Scale Construction: Developing Reliable and Valid Measurement Instruments

Timothy R. Hinkin

Cornell University

J. Bruce Tracey

Cornell University

Cathy A. Enz

Cornell University

The purpose of this paper is to describe the process for developing reliable and valid measurement instruments that can be used in any hospitality industry field research setting. Many instances exist in which the researcher cannot find an adequate or appropriate existing scale to measure an important construct. In these situations it is necessary to create a new scale. Failure to carefully develop a measurement instrument can result in invalid and uninteqretable data. Hence, a systematic seven-step process is outlined here to assist researchers in devising usable scales. Examples from the authors' own research are used to illustrate some of the steps in the process.

Overview

Much of the research in the hospitality industry is conducted in field settings where the most commonly used method of data collection is the survey questionnaire (Schmitt and Klimoski, 1991; Stone, 1978). Unfortunately, questionnaires often have lacked reliability and validity which has led to difficulties in interpreting research results (Cook, Hepworth, Wall and Warr. 1981; Schriesheim, Powers, Scandura, Gardiner and Lankau, 1993). As noted in Hinkin (1995) many of the measures currently being used in the study of management have serious flaws. Some of these measures are being used in research conducted in hospitality settings (e.g., Tracey and Hinkin, 1996). In addition, measures commonly used in research in the industry have been shown to have psychometric problems (e.g. Carman, 1990). Researchers also need to develop measures to study phenomena unique to the industry, such as guest satisfaction among business travelers (Gunderson, Heide and Olsson, 1996). A well-established framework to guide researchers through the various stages of survey scale development is lacking.

While many researchers may not be interested in measurement per se, they often must find ways of studying important questions where existing scales are either inadequate, inappropriate or unavailable. This article builds on the work of Churchill (1979) and Hinkin (1995) and presents a seven-step process for scale development and analysis, using examples from our own research to illustrate the most appropriate methods for designing reliable and valid scales. The focus will be on the development of multiple measures each of which consists of multiple items. However, the process would be the same, although less complex, for developing a single scale with multiple items. We should note that there are many different types of measures, but the vast majority of scales used by behavioral scientists in survey questionnaires are Likert scales that utilize an interval level of measurement (Cook et al., 1981; Schmitt and Klimoski, 1991). As such, this paper will describe the process of the development of multi-item, multi-subscale, interval-level scales. Figure 1 lists the seven steps necessary to produce reliable and valid scales. The following sections cover each of the steps of scale development in greater detail.

Figure 1
Guidelines for Scale Development and Analysis

Step 1: Item Generation
Create Items

Step 2: Content Adequacy Assessment
Test for conceptual consistency of items

Step 3: Questionnaire Administration
Determine the scale for items
Determine an adequate sample size
Administer questions with other established measures

Step 4: Factor Analysis
Exploratory to reduce the set of items
Confirmatory to test the significance of the scale

Step 5: Internal Consistency Assessment
Determine the reliability of the scale

Step 6: Construct Validity
Determine the convergent and criterion-related validity

Step 7: Replication
Repeat the scale-testing process with a new data set

Step 1. Item Generation

The scale development process begins with the creation of items to assess a construct under examination. This process can be conducted inductively, by generating items first, from which scales are then derived, or deductively, beginning with a theoretical definition from

which items are then generated. Both of these approaches have been used by behavioral researchers and the decision must be made about which is most appropriate in a particular situation.

The inductive approach is usually used when exploring an unfamiliar phenomenon where little theory may exist. Experts on the subject are typically asked to provide descriptions of their feelings about their organizations or to describe some aspect of behavior. Responses are then classified into a number of categories by content analysis based on key words or themes. From these categorized responses, items are then derived.

Deductive scale development uses a theoretical definition of a-construct which is then used as a guide for the creation of items (Schwab, 1980). This approach requires an understanding of the relevant literature and of the phenomenon to be investigated and helps to ensure content adequacy in the final scales. In most situations where some theory exists, the deductive approach would be most appropriate. Getty and Thompson (1994) provide a good example of the deductive approach to item development for a measure of lodging quality.

Item Development

There are a number of basic guidelines that should be followed to ensure that the items are properly constructed. Some of the most important and often overlooked practices will be presented. Items should address only a single issue; "double-barreled" items such as "My employees are dedicated and hardworking" may represent two constructs and result in confusion on the part of the 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 to or outcomes of behaviors (Harrison and McLaughlin, 1993). As an example, in the examination of supervisory behavior, "My supervisor treats me fairly" should not be included in a scale with the outcome "I feel committed to my supervisor." Statements should be simple and as short as possible and the language used should be familiar to target respondents. Negatively-worded or reverse-scored items should be used with caution as a few of these items randomly interspersed within a measure can have a detrimental effect on its psychometric properties (Harrison and McLaughlin, 1991). Items must be understood by the respondent as intended by the researcher if meaningful responses are to be obtained. Finally, remember that content redundancies are desirable when creating multiple items because they are the foundation of internal consistency reliability.

