introduction and literature review section of the article. Hypothesized relationships were usually confirmed using either correlation or regression analysis, and in four studies, using structural modeling. In many cases, authors stated that these relationships provided evidence supporting the validity of the new measure.

The issue of construct validity was specifically addressed by less than a quarter of the sample. Many authors stated that a stable factor structure provided evidence of construct validity. Discriminant validity analyses were conducted in six studies while convergent validity was assessed in seven. Pierce, Gardner, Cummings and Dunham (1989) assessed discriminant validity by factor analyzing their self-esteem with several affective measures with the resulting factor structure supporting the validity of the measure. Niehoff and Moorman (1993, p. 537) examined the “nomological network validity” of a new monitoring measure by correlating it with other leadership measures to demonstrate convergent validity. Differences on group scores were assessed in only two studies. In the majority of studies there was a potential common source/common method bias, and this issue was addressed in only six of the studies. MacKenzie et al. (1991) assessed the potential impact of this bias by creating a same-source factor in LISREL analysis to assess its effect on the overall fit of the proposed model. Butler (1991) controlled this bias by collecting data from multiple sources. Social desirability was assessed in only three studies using the Crowne and Marlowe (1964) measure, and had no significant relationship with the variables of interest.

It may be argued that, due to potential difficulties caused by common source/common method variance, it is inappropriate to use the same sample both for scale development and for assessing construct validity (e.g., Campbell, 1976). The factor analytical techniques that were used to develop the measures may result in factors that are sample specific and inclined toward high reliability (Krzystofik, Cardy & Newman, 1988). The use of an independent sample to provide an application of the measure in a substantive context will enhance the generalizability of the new measures (Stone, 1978). To the extent that hypotheses using the measure are confirmed, confidence in its construct validity will be increased. It is also recommended that when items are added or deleted from a measure, the “new” scale should then be administered to another independent sample (Anderson & Gerbing, 1991; Schwab, 1980). Twenty-five (33%) of the studies in the sample reported the use of multiple samples in the scale development and testing process. In several cases, items from one analysis that did not perform as expected were replaced by items that improved the content validity, factor structure, and reliability of a new measure. Several researchers used multiple samples to develop and test their scales, usually with
very good results. For example, Butler (1991) used nine samples and numerous techniques in the development of his measure of trust, including both field data as well as a laboratory study.

To summarize, construct validation is essential for the development of quality measures (Schmitt & Klimoski, 1991). There was a marked reliance on factor analytical techniques to infer the existence of construct validity. In the vast majority of articles no mention of validity was made at all. Due to the large sample sizes used in most studies, results supporting criterion-related validity were often statistically significant, but the magnitude of the relationship was small enough to be of little practical significance (cf. Hays, 1973). There were, however, several studies whose primary purpose was scale development that did an excellent job of demonstrating the validity of their constructs. For example, Jackson et al. (1993) administered their measure of job control to occupants of two different types of jobs that would be expected to have different levels of control. They then used analysis of variance to test for differences on the measure of control across the two groups and found that there were indeed significant differences which provided evidence of discriminant validity of the new measure. Ironson, Smith, Brannick Gibson and Paul (1989) adopted a multitrait-multimethod approach to examine the convergent and discriminant validity of their Job in General (JIG) scale with multiple samples. They first correlated their measure with four other measures of job satisfaction, resulting in correlations ranging from .67 to .80, providing evidence of convergent validity. They then correlated their measure and the Job Descriptive Index (JDI) with specific measures and general measures of satisfactions and other organizational outcomes such as trust and intent to leave, predicting that the JIG would correlate more highly with general than specific measures while the JDI would correlate more highly with specific measures. This was shown to be the case, providing evidence of discriminant validity. They also utilized regression analyses to demonstrate that the JIG added significantly to the variance explained by the JDI. Finally, they administered both measures to a sample before and after an organizational intervention and found significantly greater increases in the JDI than in the JIG measure. It was concluded that all of these analyses provided evidence in support of the construct validity of the new measure.

