A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires

TIMOTHY R. HINKIN
Cornell University

The adequate measurement of abstract constructs is perhaps the greatest challenge to understanding the behavior of people in organizations. Problems with the reliability and validity of measures used on survey questionnaires continue to lead to difficulties in interpreting the results of field research. Price and Mueller suggest that measurement problems may be due to the lack of a well-established framework to guide researchers through the various stages of scale development. This article provides a conceptual framework and a straightforward guide for the development of scales in accordance with established psychometric principles for use in field studies

In an extensive review of the organizational behavior literature, Hinkin (1995) found that inappropriate domain sampling, poor factor structure, low internal consistency reliability and poor reporting of newly developed measures continue to threaten our understanding of organizational phenomena. The creation of flawed measures may be due in part to the lack of a well-established framework to guide researchers through the various stages of scale development (Price & Mueller, 1986). If researchers determine that no measure exists with which to assess a particular phenomenon and decide to develop a new measure, some direction in scale development should prove useful. The importance of sound measurement is stated succinctly by Schoenfeldt (1984): "The construction of the measuring devices is perhaps the most important segment of any study. Many well-conceived research studies have never seen the light of day because of flawed measures" (p. 78).

Perhaps the greatest difficulty in conducting research in organizations is assuring the accuracy of measurement of the constructs under examination (Barrett, 1972). A construct is a representation of something that does not exist as an observable dimension of behavior, and the more abstract the construct, the more difficult it is to measure (Nunnally, 1978). Because researchers studying organizational behavior rely most heavily on the use of questionnaires as the primary means of data collection

Author's Note: Correspondence should be addressed to Timothy R. Hinkin, School of Hotel Administration, Cornell University, Ithaca, NY 14853-6901; telephone (607) 255-2938.
Organizational Research Methods, Vol. 1 No. 1, January 1998 104-121
© 1998 Sage Publications, Inc.

from the SAGE Social Science Collections. All Rights Reserved.

(Stone, 1978), it is crucial that the measures on these survey instruments adequately represent the constructs under examination.

The purpose of this article is to provide a conceptual framework and a straightforward guide for the development of scales in accordance with established psychometric principles for use in survey research. The article is directed toward those readers who may have limited knowledge or methodological expertise in the scale development process but who are somewhat familiar with many of the various statistical concepts and methods to be described herein. As such, no attempt will be made to describe any of the recommended techniques in great depth. Rather, the focus will be on the order in which the various analyses should be undertaken, potential problems that may arise, recommendations for reporting results, and ways in which the process may be made more effective. For the sake of brevity, the discussion will be oriented around one alternative in a step, but mention will be made of alternative ways of executing a particular step in the process. The article will describe the development of measures consisting of multiple scales. The process would be the same, although less complex, for developing a single, multi-item scale. Supplementary readings will often be recommended to provide the reader with the opportunity to examine discussed techniques in greater detail. A model for the scale development process is presented in Figure 1.

The Scale Development Process

Many criteria have been proposed for assessing the psychometric soundness of measurement instruments. The American Psychological Association (APA, 1995) states that an appropriate operational definition of the construct a measure purports to represent should include a demonstration of content validity, criterion-related validity, and internal consistency. Together, these provide evidence of construct validity—the extent to which the scale measures what it is purported to measure. There are three major aspects of construct validation: (a) specifying the domain of the construct, (b) empirically determining the extent to which items measure that domain, and (c) examining the extent to which the measure produces results that are predictable from theoretical hypotheses (Nunnally, 1978). Construct validity forms the link between theory and psychometric measurement (Kerlinger, 1986), and construct validation is essential for the development of quality measures (Schmitt & Klimoski, 1991). Each stage of the process described below will contribute to increasing the confidence in the construct validity of the new measure.

Step 1: Item Generation

The first stage of scale development is the creation of items to assess the construct under examination. The key to successful item generation is the development of a well-articulated theoretical foundation that would indicate the content domain for the new measure. At this point, the goal of the researcher is to develop items that will result in measures that sample the theoretical domain of interest to demonstrate content validity. Domain sampling theory states that it is not possible to measure the complete domain of interest, but that it is important that the sample of items drawn from potential items adequately represents the construct under examination (Ghiselli, Campbell, & Zedeck, 1981).