Example: Item Generation and Development for the Multi-factor Leadership Questionnaire

As an example of the process for generating and developing multiple item, interval-level scales, we will consider the work of Bass and his colleagues (e.g., Bass, 1985; Bass and Avolio, 1994) in their attempts to measure the transformational leadership construct. Although there are many measures of leadership in the organizational studies literature, the transformational leadership construct is relatively new, has gained a great deal of attention in recent years and has been used in several hospitality settings (e.g., Tracey and Hinkin, 1994).

Bass' work in this area began in the late 1970s and early 1980s. Initially, Bass and his colleagues utilized both an inductive and deductive approach for generating items to assess the transformational leadership construct. Bass began inductively deriving a taxonomy of the transformational leadership process, and then he developed a questionnaire for future inquiries in this domain. In his first study, Bass asked 70 industry executives from a variety of work organizations to describe individuals whom they had encountered in their own careers who "raised their awareness about issues of consequence, shifted them to higher level needs and influenced them to transcend their own self-interests for the good of the group or organization and to work harder than they originally had expected they would" (Bass, 1985, p. 29). Based on the executives' responses and a review of the existing leadership literature (e.g., Burns, 1978), Bass and his colleagues developed 142 items that assessed a wide range of leadership behaviors. These items were created to represent two major dimensions of leadership: transformational and transactional.

Based on the executives' responses from Bass' first study, 11 graduate students used a simple sorting procedure to deductively classify each item as either transformational, transactional or neither. From this content analysis, 73 items were retained and included in the first version of the Multifactor Leadership Questionnaire (MLQ). Over the past several years, Bass and his colleagues have refined their operationalization of the transformational leadership construct. Currently they argue for four distinct, though inter-related, dimensions of transformational leadership.

Number of Items

There are no specific rules about the number of items to be retained but some helpful heuristics exist. A measure needs to be internally consistent and be parsimonious, comprised of the minimum number of items that adequately assess the domain of interest (Thurstone, 1947). Adequate internal consistency reliability can be obtained with four or five items per scale (Harvey, Billings and Nilan, 1985; Hinkin and Schriesheim, 1989). Keeping a measure short is an effective means of minimizing response biases caused by boredom or fatigue (Schmitt and Stults, 1985). Additional items also demand more time in both the development and administration of a measure (Carmines and Zeller, 1979). These issues would suggest that a quality scale comprised of four to six items could be developed for most constructs or conceptual dimensions. It should be anticipated that approximately one-half of the new items will be retained for use in the final scales, so at least twice as many items should be generated than will be needed for the final scales. Once the scale has been developed it is time to pretest the scale for the content adequacy of the items.

Step 2. Content Adequacy Assessment

An often overlooked yet necessary step in the scale development process is pretesting items for content adequacy. In many instances researchers have invested substantial time and effort in collecting large data sets only to find that an important measure is flawed. Assuring

content adequacy prior to final questionnaire development provides support for construct validity as it allows the deletion of items that may be conceptually inconsistent.

Several content assessment methods have been described in the research methods literature (cf., Nunnally, 1978). One common method requires respondents to categorize or sort items based on their similarity to construct definitions. This can be conducted using experts in a content domain. Naive respondents can also be used if they are able to read and understand the definitions and items, and students can often be used during this stage of scale development. In either case, respondents are presented with construct definitions without titles and are asked to match items with a corresponding definition. An acceptable agreement index must be determined prior to administration of the items and definitions.

A more recently developed method for conducting content assessments utilizes both sorting and factor analytical techniques to quantitatively assess the content adequacy of a set of newly developed items (Schriesheim, Powers, Scandura, Gardiner and Lankau, 1993). Respondents were asked to rate the extent to which items corresponded with construct definitions. The responses were then factor- analyzed (discussed in a later section) and those items that loaded appropriately were retained for subsequent administration to an additional sample.

A third method, based on analysis of variance techniques, will be described below. This technique is very simple and straightforward and permits a statistical test of content adequacy. It can be conducted with a relatively small sample and has a very low cost both in time and in money.

None of these techniques will guarantee a content valid scale, but they will provide evidence that the items represent a reasonable measure of the construct under examination and reduce the need for subsequent scale modification. Those items that are retained from this analysis can then be used with some confidence for further data collection. If enough items are not retained then more may be generated at this stage.

Example: Content Adequacy of the MLQ

We conducted a content adequacy assessment of the current items that have been developed to assess four dimensions of the transformational leadership construct. We used Form 5-X of the MLQ, which included 39 items that are purported to measure four distinct dimensions of transformational leadership; idealized influence, individualized consideration, intellectual stimulation and inspirational motivation.

