An Analysis of Variance Approach to Content Validation

TIMOTHY R. HINKIN
J. BRUCE TRACEY
Cornell University

Although procedures for assessing content validity have been widely publicized for many years, Hinkin noted that there continue to be problems with the content validity of measures used in organizational research. Anderson and Gerbing, and Schriesheim, Powers, Scandura, Gardiner, and Lankau discussed the problems associated with typical content validity assessment and presented techniques that can be used to assess the empirical distinctiveness of a set of survey items. This article reviews these techniques and presents an analysis of variance procedure that can provide a higher degree of confidence in determining item integrity and scale content validity. The utility of this technique is demonstrated by using two samples and two different measures.

In the social sciences, a clear and comprehensive understanding of any phenomena is established in part by the psychometric quality of measures that are used for inquiry. One key indicator of quality is content validity. This type of validity, defined as the extent to which a measure's items reflect a particular theoretical content domain (Kerlinger, 1986), is a necessary precondition for establishing evidence for construct validity. Unfortunately, although the importance of content validity has been vigorously emphasized over the last several decades (e.g., Barrett, 1972; Cook, Hepworth, Wall, & Warr, 1981; Schmitt & Klimoski, 1991), many researchers have failed to use or document the procedures for assessing an instrument's content validity (Hinkin, 1995). This situation is somewhat perplexing. Assessing evidence of content validity does not necessarily require complicated, cumbersome analytical analyses or huge samples. Rather, the process can be quite straightforward and provides an efficient means for establishing and interpreting the utility of any measure.

For this article, we begin by reviewing two recent pretesting approaches to content validity assessment. We argue that although these approaches have utility, they have some methodological limitations. We then present the results from two studies that utilize a third technique that provides a simple, yet direct assessment of content validity. This technique is based on an analysis of variance (ANOVA) approach and reduces the subjective decision-making requirements that are characteristic of other types of

Organizational Research Methods, Vol. 2 No. 2, April 1999 175-186

content validity/adequacy assessment. We conclude by encouraging more focused attention on the issue of content validation.

Anderson and Gerbing's Substantive Validity

Anderson and Gerbing (1991) proposed a pretesting procedure for establishing a measure's *substantive validity*, a type of content validity defined as the extent to which a measure is judged to be reflective of, or theoretically linked to some construct of interest. They used two indices for assessing the content validity of a measure: the substantive agreement (SA) index and the substantive validity (SV) index. The SA index reflects the proportion of respondents who assign an item to its intended construct. This index is quite similar to those used in traditional Q-sort procedures (cf., Nunnally, 1978) and the authors suggest it can be used with small sample sizes ($N = 20$). The SV index is an extension of Lawshe's (1975) method for assessing substantive validity. This validity index measures the extent to which respondents assign an item to its posited construct more than to any other construct. For the SV index, Anderson and Gerbing (1991) suggested that a binomial test be conducted to determine whether an item significantly assesses one construct more than it does any other. This test simply involves an assessment of the probability that an item is properly assigned to its posited construct.

Anderson and Gerbing (1991) presented a two-step procedure to support the use of the SA and SV indices. The first step involved a confirmatory factor analysis of data from 379 respondents who completed a 35-item questionnaire that purportedly measured five first-order personality constructs, composed of 7 items each. Based on the factor loadings, each of the 35 items was then classified into one of four categories: high, moderate, ambiguous, or useless. These categories reflected judgments regarding the extent to which items assessed the posited construct and were used as a basis for making comparisons with the results from the subsequent substantive validity analysis. For the second step, Anderson and Gerbing administered the same 35 items to two student samples of 20 respondents each. This survey administration involved a sorting task in which the 35 items were assigned to one of five categories representing each of the proposed theoretical dimensions. All items were listed on one page of the survey, and construct labels and a one-sentence definition of each of the five dimensions appeared on a separate page. The respondents were asked to read each item and then assign it to the most applicable construct definition. Upon completion of this sorting task, respondents were given the opportunity to reclassify any item. SA and SV indices were then computed for each sample.