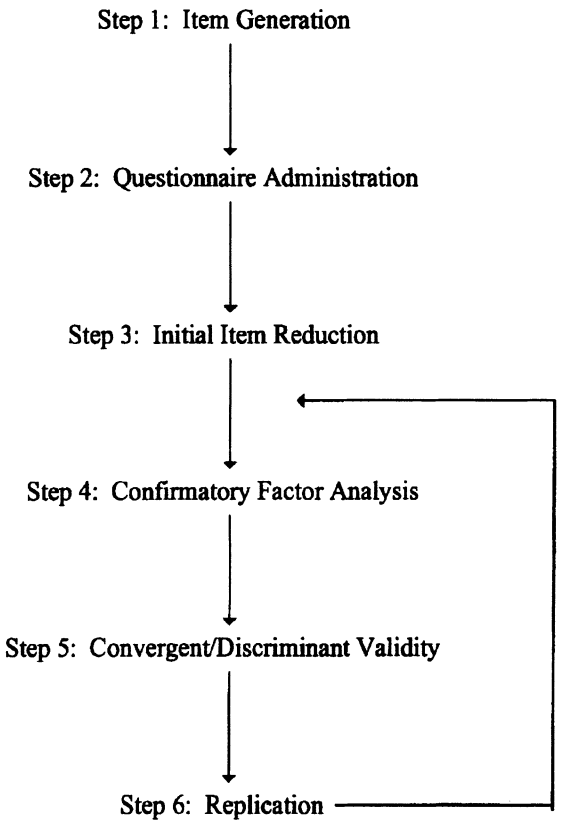

Figure 1: Scale Development Process

Once a thorough understanding of the theoretical foundation for the potential measure has been developed, there are several ways in which preliminary items may be created. The following discussion will cover two of these approaches. 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). Both of these techniques have been used by organizational researchers, and the scale developers must decide which is most appropriate in their particular situation. Each method will be briefly discussed below.

Deductive

Deductive scale development derives its name from the fact that the theoretical foundation provides enough information to generate the initial set of items. 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). For example, *expert power* might be defined as "the ability to

administer to another information, knowledge, or expertise." Items then may be generated from this definition, being sure that they are worded consistently in terms of describing a single behavior or an affective response.

Advantages and disadvantages. An advantage of the deductive approach to scale development is that if properly conducted, it will help to assure content validity in the final scales. Through the development of adequate construct definitions, items should capture the domain of interest. The disadvantages of the deductive approach are that it is very time-consuming and requires that researchers possess a working knowledge of the phenomena under investigation. In exploratory research, it may not be appropriate to attempt to impose measures onto an unfamiliar situation. In most situations in which theory does exist, the deductive approach would be most appropriate. (For an example of this approach, see Ironson, Smith, Brannick, Gibson, & Paul, 1989; Viega, 1991.)

Inductive

The inductive approach may be appropriate when the conceptual basis for a construct may not result in easily identifiable dimensions for which items can then be generated. 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 (see Williamson, Karp, Dalphin, & Gray, 1982) or a sorting process such as the Q-Sorting technique with an agreement index of some type, usually using multiple judges (see Anderson & Gerbing, 1991; Kerlinger, 1986). From these categorized responses, items are derived for subsequent factor analysis.

Advantages and disadvantages. This approach may be very useful when conducting exploratory research and when it is difficult to generate items that represent an abstract construct. The challenge arises, however, when attempting to develop items by interpreting the descriptions provided by respondents. Without a definition of the construct under examination, it can be difficult to develop items that will be conceptually consistent. This method requires expertise in content analysis and relies heavily on post hoc factor analytical techniques to ultimately determine scale construction, basing factor structure and, therefore, scales on item covariance rather than similar content. Although items may load on the same factor, there is no guarantee that they measure the same theoretical construct or come from the same sampling domain (Cortina, 1993). The researcher is compelled to rely on some theoretical framework, with little assurance that obtained results will not contain items that assess extraneous content domains (Schriesheim & Hinkin, 1990). This technique also makes the appropriate labeling of factors more difficult (Ford, MacCallum, & Tait, 1986). (For an example of this approach, see Butler, 1991; Kipnis, Schmidt, & Wilkinson, 1980.)

Item Development

There are a number of guidelines that one should follow in writing items. Statements should be simple and as short as possible, and the language used should be

familiar to target respondents. It is also important to keep all items consistent in terms of perspective, being sure not to mix items that assess behaviors with items that assess affective responses (Harrison & McLaughlin, 1993). Items should address only a single issue; "double-barreled" items such as "My manager is intelligent and enthusiastic" should be not be used. Such items may represent two constructs and result in confusion on the part of the respondents. Leading questions should be avoided, as they may bias responses. Items that all respondents would answer similarly should not be used, as they will generate little variance. The issue of negatively worded, reverse-scored items has stimulated much discussion and has strong proponents both for and against their use. Some researchers argue that the use of reverse-scored items may reduce response set bias (e.g., Price & Mueller, 1986). Others, however, have found that the use of a few of these items randomly interspersed within a measure may have a detrimental effect on psychometric properties of a measure (Harrison & McLaughlin, 1991). If the researcher does choose to use reverse-scored items, they must be very carefully worded to assure appropriate interpretation by respondents, and careful attention should be paid to factor loadings and communalities at the factor analytical stage of scale development (Schriesheim, Eisenbach, & Hill, 1989). (For a more detailed discussion of writing items, see Edwards, 1957; Warwick & Lininger, 1975).

Content Validity Assessment

After items have been generated, they should be subjected to an assessment of content validity. This process will serve as a pretest, permitting the deletion of items that are deemed to be conceptually inconsistent. There seems to be no generally accepted quantitative index of content validity of psychological measures, and judgment must be exercised in validating a measure (Stone, 1978). Methods do exist, however, to examine the consistency of judgments with respect to content validity.