The sample used for this content adequacy assessment consisted of 57 graduate hospitality management students at a large northeastern university, all of whom had worked for several years in the industry. The average age of the students was 28, 46% were female and they had an average of seven years- of work experience. Questionnaires were administered during normal class time and took approximately 20 minutes to complete. Both verbal and

written instructions were provided prior to administration and the respondents completed the surveys anonymously.

The respondents rated each of the 39 transformational leadership items on the extent to which they believed the items were consistent with each of the four dimensions of transformational leadership. Response choices ranged from 1 (not at all) to 5 (completely). A brief description of each transformational leadership dimension was presented at the top of each page of the questionnaire, followed by the list of 39 transformational leadership items (see Appendix A for an example of this questionnaire format). Four versions of the questionnaire were administered, each with the definitions presented in a different order. This was done to control for response bias that may be due to order effects.

To determine if the items were categorized according to Bass' propositions, an analysis of variance was conducted. First, the mean score for all items on each of the four transformational leadership scales was calculated. Then, a comparison of means across the four dimensions was conducted to identify those items that were evaluated appropriately (i.e., to identify whether an item was statistically significantly higher on the appropriate definition; $p < .05$)

The results from this analysis revealed that 23 of the 39 items were classified in a manner consistent with Bass' conceptualization. These results provided some support for the proposed dimensionality of the transformational leadership construct. Three idealized influence items, four inspirational motivation items, eight intellectual stimulation items and eight individualized considerations items were judged to reflect the proposed transformational leadership dimensions. Table 1 presents the mean ratings for all items and highlights those items that were rated appropriately according to Bass' theory.

At this point in the process, the researcher retains the set of items that have been carefully devised and reviewed by experts or modified according to the results of a quantitative pretest. In the example above, this was 23 items.

Step 3. Questionnaire Administration

The retained items are then presented to an appropriate sample with the objective of examining how well those items confirmed expectations regarding the psychometric properties of the new measure. The new items should be administered with other established measures to later assess the distinction or overlap among the proposed and existing scales. These would include measures with which the new scales would be hypothesized to be strongly related and unrelated to examine discriminant, convergent and criterion-related validity, discussed in a following section. In addition, data from existing measures will later be used for preliminary examination of construct and criterion-related validity of the new scales.

Item Scaling

As previously mentioned, Likert scales are the most commonly used in survey research

Table 1
Mean Ratings from Content Adequacy Assessment

Scale:	II	IM	IS	IC
Item:				
II1	3.94	3.87	2.80	3.11
IM1	4.09	4.24	3.28	3.61
IS1	2.93	3.04	**4.63**	3.06
IC1	3.31	2.87	3.81	**4.61**
II2	3.91	4.35	3.22	3.54
IM2	3.07	**4.37**	4.30	3.19
IS2	2.89	2.83	**4.57**	3.04
IC2	3.56	3.02	3.76	**4.52**
II3	3.96	4.22	3.09	3.70
IM3	3.59	**4.63**	3.24	3.43
IS3	2.96	3.09	**4.69**	3.50
IC3	3.46	2.54	3.09	**4.59**
II4	**4.61**	2.83	2.69	3.02
IM4	3.70	**4.41**	3.35	3.94
IS4	2.61	2.81	**4.50**	2.87
IC4	3.37	2.87	3.30	**4.65**
II5	4.46	4.13	2.72	3.02
IM5	3.89	4.13	3.72	4.48
IS5	2.94	2.78	**4.56**	3.26
IC5	3.41	3.15	3.20	**4.57**
II6	**3.80**	3.04	2.57	2.72
IM6	3.26	3.19	2.54	3.96
IS6	2.91	3.02	**4.39**	3.30
IC6	3.61	2.93	3.63	**4.69**
II7	3.65	4.07	3.00	3.37
IM7	3.74	**4.48**	3.33	3.46
IS7	3.35	3.31	4.56	4.43
IC7	3.33	3.02	3.17	3.72
II8	3.65	3.43	2.74	3.33
IM8	3.22	3.11	2.94	3.35
II9	3.59	4.69	2.78	3.09
IM9	3.61	**4.70**	3.04	2.85
IS8	2.85	2.89	**4.59**	3.50
IC8	3.50	2.74	3.81	**4.50**
II10	**4.70**	3.74	3.11	3.48
IM10	4.37	3.96	3.00	3.20
IS9	2.85	2.78	**4.69**	3.17
IC9	3.57	2.85	2.93	**4.39**
IS10	3.41	3.04	3.20	3.28

Table 1 (cont'd)
Mean Ratings from Content Adequacy Assessment

Note: Items in bold and underline were rated significantly higher than other items on the appropriate dimension ($p<.05$)
Note: The number associated with each item refers to the order in which the item appeared in the questionnaire.