To compare the results from the confirmatory factor analysis and the SA and SV indices, three types of analyses were conducted. First, zero-order correlations between the item factor loadings from the confirmatory factor analysis results (based on the responses to the 35-item questionnaire) and the SA and SV values from the two student samples were calculated. Second, a one-way analysis of variance (ANOVA) and follow-up *t*-tests were used to compare SA and SV index differences across the four categories created from the initial confirmatory factor analysis results. Finally, signal-detection matrices (Green & Swets, 1966) were computed to assess the relationship between items with significant ($p \geq .05$) and nonsignificant SV ($p < .05$) values. Items judged to be retained or deleted based on SV values were correlated with

those items judged to be high or moderate and ambiguous or useless, respectively, in the confirmatory factor analysis described above. In summary, the results demonstrated a high degree of convergence between the confirmatory factor analysis and the SA and SV indices. There was strong agreement between the samples on item retention and deletion, and high correlations among the indices.

Although the SA and SV indices appear to be quite relevant for establishing content validity, there are some concerns. The authors used an ipsative, or forced-choice, response format, which does not take into account the extent to which an item may correspond to a given dimension and may bias results (Mehrens & Lehman, 1978; Schriesheim, Hinkin, & Podsakoff, 1991). The analyses of variance and follow-up *t*-tests were conducted using inductively derived classification categories (e.g., high, moderate, ambiguous, and useless) as the grouping (i.e., independent) variable, and not the theoretically-posited dimension. As such, no specific information was generated regarding differences or similarities among individual items and the underlying conceptual domain. Finally, the lack of significant differences among item classifications suggests a reliance on heuristics rather than statistics to categorize items.

Schriesheim, Powers, Scandura, Gardiner, and Lankau's Content Adequacy

Similar to Anderson and Gerbing (1991), Schriesheim, Powers, Scandura, Gardiner, and Lankau (1993) argued that content validity is an important first step in the construct validation process, and should be assessed immediately after a measure has been developed but prior to utilization in a research study. To address concerns about the subjective nature of traditional content validity procedures, Schriesheim et al. described two sorting procedures and two types of factor analyses that can be used to assess the empirical distinctiveness of items that measure proposed theoretical dimensions.

To examine the content *adequacy* (a term similar to, but distinct from, content validity) of an existing measure, Schriesheim et al. (1993) used a set of 20 items from the Minnesota Satisfaction Questionnaire (MSQ-S; Weiss, Dawis, England, & Lofquist, 1967) that had been developed to assess multiple dimensions of job satisfaction. All items were listed on each of three pages of a questionnaire, with a definition of intrinsic satisfaction, extrinsic satisfaction, and neither intrinsic nor extrinsic satisfaction as headings, each on an individual page. In contrast to Anderson and Gerbing, however, Schriesheim et al. did not use an ipsative sorting procedure. Instead, respondents (N = 150 M.B.A. students) were asked to rate each item on a Likert-type scale to indicate the extent to which the items corresponded to each construct definition. The *neither* category was eliminated from further analysis as none of the items had their highest mean in that category.

Schriesheim et al. (1993) first computed a Q-correlation matrix (item by item) of the data. This matrix was then subjected to a principal components analysis, extracting the number of factors corresponding to the theoretical dimensions under examination. Those items that met Ford, MacCallum, and Tait's (1986) heuristic for retention (.40 or greater on the appropriate factor with no major cross-loadings) were judged as meaningful and representative of the construct under examination. As a second approach, Schriesheim et al. computed correlations among the items that were included in the

extended matrix of ratings (i.e., across respondents and items), and then conducted a principal axis factor analysis using squared multiple correlations as initial estimates of communality. Both of the factor analyses yielded identical results and demonstrated that 17 of 20 MSQ items loaded exclusively on the posited dimension. Schriesheim et al. also gathered and analyzed data from a second sample (N = 67 undergraduate business students) using Q-factor analysis and found similar results, suggesting that several MSQ items have been theoretically misclassified.