Conclusion

Cronbach and Meehl (1955) describe the complexity and challenge of establishing construct validity for a new measure. Inadequate measures may continue to be developed and used in organizational research for several reasons. First, researchers may not understand the importance of reliability and validity to sound measurement, and may rely on face validity if a measure appears to capture the construct of interest. It has been shown, however, that “...a measure may appear to be a valid index of some variable, but lack construct and/or criterion-related validity” (Stone, 1978, pp. 54-55). Second, developing sound scales is a difficult and time-consuming process (Schmitt & Klimoski, 1991).
Given the desire to complete research for submission for publication, the development of sound measures may not seem like an efficient use of a researcher's time. Third, the profession may place too great an emphasis on statistical analysis (Schmitt, 1989), while overlooking the importance of accuracy of measurement. Statistical significance is of little value, however, if the measures utilized are not reliable and valid (Nunnally, 1978). Finally, there seems to be no well-established framework to guide researchers through the various stages of scale development (Price & Mueller, 1986). As a result, even though a researcher may possess a strong quantitative foundation, the process of scale development may not be well understood, thus scale development efforts may be fragmented and incomplete. Theoretical progress, however, is simply not possible without adequate measurement (Korman, 1974; Schwab, 1980).

Taken in isolation, it does not seem that any one study in this sample is severely problematic. Taken in aggregate, however, it is apparent that significant problems in the process and reporting of measurement development continue to exist. That does not mean that there are not excellent examples of scale development, such as Butler (1991), MacKenzie et al. (1991) and Jackson et al. (1993), but it does mean that these studies should set an example of the process for others to follow. If one believes that the problems stem from ignorance rather than negligence, the solution is education rather than admonishment. It is probably true that there is a little of both at work, and the guilty parties sit at both the researchers' and reviewers' desks.

It may be useful to reflect back on what has been learned from this review using Schwab's (1980) guidelines. First, with respect to item development, much more care and attention must be given to manner in which items are created. Whether inductively or deductively derived there must be strong and clear links established between items and a theoretical domain. Enough items must be developed to allow for deletion, as some items that appear to be valid are not judged by others to be so and factor and reliability analyses often necessitate the deletion of items. A sorting process that assures content validity is not only necessary but relatively simple to accomplish. Oddly enough, this is probably the easiest and least time consuming part of conducting survey research as it does not require large numbers nor complex questionnaire development and administration, yet is often the most neglected (Schriesheim et al. 1993).

Assuming that items have been developed that provide adequate content validity, the primary concern in scale construction is scale length to assure adequate domain sampling, reliability, and to minimize response biases. The reporting of factor analysis results could be greatly improved if researchers followed the framework suggested by Ford et al. (1986). Confirmatory factor analytical techniques such as LISREL should be used more frequently in scale development. Interestingly, internal consistency reliabilities are usually reported in published research, but reliability simply does not assure validity (Nunnally, 1978). In many cases, scales with coefficient alphas of less than .70 were reported which is simply unacceptable. There are also other methods available to assess reliability, particularly stability over time, that are seldom used or reported.
The demonstration of construct validity of a measure is the ultimate objective of the scale development (Cronbach & Meehl, 1955). In the scale evaluation stage the use of means other than within-measure factor analysis and relationships with criterion variables should be encouraged to provide evidence of validity, and the items comprising new scales should be presented. The use of multiple methods and samples might be a necessary requirement in the development and use of new measures. Reviewers should more closely examine the measures being used in a study, and may require that logic for the use and empirical support of new scales be strong, particularly where scales already exist. If progress is to be made in understanding behavior in organizations, scales must be developed that accurately measure the dynamic under investigation, as quality research must begin with quality measurement. This is the responsibility of both the researcher and the reviewer.

In a larger sense, one of the objectives of researchers in the behavioral sciences might be to develop standardized measures based on multiple large samples to reduce the generation of equivocal results. Use of standardized measures would make it easier to compare findings and facilitate the development and testing of theory (Price & Mueller, 1986). First, however, it would be necessary to decide upon construct definitions, and there seems to be much disagreement on this subject. For example, feedback has been defined and operationalized in a number of ways, and, as a result, few generalizations can be made about the effects of feedback on individuals (Ilgen, Fisher & Taylor, 1979). The development of measures could be a fruitful area for collaborative research in specific areas in the future. It may also be appropriate to question the continued heavy reliance on survey questionnaires in organizational research. Alternative methodological such as ethnographic studies may help to dig beneath the surface of organizational phenomenon where survey research cannot take us.

Acknowledgment: An earlier version of this paper was presented at the 1992 Academy of Management Annual Meetings in Las Vegas. The research was sponsored in part by funding from the Center for Hospitality Research, Cornell University School of Hotel Administration. The author would like to thank Larry Williams and three anonymous reviewers for their helpful comments on earlier drafts of this article.

Appendix

Articles Included in Study


References