II = Idealized Influence - a follower's respect and admiration for the leader.
IM = Inspirational Motivation - communication of expectations and confidence in the leader's vision and values.
IS = Intellectual Stimulation - the extent the leader provides followers with interesting and challenging tasks and encourages them to solve problems.
IC = Individualized Consideration - the degree the leader shares the individual follower's concerns and developmental needs

using questionnaires (Cooket al., 1981: Schmitt and Klimoski, 1991). Likert scales include several "points" along a continuum that define various amounts or levels of the measured attribute or variable (e.g., agreement, frequency, importance etc.). An example of a seven-point Likert response format is as follows:

I offer helpful suggestions to fellow workers:

Strongly Disagree	*Disagree*	*Slightly Disagree*	*Neither Agree nor Disagree*	*Slightly Agree*	*Agree*	*Strongly Agree*
1	*2*	*3*	*4*	*5*	*6*	*7*

It is suggested that the new items be scaled using five- or seven-point Likert scales. Measures with five- or seven-point scales have been shown to create variance that is necessary for examining the relationships among items and scales and create adequate coefficient alpha (internal consistency) reliability estimates (Lissitz and Green, 1975).

Sample Size

The data must be collected from an adequate sample size to appropriately conduct subsequent analyses. There has been substantial debate over the sample size needed to appropriately conduct tests of statistical significance. It appears that the number of variables or items to be assessed will dictate the sample size needed to obtain robust results. Earlier recommendations for item-to-response ratios ranged from 1:4 (Rummel, 1970) to at least 1:10 (Schwab, 1980) for each set of scales to be factor analyzed. Recent studies have 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 and Velicer, 1988). For confirmatory factor analysis, we recommend a minimum sample size of 100 (cf., Bollen, 1989). However, we suggest that a conservative approach be adopted. As the number of items increases, it may be necessary to increase the number of respondents. As sample size increases, the likelihood of attaining statistical significance

increases, which in turn may distort the practical meaning of the results. As such, it is important to note the difference between statistical and practical significance (Cohen, 1969).

Upon completion of data gathering it is essential to evaluate the performance of the items to determine whether they adequately constitute the scale. Item evaluation through factor analysis is one of the most critical steps in determining the viability of the scale.

Step 4. Factor Analysis

There are two basic types of factor analyses available for the scale development process. The first is termed exploratory and is commonly used to reduce the set of observed variables to a smaller, more parsimonious set of variables. The second type is called confirmatory and is used to assess the quality of the factor structure by statistically testing the significance of the overall model (e.g., distinction among scales), as well as the relationships among items and scales. When using the inductive approach, exploratory factor analysis may be most helpful for identifying those items that load as predicted. For deductive studies confirmatory analysis may be most useful. Both types of analyses can be used, however, in both inductive and deductive studies. Prior to conducting the factor analysis, the researcher may find it useful to examine the inter-item correlations among the variables and any variable that correlates at less than .4 with all other variables may be deleted from the analysis (Kim and Mueller, 1978). Low correlations indicate items that are not drawn from the appropriate domain and that are producing error and unreliability (Churchill, 1979).

Exploratory Factor Analysis

A common factoring method such as principal axis is recommended because the principal-components method of analysis accounts for common, specific and random error variances (Ford, MacCallum and Tait, 1986; Rummel, 1970). The number of factors to be retained depends on both underlying theory and empirical results. There are no specific rules for retaining items, however. Eigenvalues greater than 1 (Kaiser criterion) or a scree test of the percentage of variance explained (cf., Cattell, 1966) are commonly used to determine the number of factors to retain, if the factors are assumed to be largely uncorrelated, an orthogonal rotation should be used; if the factors are assumed to be correlated, an oblique rotation should be used. It may be useful to conduct both types of analyses to determine which items to retain. However, if the intent is to develop scales that are reasonably independent of one another, more reliance should be placed on the orthogonal analyses when eliminating items.

The objective is to identify those items that most clearly represent the content domain of the underlying construct. Only those items that clearly load on a single factor should be retained. Again, 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% may serve as a minimum acceptable target. At this stage items loading inappropriately can be deleted and the analysis repeated, until a clear factor structure matrix that explains a high percentage of total item variance is obtained. Getty and Thompson (1994) provide a very good example of data reduction using exploratory factor analysis.

Confirmatory Factor Analysis

Although an exploratory factor analysis can be quite useful for assessing the extent to which a set of items assesses a particular content domain (or set of scales), a major weakness of this technique is the inability to quantify the goodness-of-fit of the resulting factor structure (Long, 1983). in addition, exploratory factor analysis involves a post hoc interpretation of the results, whereas confirmatory factor analysis specifies a priori relationships and distinctions among the scales or variables of interest. 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 and Anderson, 1988). As such, it is recommended that new scales be subjected to confirmatory factor analysis, whether or not exploratory analyses have been conducted. In scale development, confirmatory factor analysis should be just that—a confirmation that the prior analyses have been conducted thoroughly and appropriately.