The factor analytic procedures suggested by Schriesheim et al. (1993) have made an important contribution to the scale development process. Their approach focuses on the relative adequacy of each item, as well as the correspondence between items and the posited theoretical constructs. However, it stops short of providing a true statistical test of an item's content validity, primarily due to the subjective criteria that are often employed to determine factor and item retention. For example, a scree plot may suggest that five factors be used to define the dimensionality of a particular item set, whereas a Kaiser criterion may suggest that up to seven factors be retained. At this point, the researcher has to make a judgment regarding the number of factors to retain (i.e., use the scree plot or Kaiser criterion) and about item loadings. Unfortunately, this type of judgment relies on heuristics and/or convention such as "positive and meaningful loadings" (Schriesheim et al., 1993, p. 400), and subsequently introduces a degree of uncertainty to the interpretation and meaning of the focal construct(s). In addition, factor analytic techniques typically require larger sample sizes to achieve an adequate respondent-to-item ratio. Although sample size is not an inherent concern from a methodological standpoint, there may be administrative difficulties in obtaining enough data to yield robust results.

The Current Study

The current study builds on the work described above and presents the use of an analysis of variance technique that can add a higher degree of confidence in item integrity and scale content validity. This procedure has several advantages over other analyses. First, it virtually eliminates the use of subjective judgment for item retention. Analysis of variance provides a direct empirical test for determining item distinctiveness, and the only judgment call concerns the p value for determining significance. Second, this technique can be used with small sample sizes. The Central Limit Theorem holds that a sample size of 30 is usually sufficient to obtain a normal sampling distribution (Agresti & Agresti, 1979), however factor analytical techniques typically require much larger samples. In addition, the use of small samples provides a more conservative means of distinguishing practical significance from statistical significance (Runkel & McGrath, 1984; Schmitt & Klimoski, 1991; Stone, 1978). Using small samples may result in the elimination of a few false negative items that might be retained using factor analytic procedures, however, it would be much more difficult to retain a false positive item, a far worse consequence (Anderson & Gerbing, 1991). Third, it is very simple and straightforward as the analysis involves only one procedure. We will demonstrate the utility of this methodology by comparing the results of the Schriesheim et al. (1993) technique with those obtained from analysis of variance, using the same data from two different samples.

Study 1

Measure

For Study 1, we used the Multifactor Leadership Questionnaire (MLQ), Form 5-X, developed by Bass and Avolio (1990). This form includes 39 items that purportedly measure four dimensions of transformational leadership: idealized influence, individualized consideration, intellectual stimulation, and inspirational motivation.

Sample and Procedure

The sample consisted of 57 graduate business students at a large northeastern university. The average age of the students was 28, 46% were female, and they had an average of 7 years of work experience. As noted in Schriesheim et al. (1993), the requirements to complete a task such as this are sufficient intellectual ability to rate the correspondence between items and definitions of various theoretical constructs, and the lack of any pertinent biases. As such, the use of college students was deemed appropriate. The researchers administered questionnaires during normal class time, taking approximately 15 minutes to complete. Explicit written and verbal instructions were provided prior to administration, and the respondents were asked not to sign their names.

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*). The definition of one of the four transformational leadership dimensions was presented at the top of each page of the questionnaire, followed by a randomized listing of all transformational leadership items. Four versions of the questionnaire were administered, each with the definitions presented in a different order, to control for response bias that may occur from order effects. No statistically significant differences among responses across the versions were found. Extreme care was taken to ensure that the definitions were consistent with Bass and Avolio's (1990) conceptualization of the four transformational leadership dimension.

Factor Analysis

Consistent with Schriesheim et al. (1993), the first step was to calculate an item-by-item Q-correlation matrix. The matrix was then subjected to a principal components analysis. Four factors were extracted and then subjected to a varimax rotation. (Note: The results yielded six eigenvalues greater than 1.0: 10.89, 7.48, 4.97, 2.07, 1.08, and 1.05. However, both theoretical parsimony and a scree test suggested retaining only four factors.) These factors explained 65.2% of the total item variance. Item loadings are presented in Table 1.