Perhaps the most contemporary approach is that developed by Schriesheim and colleagues (Schriesheim, Powers, Scandura, Gardiner, & Lankau, 1993). The first step is to administer a set of items that have been developed to measure various constructs, along with definitions of these various constructs, to respondents. All items are included on every page, with a different definition at the top of each page. Respondents are then asked to rate on a Likert-type scale the extent to which each item corresponds to each definition. Schriesheim et al. (1993) also included a "does not match any definition" option but eliminated this category from analysis as responses to this were very infrequent. A Q-correlation matrix (item by item) of the data was then calculated, and that matrix was subjected to principal components analysis, extracting the number of factors corresponding to the theoretical dimensions under examination. Those items that met Ford et al.'s (1986) heuristic guideline for factor loadings .40 or greater on the appropriate factor with no major cross loadings were judged as meaningful and representative of the construct under examination.

A second recent advance in establishing content validity is the technique of substantive validity analysis developed by Anderson and Gerbing (1991). They propose two indexes to assess substantive validity. The first is the proportion of respondents who assign an item to its intended construct. The second is the degree to which each rater assigned an item to its intended construct. The latter index can be tested for statistical significance. This technique allows a researcher to pretest items

with small samples to quantitatively identify items that would be retained for use in field research and subsequent confirmatory factor analysis.

A similar technique of assessing content validity is to provide naive respondents with construct definitions, asking respondents to match items with their corresponding definition, also providing an "unclassified" category for items that are determined not to fit one of the definitions. This technique was employed recently by MacKenzie, Podsakoff, and Fetter (1991; see Hinkin, 1985, for a detailed description of this process). An acceptable agreement index—the percentage of respondents who correctly classify an item (minimum of 75%)—must be determined prior to administration of the items and definitions. Alternate forms with items in varying order should be used to maximize the efficacy of the sorting process.

Although none of these techniques will guarantee that content validity has been obtained, they will provide evidence of "content adequacy" as described by Schriesheim et al. (1993). The retained items should represent a reasonable measure of the construct under examination and reduce the need for subsequent scale modification. In the content validity assessment, it may be appropriate to use a small sample of students as this is a cognitive task not requiring an understanding of the phenomena under examination (Anderson & Gerbing, 1991; Schriesheim et al., 1993).

Number of Items

A very common question in scale construction is, "How many items?" There are no hard-and-fast rules guiding this decision, but keeping a measure short is an effective means of minimizing response biases caused by boredom or fatigue (Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1990). Additional items also demand more time in both the development and administration of a measure (Carmines & Zeller, 1979). Harvey, Billings, and Nilan (1985) suggest that at least four items per scale are needed to test the homogeneity of items within each latent construct. 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). It is difficult to improve on the internal consistency reliabilities of five appropriate items by adding items to a scale (Hinkin, 1985; Hinkin & Schriesheim, 1989; Schriesheim & Hinkin, 1990). Cortina (1993) found that scales with many items may have high internal consistency reliabilities even if item intercorrelations are low, an argument in favor of shorter scales with high internal consistency. It is also important to assure that the domain has been adequately sampled, as inadequate sampling is a primary source of measurement error (Churchill, 1979). As Thurstone (1947) points out, scales should possess simple structure, or parsimony. Not only should any one measure have the simplest possible factor constitution, but any scale should require the contribution of a minimum number of items that adequately tap the domain of interest. These findings would suggest that the eventual goal will be the retention of four to six items for most constructs, but the final determination must be made only with accumulated evidence in support of the construct validity of the measure. It should be anticipated that approximately one half of the created items will be retained for use in the final scales, so at least twice as many items as will be needed in the final scales should be generated to be administered in a survey questionnaire.

Item Scaling

With respect to scaling the items, it is important that the scale used generate sufficient variance among respondents for subsequent statistical analyses (Stone, 1978). Although there are a number of different scaling techniques available, such as Guttman and Thurstone, Likert-type scales are the most frequently used in survey questionnaire research (Cook et al., 1981) and are the most useful in behavioral research (Kerlinger, 1986). They also are most suitable for use in factor analysis. Although researchers have used 7-point and 9-point scales, Likert (1932) developed the scales to be composed of five equal appearing intervals with a neutral midpoint, such as *strongly disagree, disagree, neither agree nor disagree, agree, strongly agree.* Coefficient alpha reliability with Likert scales has been shown to increase up to the use of five points, but then it levels off (Lissitz & Green, 1975). Accordingly, it is suggested that the new items be scaled using 5-point Likert-type scales. If the scale is to be assessing frequency in the use of a behavior, it is very important that the researcher accurately benchmark the response range to maximize the obtained variance on a measure (Harrison & McLaughlin, 1991). For example, if available responses range from *once* to *five or more times* on a behavior that is very frequently used, most respondents will answer at the upper end of the range, resulting in minimal variance and the probable elimination of an item that might in fact have been important but was scaled incorrectly. (For an in-depth discussion of item scaling, see Bass, Cascio, & O'Connor, 1974.)