Confirmatory factor analysis is a type of structural equations analysis that is designed 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 (Joreskog and Sorbom, 1993). The multi-trait model restricts each item to load only on its appropriate factor. It is recommended that confirmatory factor analysis be conducted by using the item variance-covariance matrix (Harvey et al., 1985).

There are several statistics that can be used to assess goodness-of-fit. 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 and Mclver, 1981), but the fit is considered better the closer the chi-square value is to the degrees of freedom for a model (Thacker, Fields, and Tetrick, 1989). A nonsignificant chi-square is desirable, indicating that differences between the variance-covariance matrix of the specified (i.e., *a priori)* model and the variance-covariance matrix of the observed model are small enough to be due to sampling fluctuation. In addition, it is desirable to have a significantly smaller chi-square for the specified model than for competing models. However, chi-square is quite sensitive to sample size. As such, a significant chi- square may not be problematic if additional fit indices are adequate.

In addition to chi-square, there are currently about 30 goodness-of-fit indices that can be used to assess confirmatory factor analytic results (MacKenzie, Podsakoff, and Fetter, 1991). Muliak, James, Van Alstine, Bennet, Lind, and Stilwell (1989) recommend the use of the Adjusted Goodness of Fit Index, Normalized Fit Index, and Tucker-Lewis Index to assess the correspondence between the proposed model and the data. In addition, the use of relative fit indices, such as the Comparative Fit Index, has been suggested to control for the effects of sample size. Each of these indices measures the amount of variance and covariance accounted for in the model, and values range from 0 to 1. Unlike chi-square, there is no statistical test of fit. As such, the interpretation of these indices is somewhat subjective. As a heuristic, a value over .90 indicates a reasonably good model fit (Widaman, 1985). An examination of Root Mean Square Residual may also be useful, with a value of less than 0.05 considered acceptable (Bagozzi, Yi and Phillips, 1991). The two most commonly used software packages for conducting confirmatory factor analyses are Joreskog and Sorbom's LISREL and Bentler's EQS, and each provides a fairly comprehensive list of fit indices (Bollen, 1989).

Once the overall fit of the model has been examined, additional interpretation is necessary. First, each model coefficient (e.g., item) should be individually examined for degree of fit. By selecting a desired level of significance, the researcher can use the t-values to test the null hypothesis that the true value of specified parameters is zero and determine if the items are good indicators of a particular scale. Those items that are not significant may need to be eliminated. Second, modification indices should be considered. While t-values provide an estimate of fit for specified parameters, the modification indices provide information regarding unspecified parameters or cross-loadings. A large modification index indicates that a parameter might also contribute explained, but unspecified, variance in the model.

If the output reveals large modification indices, the model should be respecified and the analysis repeated, allowing the items with the largest indices to load on the specified corresponding factor. However, these modifications should only be made if they are theoretically plausible. The output should then be re-examined, with special attention to t-values for all specified loadings. Again, there are no hard or fast rules, but the fewer modifications made to the initial model the better. If all appropriate loadings are significant at $p < .01$ or less and the magnitude and significance level of any inappropriate cross-loadings are relatively small, the researcher can be assured that the data fit the model quite well. However, if an inappropriate item demonstrates a significant loading then the item may not be tapping a single underlying construct and should be deleted and the model respecified. Performing this model respecification should result in a smaller chi-square and larger goodness-of-fit indexes. Sweeney and McFarlin (1993) provide a very good example of describing procedures and presenting results using confirmatory factor analysis.

Example: Exploratory and Confirmatory Factor Analyses of the MLQ

The content adequacy of the MLQ revealed that 23 of the 39 items appear to be consistent with Bass' propositions. These items were retained and administered to an

independent sample of 123 general managers and middle-level managers from a large U.S. hotel management organization. The average age of these respondents was 38 and 50% were females. Most of the individuals (70%) had been in their current job longer than one year and most (78%) had at least some undergraduate college experience.

Questionnaires were administered directly to 46 of the participants. An additional 140 questionnaires were distributed through the mail. Of these, 77 usable questionnaires (56%) were returned. There were no significant differences between the two sub-samples on any of the demographic variables collected for this study. Therefore, all analyses were based on a total sample of 123 cases. All participants responded on a voluntary basis and were assured that their individual responses would remain confidential. The referent leader for this study was the respondents' superior with whom they interacted on a frequent basis.

To determine whether distinctions among the four MLQ scales were justified, an exploratory factor analysis of the 23 "good" transformational leadership items was conducted. Using an oblique rotation and a principal axis method for extraction, the results yielded a 4-factor solution that accounted for 64.6% of the variance. A scree test and an eigen-value of 1.0 were used to select the number of factors, and items with factor loadings of .40 or higher on only one factor were used to define the factor. In general, the results were not very interpretable. For example, factor one consisted of 10 items and included five individualized consideration items, three intellectual stimulation items, one inspirational motivation item and one idealized influence item. In addition, factor two was defined by a single item. Factor loadings for all items are listed in Table 2.