Summary

Reporting the manner in which items are derived is very important. The items must be clearly linked to the theoretical construct being assessed. Relationships that may seem obvious to the researcher due to immersion in the content area may not be as apparent to reviewers and readers. Investing the time and effort in assessing content adequacy in the beginning of a research project can minimize later problems with poor psychometric properties in a new measure. Keeping in mind that the final scales should be composed of four to six items each, the larger the number of items that have demonstrated "content adequacy" (Schriesheim et al., 1993) the better for the next step in the scale development process.

Step 2: Questionnaire Administration

In this stage of scale development, the researcher will use the items that have survived the content validity assessment described above to measure the construct under examination. The items should now be presented to a sample representative of the actual population of interest, such as managers or employees, with the objective of examining how well those items confirmed expectations regarding the psychometric properties of the new measure. (For a detailed discussion of sampling techniques, see Stone, 1978.) The new items should be administered along with other established measures to examine the "nomological network"—the relationship between existing measures and the newly developed scales. If possible, it would be advantageous to

collect information from sources other than the respondent to ameliorate the common source/common method concerns raised when collecting data from a single source. For example, performance data or peer assessment might be appropriate for use. Theory should dictate those variables with which the new measures should correlate or be independent. These other measures will be used later in subsequent analyses to provide preliminary evidence of criterion-related, convergent, and discriminant validity and, hence, the construct validity of the new scales.

Sample Size

In the scale development process, it will be necessary to use several independent samples. In the content validity pretest step of the process, both Schriesheim et al. (1993) and Anderson and Gerbing (1991) have suggested that small samples may be appropriate for their analyses, the former using a sample of 65 and the latter using two samples of 20. There has been substantial debate over 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). 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. Use of large samples assists in obtaining stable estimates of the standard errors to assure that factor loadings are accurate reflections of the true population values. At this stage of scale development, the researcher must ensure that data are collected from a sample of adequate size to appropriately conduct subsequent analyses. 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. Based on the latter recommendation, if 30 items were retained to develop three measures, at least 300 respondents would be needed for data collection. 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). It is suggested that the more conservative approach of at least 200 respondents be adopted for factor analysis. As the number of items increases, it may be necessary to increase the number of respondents. With larger samples, smaller differences tend to be detectable as more than mere sampling fluctuation (Hayduk, 1987).

Summary

It is at this stage that the researcher collects data to both evaluate the new measure's factor structure and also for subsequent examination of convergent, discriminant, and criterion-related validity with other measures. Selection of an appropriate type of sample is very important to assure enough variance in responses and avoid the effects of an idiosyncratic context. The sample used for the subsequent data collection should be of adequate size and be representative of the population of interest and be clearly described.

Step 3: Initial Item Reduction

Exploratory Factor Analysis

Once the data have been collected, it is recommended that factor analysis is used to further refine the new scales. Factor analysis allows the reduction of a set of observed variables to a smaller set of variables. This creates a more parsimonious representation of the original set of observations providing evidence of construct validity (Guadagnoli & Velicer, 1988). Because the principal-components method of analysis mixes common, specific, and random error variances, a common factoring method such as principal axis is recommended (Ford et al., 1986; Rummel, 1970). Prior to conducting the factor analysis, the researcher may find it useful to examine the interitem correlations of the variables, and any variable that correlates at less than .4 with all other variables may be deleted from the analysis (Kim & Mueller, 1978). A key assumption in the domain sampling model is that all items belonging to a common domain should have similar average intercorrelations. Low correlations indicate items that are not drawn from the appropriate domain and that are producing error and unreliability (Churchill, 1979).

The number of factors to be retained depends on both underlying theory and quantitative results. The researcher should have a strong theoretical justification for determining the number of factors to be retained, and the examination of item loadings on latent factors provides a confirmation of expectations. Eigenvalues of greater than 1 (Kaiser criterion) and a scree test of the percentage of variance explained (see Cattell, 1966) should be used to support the theoretical distinctions. If the items have been carefully developed, the number of factors that emerge on both Kaiser and scree criteria should equal the number of scales being developed. The intent is to develop scales that are reasonably independent of one another, so an orthogonal rotation is recommended for this analysis.

Keeping in mind that parsimony and simple structure are desired for the scales, the researcher should retain only those items that clearly load on a single appropriate factor. The objective is to identify those items that most clearly represent the content domain of the underlying construct. There are no hard-and-fast rules for this, but the .40 criterion level appears most commonly used in judging factor loadings as meaningful (Ford et al., 1986). A useful heuristic might be an appropriate loading of greater than .40 and/or a loading twice as strong on the appropriate factor than on any other factor. It may also be useful to examine the communality statistics to determine the proportion of variance in the variable explained by each of the items, retaining the items with higher communalities. The percentage of the total item variance that is explained is also important; the larger the percentage the better. Once again, there are no strict guidelines, but 60% could serve as a minimum acceptable target. At this stage, inappropriately loading items can be deleted, and the analysis repeated, until a clear factor structure matrix that explains a high percentage of total item variance is obtained. (For an in-depth discussion of factor analysis, see Ford et al., 1986; Kim & Mueller, 1978.)

Internal Consistency Assessment

Reliability is the accuracy or precision of a measuring instrument and is a necessary condition for validity (Kerlinger, 1986). The reliability of a measure should be assessed

after unidimensionality has been established (Gerbing & Anderson, 1988). Reliability may be calculated in a number of ways, but the most commonly accepted measure in field studies is internal consistency reliability using Cronbach's alpha (Price & Mueller, 1986). Use of this statistic is also recommended when used in conjunction with factor analysis (Cortina, 1993). At this step, the internal consistency reliabilities for each of the new scales is calculated. A large coefficient alpha (.70 for exploratory measures; Nunnally, 1978) provides an indication of strong item covariance and suggests that the sampling domain has been captured adequately (Churchill, 1979). If the number of retained items at this stage is sufficiently large, the researcher may want to eliminate those items that will improve or not negatively affect the reliability of the scales. This step is justified because the unidimensionality of individual scales has been established through the factor analyses previously conducted. Carmines and Zeller (1979) point out that the addition of items makes progressively less of an impact on the reliability and may in fact reduce the reliability if the additional items lower the average interitem correlation. Cortina (1993) found that alpha is very sensitive to the number of items in a measure, and that alpha can be high in spite of low item intercorrelations and multidimensionality. This suggests that .70 should serve as an absolute minimum for newly developed measures, and that through appropriate use of factor analysis, the internal consistency reliability should be considerably higher than .70. Most statistical software packages produce output that provides reliabilities for scales with individual items removed. At this stage, it is possible to retain those items that contribute to the internal consistency reliability and adequately capture the sampling domain. Reporting internal consistency reliability should be considered absolutely necessary.

Summary

Once again, reporting of results is very important in the development of a new measure. Ford et al. (1986) have made specific recommendations regarding reporting factor analytical results that bear repeating more than 10 years later. They suggest that "information that should be presented includes the

a. factor model;
b. method of estimating communalities (if applicable);
c. method of determining the number of factors to retain;
d. rotational method;
e. strategy of interpreting factors;
f. eigenvalues for all factors (if applicable);
g. percentage of variance accounted for (if using orthogonal rotation);
h. complete factor loading matrix;
i. descriptive statistics and correlation matrix if the number of variables is small;
j. computer program package;
k. method of computation of factor scores;
l. pattern matrix and interfactor correlations when oblique rotation is used." (p. 311)

Internal consistency reliability is a necessary but not sufficient condition for construct validity (APA, 1995) and should be clearly reported for all measures in a study. The focus of this section has been on a single type of reliability assessment, but

this by no means is suggesting that researchers should not attempt to assess the reliability of a measure with other means, particularly test-retest reliability.

Step 4: Confirmatory Factor Analysis

If the above steps are all carefully followed, it is highly likely that the new scales will be internally consistent and possess content validity. One of the weaknesses of typical factor analytical techniques is their inability to quantify the goodness of fit of the resulting factor structure (Long, 1983). Items that load clearly in an exploratory factor analysis may demonstrate a lack of fit in a multiple-indicator measurement model due to lack of external consistency (Gerbing & Anderson, 1988). A computer program such as LISREL provides a technique allowing the researcher to assess the quality of the factor structure by statistically testing the significance of the overall model and of item loadings on factors. This affords a stricter interpretation of unidimensionality than does exploratory factor analysis. In scale development, confirmatory factor analysis should be just that—a confirmation that the prior analyses have been conducted thoroughly and appropriately.

It is recommended that confirmatory factor analysis be conducted using the item variance-covariance matrix computed from data collected from an independent sample. Differences in item variances are lost in the analysis of correlations because all variables are standardized to a common variance (Harvey et al., 1985). If the initial sample was large enough, it may be possible to split the sample randomly in halves and conduct parallel analyses for scale development, using exploratory and confirmatory factor analyses (Krzystofiak, Cardy, & Newman, 1988). The purpose of the analysis is twofold. First, to assess the goodness of fit of the measurement model comparing a single common factor model with a multitrait model with the number of factors equal to the number of constructs in the new measure (Jöreskog & Sörbom, 1989). The multitrait model restricts each item to load only on its appropriate factor. The second purpose it to examine the fit of individual items within the specified model using the modification indices and *t* values.

The chi-square statistic permits the assessment of fit of a specific model as well as the comparison between two models. 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). A nonsignificant chi-square is desirable, indicating that differences between the model-implied variance and covariance and the observed variance and covariance are small enough to be due to sampling fluctuation. A model with a large chi-square may still be a good fit if the fit indices are high, as this measure is particularly dependent on sample size (Jöreskog & Sörbom, 1989). It is desirable to have a significantly smaller chi-square for the specified model than for competing models.