A closer look at the results shows that the idealized influence (II) items did not load at all as predicted. Based on the strength of factor loadings, factor one consists primarily of five individualized consideration (IC) items while factor three is comprised mainly of four intellectual stimulation (IS) items. Factor four is comprised primarily of individualized consideration (IC) and inspirational motivation (IM) items. At this stage it would be recommended that the remaining items be deleted and the factor analysis repeated. It could be expected that a reasonable three-factor solution would emerge, but it is apparent that the idealized influence construct was not identified by the respondents in this sample and the factor structure proposal by Bass was not supported.

Although the exploratory factor analytical results provide some evidence of a three-factor solution, confirmatory factor analysis provides a more rigorous test of item loadings. To conduct a confirmatory factor analysis of the 23 MLQ items, another independent sample was obtained. We administered questionnaires directly to 158 fulltime employees of a large western U.S. resort hotel. As in the previous sample, the referent leader was the respondent's direct superior with whom they interacted on a daily basis. The average age of these respondents was 39 and 41% were females. Most of the individuals (63%) had been in their current job longer than one year and most (60%) had at least some undergraduate college experience.

Table 2
Factor Loadings from the Exploratory Factor Analysis of 23 Content Adequate MLQ Items

	Factor 1	Factor 2	Factor 3	Factor 4
IC5	.86			
IC3	.85			
IC4	.81			
IS6	.74			
IC9	.70			
IC8	.70			
IM9	.69		.33	
IS9	.59			
IS8	.54	.27		
II6	.53			
IS4		.87		
IS2			.85	
IS1			.64	
IS3	.29	.28	.57	
II4		-.40	.46	
IS5			.43	.31
IC1				.83
IC6				.69
IC2				.65
IM7				.61
IM4				.46
II10			.30	.44
IM3			.35	.37

Note: The number associated with each item refers to the order in which the item appeared in the questionnaire.
Note: Factor loadings <.25 are omitted.

The confirmatory factor analysis of the 23 items was conducted using LISREL 8.03 (Joreskog and Sorbom, 1993). The fit of the four-factor model (i.e., multi-trait model to determine whether the four scales are indeed distinct) was evaluated using the sample variance-covariance matrix as input and a maximum likelihood solution. The overall chi-square was statistically significant ($X^2=532.28$; $df=224$; $p<.01$), the Goodness of Fit Index was 0.76, the Adjusted Goodness of Fit Index was 0.71, the Normed Fit Index was 0.78, the Comparative Fit Index was 0.86 and the Root Mean Square Residual for the predicted minus observed correlation matrices was 0.09. As these indices were not within the range of conventionally accepted values (cf., Bollen, 1989), the four-factor model was not supported.

However, modification indices for the lambda matrix (i.e., a matrix that indicates which of the observed variables or items serve as indicators of the latent variables or scales) suggested that fit could be improved. One approach that can be taken to enhance model fit is to eliminate items that load on multiple factors. According to Medsker, Williams and Holohan (1994), values less than four are acceptable for defining a factor, while values higher than five indicate that the items are loading on multiple factors and that error terms may be correlated. The modification indices showed that 12 items exceeded the suggested cutoff. Using the

criteria suggested by Medsker et al. (1994), all three idealized influence items were eliminated, one inspirational motivation item was eliminated, five individualized consideration items were eliminated and three intellectual stimulation items were eliminated. The remaining three factors were defined by 11 items: three inspirational motivation items; three individualized consideration items; and five intellectual stimulation items.

Results from the confirmatory factor analysis of the revised scales provided strong support for a three-factor model. Using the sample variance-covariance matrix as input and a maximum likelihood solution, the overall chi-square was statistically non-significant ($X^2 = 62.86;\ df = 41;\ p > .01$), the Goodness of Fit Index was .93, the Adjusted Goodness of Fit Index was 0.90, the Normed Fit Index was .93, the Comparative Fit Index was .98 and the Root Mean Square Residual for the predicted minus observed correlation matrices was .05. All these values suggest good model fit for the three factor model. In addition, each item was a good indicator of the corresponding scale (i.e., all had significant t-values) and all modification indices were low. Our results do not confirm that the items devised by Bass constitute the four scales as he intended.

The factor analysis step helps to determine how many factors or subscales exist for a set of items. In this example, neither the exploratory nor the confirmatory factor analyses provided empirical support for the four-factor typology proposed by Bass, but did support a very similar three factor solution in both samples. As such, these three factors can be accepted with some confidence as representing the constructs under examination.

Step 5. Internal Consistency Assessment

After unidimensionality of each scales has been established (Gerbing and Anderson, 1988). Reliability may be calculated in a number of ways, but the most commonly accepted measure in field studies for assessing a scale's internal consistency is Cronbach's alpha which tells how well the items measure the same construct (Price and Mueller, 1986). After the exploratory and confirmatory factor analyses have been conducted and all "bad" items have been deleted, the internal consistency reliabilities for each of the scales should be calculated. A large coefficient alpha (.70 for exploratory measures; Nunnally, 1978) provides an indication of strong item covariance or homogeneity and suggests that the sampling domain has adequately been captured (Churchill, 1979). If the number of retained items at this stage is sufficiently large, the researcher may want to eliminate those items that do not share equally in the common core dimension by deleting items that will improve or not negatively impact 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 impact on the reliability and may in fact reduce the reliability if the additional items lower the average inter-item correlation. Most statistical software packages produce output that provides reliabilities for scales with individual items removed. Reporting internal consistency reliability should be considered absolutely necessary.

Example: Internal Consistency of the Revised MLQ

The reliability analysis showed that the revised MLQ scales, based on the confirmatory factor analysis, had good internal consistency (.81 to .87). It should be emphasized that even though two of the revised scales included only three items each, the content adequacy assessment and factor analyses helped retain items that were consistent with the corresponding construct domain.

The strong internal consistency reliability for the revised scales tells us that the retained items measure the same constructs. Reliability testing is critical for new scale development before you attempt to draw inferences based on a scale. If a scale has low reliability it may be necessary to add or reexamine the existing items. The sixth step is to examine validity or how well a scale measures what it says it is measuring.

Step 6. Construct Validation

At this point, the new scales should demonstrate content validity (see Step 2) and internal consistency reliability (see Step 5), both of which provide supportive evidence of construct validity. 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 is also useful to examine relationships with variables that are theorized to be outcomes of the focal measure (criterion-related validity).

Example: Convergent and Criterion-Related Validity of the MLQ

To assess the convergent and criterion-related validity of the MLQ, we gathered some additional data from our third sample (N=158). For the convergent validity assessment, we gathered information on four scales from the Managerial Practices Survey (MPS) (Yukl, 1990). The first scale, clarifying, focuses on task assignment and providing direction in how to complete work. The second scale, inspiring, is based on influence techniques that appeal to emotion or logic to generate enthusiasm for work. The third scale, supporting, includes behaviors such as listening to complaints and looking out for someone's best interests. The fourth scale, team building, focuses on cooperation, teamwork and constructive conflict resolution. The items asked respondents to indicate the extent to which their immediate supervisor demonstrated the behavior described and each scale had five to six items. The response choices ranged from 1 (never, not at all) to 4 (usually, to a great extent).

Conceptually, these MPS scales appear to be quite similar to the defining elements of the MLQ scales. For example, inspirational motivation (one of the four MLQ scales) was defined as behaviors that communicate expectations (clarify) and create a team spirit (team building) through enthusiasm (inspiring). However, while there are several similarities, it also appears that the MLQ and MPS assess distinct constructs, most important of which is the distinction

between leadership and managerial practices. As such, the MPS serves as an appropriate convergent validity referent.

For the criterion-related validity assessment, we collected information on two relevant outcome variables. The first outcome variable was subordinate ratings of satisfaction with their leader. This outcome was assessed using the nine-item scale from the *Job Description Index* (Smith, Kendall and Hulin, 1969) which asked respondents to rate the extent to which they were satisfied with their leader. The response choices ranged from 1 (very dissatisfied) to 5 (very satisfied). The second outcome variable was subordinate ratings of leader effectiveness. This outcome was assessed using the six-item scale developed by Hinkin and Tracey (1994). The response choices ranged from 1 (highly ineffective) to 7 (highly effective).

The results from the convergent validity analysis showed moderately high correlations among the MLQ and MPS scales (.55 to .69; all p<.01). These findings support the expected conceptual overlap between the MLQ and MPS. However, because less than 50% of the variance is accounted for by any single correlation, the scales appear to assess distinctive constructs. The results from the criterion-related validity analysis showed significant correlations between each of the revised MLQ scales and the satisfaction with leader and leader effectiveness scales (.58 to .75, all $p < .01$).

Overall, the convergent and criterion-related validity analyses provide further support for the construct validity of the revised MLQ scales. Descriptive statistics, internal consistency reliability estimates and correlations among all scales used for the convergent and criterion-related validity analyses (except the performance appraisal indicators) are listed in Table 3. While the correlations of the MLQ scales with other variables helps us to see the degree to which this scale is related to established measures, the process of determining construct validity is difficult and on-going because it is the result of continual use of a scale in different settings. The final step in our scale development process is replication.

Step 7. Replication

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. 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 (Krzystofiak, Cardy and Newman, 1988). To avoid the common source problem, it is recommended that data from sources other than the respondent, such as performance appraisals, be collected where possible. The replication should include confirmatory factor analysis, assessment of internal consistency reliability and construct validation. 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.