Recently, there has been increased attention on goodness-of-fit indices, and more than 30 have been used in confirmatory factor analysis (MacKenzie et al., 1991). It has been shown, however, that many measures of fit are sensitive to sample size differences. Medsker, Williams, and Holahan (1994) recommend that the chi-square statistic be used with caution and that the Comparative Fit Index (CFI) and the Relative Noncentrality Index (RNI) may be most appropriate to determine the quality of fit of

each model to the data. These indices involve a correction factor for the degrees of freedom by subtracting the relevant degrees of freedom from each chi square value used in the original Normed Fit Index (NFI). This accounts for the lower NFI resulting from models with a smaller sample size. The CFI ranges from 0 to 1 and is recommended when assessing the degree of fit of a single model. Evaluation of this index is somewhat subjective, however, as a heuristic; a value greater than .90 indicates a reasonably good model fit. The RNI may be most appropriate when comparing the fit of competing models, and the range is slightly different than the CFI, from –1 to 1.

The quality of the model can be further assessed by the item *t* values and modification indices. Once the overall fit of the model has been examined, each model coefficient should be individually examined for degree of fit. By selecting a desired level of significance (usually $p < .05$, Bagozzi, Yi, & Phillips, 1991), the researcher can use the *t* values to test the null hypothesis that the true value of specified parameters is zero, and those items that are not significant may need to be deleted.

Although the *t* values provide an estimate of fit for specified parameters, the modification indices provide information regarding unspecified parameters, or cross loadings, with a large modification index indicating that a parameter might also contribute explained variance to the model. Those items with indices of .05 or greater should be deleted and the analysis repeated. The results should then be examined with special attention to *t* values for all specified loadings. If all appropriate loadings are significant at $p < .05$ or less, and the magnitude of any inappropriate cross loadings as indicated by modification indices are relatively small, the researcher can be assured that the data fit the model quite well. Performing this model respecification should result in a smaller chi-square and larger goodness-of-fit indices. Keeping in mind that the objective is to retain four to six items per scale, this analysis, in conjunction with an assessment of internal consistency, offers the opportunity to delete items that contribute less explained variance to the measure. (For an in-depth discussion of confirmatory factor analysis, see Byrne, 1989; Jöreskog & Sörbom, 1989; Medsker et al., 1994.)

Summary

Confirmatory factor analysis allows the researcher to quantitatively assess the quality of the factor structure providing further evidence of the construct validity of the new measure. It is still subject to the use of judgment, however, and thoroughly and clearly reporting confirmatory factor analyses is very important. Results should include at a minimum the chi-square statistic, degrees of freedom, and the recommended goodness-of-fit indices used for each competing model. It may also be appropriate to report factor loadings and *t* values. If models are respecified based on *t* values and/or modification indices, this should also be reported, along with the final statistics.

Step 5: Convergent/Discriminant Validity

Up to this point, the researcher can be relatively assured that the new scales possess content validity and internal consistency reliability. Although the prescribed scale development process will build in a certain degree of construct validity, gathering further evidence of construct validity can be accomplished by examining the extent to

which the scales correlate with other measures designed to assess similar constructs (convergent validity) and to which they do not correlate with dissimilar measures (discriminant validity). It would also be useful to examine relationships with other variables with which the measures would be expected to correlate (criterion-related validity). The data collected from the samples used in the previous analyses would now be put to use.

Multitrait-Multimethod Matrix (MTMM)

Convergent and discriminant validity are most commonly examined by using the MTMM developed by Campbell and Fiske (1959; Schmitt & Klimoski, 1991). Although the original MTMM guidelines have been criticized by a number of researchers for use of unrealistic assumptions and reliance on a qualitative assessment of comparisons of correlations (e.g., Bagozzi et al., 1991), they are still useful in determining convergent and discriminant validity (Hollenbeck, Klein, O'Leary, & Wright, 1989; Marsh & Hocevar, 1988). The data from the additional measures obtained during the original questionnaire administration are used at this stage. A matrix is obtained by correlating the newly developed scales with the other measures and by examining the magnitudes of correlations that are similar and dissimilar.

Convergent validity is achieved when the correlations between measures of similar constructs using different methods, such as self-reported performance and performance evaluation data (monotrait-heteromethod), are "significantly different from zero and sufficiently large" (Campbell & Fiske, 1959, p. 82). Discriminant validity is achieved when three conditions are satisfied: First, when correlations between measures of the same construct with different methods (monotrait-heteromethod) are greater than correlations between different constructs measured with different methods (heterotrait-heteromethod); second, when correlations between the same construct using different methods (monotrait-heteromethod) are larger than correlations between different constructs measured with common methods (heterotrait-monomethod); and finally, when similar patterns of correlations exist in each of the matrices formed by the correlations of measures of different constructs obtained by the same methods (heterotrait-monomethod) and the correlations of different constructs obtained by different methods (heterotrait-heteromethod). (For an in-depth discussion of MTMM, see Schmitt & Stults, 1986.)