Table 3
Means, Standard Deviations, Internal Consistency Reliability Estimates and Inter-Correlations for all Measures

	Mean	SD	a	1	2	3	4	5	6	7	8
1. IM	2.14	1.06	.84								
2. IS	2.11	.83	.81	.64							
3. IC	1.62	1.10	.85	.64	.67						
4. CL	3.11	.76	.93	.58	.63	.64					
5. IN	2.86	.72	.88	.69	.68	.68	.72				
6. SU	3.04	.78	.90	.55	.63	.64	.71	.68			
7. TE	3.04	.75	.89	.62	.69	.64	.69	.71	.80		
8. SA	3.45	.95	.93	.58	.60	.68	.69	.74	.73	.72	
9. LE	5.01	1.32	.90	.63	.64	.75	.72	.73	.77	.73	.80

Note: All correlations are significant at $p<.01$

IM = Inspirational Motivation (revised)
IS = Intellectual Stimulation (revised)
IC = Individualized Consideration (revised)
CL = Clarifying
IN = Inspiring
SU = Supporting
TB = Teambuilding
SA = Satisfaction with Leader
LE = Leader Effectiveness

Conclusion

We have provided a seven-step process guide for scale development and analysis in the hopes that hospitality researchers will utilize a systematic approach to item and scale creation. As the quantity of empirical industry-specific research increases, we need to ensure the quality of research to instill confidence in the results by both academic and practitioner audiences. Good research begins with good measurement. The example of the MLQ shown in this paper illustrates both the techniques used in scale development and the problems that can arise if new measures are not given serious psychometric examination. Poor scale construction brings into question the reliability and validity of the research results, no matter how careful the design of the study. In contrast, carefully constructed measures help to advance our

understanding and ensure that the study will provide accurate and usable data. By using the seven steps suggested, a researcher more likely can create scales that will provide critical information and enhance the future of hospitality research.

References

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

Bass. B.M. (1985). *Leadership and performance beyond expectations.* New York: Free Press.

Bass, B.M. & Avolio, B.J. (1994). *Improving organizational effectiveness through transformational leadership.* Thousand Oaks, CA: Sage.

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: John Wiley & Sons, Inc.

Bums, J.M. (1978). *Leadership.* New York: Harper and Row.

Carman, J.M. (1990). Consumer perceptions of service quality: An assessment of the SERVQUAL dimensions. *Journal of Retailing, 66,* 33-55.

Carmines, E.G. & Mclver, J. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. Bohmstedt and E. Borgatta (eds.) *Social measurement: Current issues.* Beverly Hills: Sage.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and validity assessment* Beverly Hills: 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., Wail, T.D. & Warr, P.B. (1981). *The experience of work.* San Diego: Academic Press.

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.

Getty, J.M. & Thompson, K.N. (1994). A procedure for scaling perception of lodging quality. *Hospitality Research Journal, 18,* 75-96.

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

Gundersen, M.G., Heide, M. & Olsson, U.H. (1996). Hotel guest satisfaction among business travelers. *Cornell Quarterly, 376(2),* 72-81.

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.

Hinkin, T.R. (1995). A review of scale development practices in the study of 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 (4),* 561-567.

Hinkin, T.R. & Tracey, J.B. (1994). Transformational leadership in the hospitality industry. *Hospitality Research Journal, 18,* 49-63.

Joreskog, K.G. SSorbom, D. (1993). *LISREL8.03.* Morresville, IN: Scientific Software, Inc.

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

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

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: Sage Publications.

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.

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.

Muliak, S., James, L., Van Alstine, J., Bennet, N., Lind, S. & Stilwell, C. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430-445.

Nunnally, J.C. (1978). *Psychometric theory (2nd ed.).* NY: McGraw-Hill.

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

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, South-Western Publishing.

Schmitt, N.W. & Stults, D.M. (1985). Factors defined by negatively keyed items: The results of careless respondents? *Applied Psychological Measurement, 9,*367373.

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 adequacy of paper-and-pencil survey-type instruments. *Journal of Management, 19,* 385417.

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

Smith, P.C., Kendall, L.M. & Hulin, C.L. (1969). *The measurement of satisfaction with work and retirement.* Chicago: Rand McNally.

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

Sweeney, P.D. & McFarlin, D.B. (1993). Workers' evaluation of the "ends" and the "means": An examination of four models of distributive and procedural justice. *Organizational Behavior and Human Decision Processes, 55,* 23-40.

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.

Tracey, J.B. & Hinkin, T.R. (1994). Transformational leaders in the hospitality industry. *Cornell Hotel and Restaurant Administration Quarterly, 35,* 18-24.

Tracey, J.B. & Hinkin, T.R. (1996). How transformational leaders lead in the hospitality industry. *International Journal of Hospitality Management,* 165-176.

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

Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1-26.

Yukl, G.A. (1990). *COMPASS: The Managerial Practices Survey.* New York: Gary Yukl and Manus Associates.