Alternative Methods

There have been several recent advances in techniques to assess convergent and discriminant validity. An analysis of variance approach has been developed to quantitatively analyze the MTMM data (Kavanagh, MacKinney, & Wolins, 1971). Factor analytical techniques also have been used to examine discriminant validity. For example, Grover (1991) used principal components analysis with multiple constructs to demonstrate discriminant validity. Becker and Vance (1993) have developed a refined model of the direct product approach to assess construct validity. Recent developments have been made in the use of confirmatory factor analysis for what Bagozzi et al. (1991) term "second-generation methods for approaching construct validity" (p. 429). The methodology is similar to that described in the confirmatory factor analysis section above, with the additional measures used in the analysis.

Bagozzi et al. provide evidence that the use of confirmatory factor analysis in construct validation overcomes the weaknesses of the Campbell and Fiske (1959) technique by providing a quantitative assessment of convergent and discriminant validity, and they recommend its use in future research. This technique recently has been adopted by other researchers (e.g., Shore & Tetrick, 1991) and may indeed eventually replace use of the MTMM. The use of the MTMM, however, has long been a well-accepted technique for establishing convergent and discriminant validity and should serve as a good starting point for establishing construct validity (Schmitt & Klimoski, 1991; Schoenfeldt, 1984).

Criterion-Related Validity

The researcher should also examine relationships between the new measures and variables with which they could be hypothesized to relate to develop a nomological network and establish criterion-related validity (Cronbach & Meehl, 1955). These relationships should be based on existing theory and may be examined using correlation or regression analyses. If hypothesized relationships attain statistical significance, evidence of criterion-related validity is provided. Also, null relationships should exist where hypothesized, such as between the new scales and measures of social desirability (Crowne & Marlowe, 1964).

Summary

The demonstration of construct validity of a measure is the ultimate objective of the scale development (Cronbach & Meehl, 1955). Attempts to demonstrate discriminant, convergent, and criterion-related validity should be clearly and succinctly reported. Theoretically justified relationships between variables in the current study could provide evidence of concurrent validity of the new measure. Although discriminant and convergent validity can be demonstrated using the MTMM, it is recommended that one of the more recently developed factor analytical techniques be used for this purpose.

Step 6: Replication

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 the psychometric properties of a new measure (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 (Krzystofiak et al., 1988). The use of an independent sample will enhance the generalizability of the new measures (Stone, 1978). 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). The use of a new sample would also allow the application of the measure in a substantive test. It would now be necessary to collect another set of data from an appropriate sample and repeat the scale-testing process with the new scales. To avoid the common source/common method problem, it is recommended that data from sources other than the respondent, such as peers or superiors, be collected where possible to provide evidence for

construct validity. The replication should include confirmatory factor analysis, assessment of internal consistency reliability, and convergent, discriminant, and criterion-related validity assessment. These analyses should provide the researcher with the confidence that the finalized measures possess reliability and validity and would be suitable for use in future research.

Conclusion

Scale development clearly involves a bit of art as well as a lot of science. Anyone who has gone through a process similar to that described above will understand the difficulty of developing sound measures. Use of a process similar to this has resulted in measures that appear to be psychometrically sound (e.g., Hinkin & Schriesheim, 1989; Kumar & Beyerlein, 1991; MacKenzie et al., 1991). By carefully following the process outlined in this article, the researcher should end up with measures that are efficient and effective to use and also satisfy APA standards for psychometric adequacy.

References

American Psychological Association (APA). (1995). *Standards for educational and psychological testing*. Washington, DC: Author.

Anderson J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76*, 732-740.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*, 421-458.

Barrett, G. V. (1972). New research models of the future for industrial and organizational psychology. *Personnel Psychology, 25*, 1-17.

Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology, 59*, 313-320.

Becker, T. E., & Vance, R. J. (1993). Construct validity of three types of organizational citizenship behavior: An illustration of the direct product model with refinements. *Journal of Management, 19*, 663-682.

Butler, J. K. (1991). Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *Journal of Management, 17*, 643-663.

Byrne, B. M. (1989). *A primer of LISREL*. New York: Springer-Verlag.

Campbell, J. P. (1976). Psychometric theory. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally.

Campbell, J. P., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carmines, E. G., & McIver, J. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. Bohrnstedt & E. Borgatta (Eds.), *Social measurement: Current issues*. Beverly Hills, CA: Sage.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*, 64-73.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cook, J. D., Hepworth, S. J., & Warr, P. B. (1981). *The experience of work*. San Diego: Academic Press.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

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

Crowne, D., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: John Wiley.

Edwards, A. L. (1957). *Techniques of attitude scale construction*. Norwalk, CT: Appleton-Century-Crofts.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*, 291-314.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25*, 186-192.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman.

Grover, S. L. (1991). Predicting the perceived fairness of parental leave policies. *Journal of Applied Psychology, 76*, 247-255.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Harrison, D. A., & McLaughlin, M. E. (1991). Exploring the cognitive processes underlying responses to self-report instruments: Effects of item content on work attitude measures. *Proceedings of the 1991 Academy of Management annual meetings*, 310-314.

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology, 78*, 129-140.

Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the job diagnostic survey: Good news and bad news. *Journal of Applied Psychology, 70*, 461-468.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL*. Baltimore: Johns Hopkins University Press.

Hinkin, T. R. (1985). *Development and application of new social power measures in superior-subordinate relationships*. Unpublished doctoral dissertation, University of Florida.

Hinkin, T. R. (1995). A review of scale development in the study of behavior in organizations. *Journal of Management, 21*, 967-988.

Hinkin, T. R., & Schriesheim, C. A. (1989). Development and application of new scales to measure the French and Raven (1959) bases of social power. *Journal of Applied Psychology, 74*, 561-567.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11*, 325-344.

Hollenbeck, J. R., Klein, H. J., O'Leary, A. M., & Wright, P. M. (1989). Investigation of the construct validity of a self-report measure of goal commitment. *Journal of Applied Psychology, 74*, 951-956.

Hunt, S. D. (1991). *Modern marketing theory*. Cincinnati, OH: South-Western.

Ironson, G. H., Smith, P. C., Brannick, M. T., Gibson, W. M., & Paul, K. B. (1989). Construction of a job in general scale: A comparison of global, composite, and specific measures. *Journal of Applied Psychology, 74*, 193-200.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications*. Chicago: SPSS.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin, 75*, 34-49.

Kerlinger, F. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart & Winston.

Kim, J., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it.* Beverly Hills, CA: Sage.

Kipnis, D., Schmidt, S. M., & Wilkinson, I. (1980). Intraorganizational influence tactics: Explorations in getting one's way. *Journal of Applied Psychology, 65*, 440-452.

Krzystofiak, F., Cardy, R. L., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluation behavior. *Journal of Applied Psychology, 73*, 515-521.

Kumar, K., & Beyerlein, M. (1991). Construction and validation of an instrument for measuring ingratiatory behaviors in organizational settings. *Journal of Applied Psychology, 76*, 619-627.

Likert, R. (1932). A technique for the measurement of attitude scales. *Archives of Psychology, 140.*

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*, 10-13.

Long, J. S. (1983). *Confirmatory factor analysis.* Beverly Hills, CA: Sage.

Mackenzie, S. B., Podsakoff, P. M., & Fetter, R. (1991). Organizational citizenship behavior and objective productivity as determinants of managerial evaluations of salespersons' performance. *Organizational Behavior and Human Decision Processes, 50*, 123-150.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*, 107-117.

Medsker, G. J., Williams, L. J., & Holohan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management, 20*, 439-464.

Nunnally, J. C. (1976). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Price, J. L., & Mueller, C. W. (1986). *Handbook of organizational measurement.* Marshfield, MA: Pitman.

Rummel, R. J. (1970). *Applied factor analysis.* Evanston, IL: Northwestern University Press.

Schmitt, N. W., & Klimoski, R. J. (1991). *Research methods in human resources management.* Cincinnati, OH: South-Western.

Schmitt, N. W., & Stults, D. M. (1985). Factors defined by negatively keyed items: The results of careless respondents? *Applied Psychological Measurement, 9*, 367-373.

Schmitt, N. W., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1-22.

Schoenfeldt, L. F. (1984). Psychometric properties of organizational research instruments. In T. S. Bateman & G. R. Ferris (Eds.), *Method & analysis in organizational research* (pp. 68-80). Reston, VA: Reston.

Schriesheim, C. A., & Eisenbach, R. J. (1990). Item wording effects on exploratory factor-analytic results: An experimental investigation. *Proceedings of the 1990 Southern Management Association annual meetings*, 396-398.

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1989, August). *An experimental investigation of item reversal effects on questionnaires.* Paper presented at the annual meeting of the Academy of Management, Washington, D.C.

Schriesheim, C. A., & Hinkin, T. R. (1990). Influence tactics used by subordinates: A theoretical and empirical analysis and refinement of the Kipnis, Schmidt, and Wilkinson subscales. *Journal of Applied Psychology, 75*, 246-257.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management, 19*, 385-417.

Schwab, D. P. (1980). Construct validity in organization behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 2, pp. 3-43). Greenwich, CT: JAI.

Shore, L. M., & Tetrick, L. E. (1991). A construct validity study of the survey of perceived organizational support. *Journal of Applied Psychology, 76*, 637-643.

Stone, E. (1978). *Research methods in organizational behavior*. Glenview, IL: Scott, Foresman.

Thacker, J. W., Fields, M. W., & Tetrick, L. E. (1989). The factor structure of union commitment: An application of confirmatory factor analysis. *Journal of Applied Psychology, 74*, 228-232.

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

Viega, J. F. (1991). The frequency of self-limiting behavior in groups: A measure and an explanation. *Human Relations, 44*, 877-894.

Warwick, D. P., & Lininger, C. A. (1975). *The sample survey: Theory and practice*. New York: McGraw-Hill.

Williamson, J. B., Karp, D. A., Dalphin, J. R., & Gray. P. S. (1982). *The research craft* (2nd ed.). Boston: Little, Brown.

Timothy R. Hinkin is an associate professor of management and director of undergraduate studies at the Cornell University School of Hotel Administration. He received his Ph.D. in organizational behavior from the University of Florida. His research interests focus on leadership, influence and power, and the dynamics between supervisors and front-line employees.