The results from this Q-factor analysis showed that the individualized consideration (IC) and intellectual stimulation (IS) dimensions emerged clearly. However, the idealized influence (II) and inspirational motivation (IM) dimensions were confounded. Based on conventional heuristics for interpreting exploratory factor analysis (Ford et al., 1986), the appropriate action would be to retain Factor 2, keeping all but

Table 1
Item Loadings From the Factor Analysis of the Transformational Leadership Items

Scale	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*
IC4	.87520			
IC9	.83621			
IC3	.83485			
IC6	.82850			
IC2	.82351			
IC5	.82036			
IC1	.80902			
IC8	.74787			
IS7	.60150	.53604		
IM5	.58213		.43641	
IC7	.54373			.31568
IM6	.51415			
IM8	.40169		.36464	.33694
IS9		.89624		
IS2		.88700		
IS3		.87814		
IS1		.86655		
IS4		.86435		
IS5		.85686		
IS8		.84986		
IS6		.83326		
IM3			.84929	
IM7			.84169	
II9			.80896	.32160
IM9			.79967	
IM4	.30211		.77132	
II7			.65705	.37975
II3			.62895	.32312
IM2		.56047	.57768	
II2			.54195	.41859
IM1			.49201	.43983
II8	.38572		.40299	.39039
II10				.80811
II4				.78357
II5			.31076	.71003
II6				.70304
IM10			.47800	.69576
II1				.54480
IS10	.32463	.31950		.39331

Note. Only item loadings of .30 or higher are listed. II = idealized influence; IM = inspirational motivation; IS = intellectual stimulation; and IC = individualized consideration.

one of the IS items (IS10), and deleting the four non-IC items that loaded on Factor 1 and the two non-II items from Factor 4. Factor 3, although more confounded, includes primarily IM items, four of which could be retained. Therefore, IC would then be comprised of 9 items, IS comprised of 8 items, IM comprised of 4 items, and II comprised of 5 items, for a total of 26 items in the measure.

Analysis of Variance

As an alternative to making item retention and deletion decisions, an analysis of variance (ANOVA) procedure was employed using the same data. A one-way ANOVA provides a direct method for assessing an item's content validity by comparing the item's mean rating on one conceptual dimension to the item's ratings on another comparative dimension. Thus, it can be determined whether an item's mean score is statistically significantly higher on the proposed theoretical construct. ANOVA provides a robust assessment of item distinctiveness because it is tolerant of moderate departures from normality and unequal variances, particularly if cell sample sizes are equal (Agresti & Agresti, 1979). In addition, concerns regarding type I error rates are addressed by using Duncan's Multiple Range Test, which provides simultaneous comparisons by holding the probability of making a type I error for the entire set of comparisons to the a priori α (that is, the confidence coefficient that applies to the entire set of comparisons is $1 - \alpha$).

The data were formatted such that each case included four lines of data that listed the item ratings for each of the transformational leadership dimensions. In addition, a dummy variable (in this case, 1, 2, 3, or 4) was inserted at the end of each line of data to provide a grouping code for each dimension. Then, one-way ANOVA and Duncan's multiple comparison tests (using SPSS 8.0 for Windows), were conducted to compare mean item ratings across the four dimensions (i.e., four "groups") to identify items that were statistically significantly higher on the appropriate definition (i.e., consistent with the proposed theoretical construct).

It should be noted that this procedure differs from that used by Anderson and Gerbing (1991), who compared the validity indices across the four classification categories that were derived from a confirmatory factor analysis of a previously collected data. For this study, we compared item means across the theoretically based dimensions.

The results from this analysis revealed that 23 of the 39 items were classified correctly. Three items (IC7, II5, II1), which would have been judged as acceptable by standard factor-analytical heuristics, did not have statistically significantly higher means on the appropriate dimensions. Thus, it appears that at least 3 items that might have been retained using factor analytic procedures do not possess adequate distinctiveness using significance criteria of .05. Item means are presented in Table 2.

Study 2

Measure

For Study 2 we created a new teaching evaluation instrument and conducted the same analyses that were used in Study 1. To develop this measure, we conducted a review of the teaching evaluation literature (e.g., Arreola, 1995) and examined existing surveys used at other universities to identify the dimensions of teaching effectiveness that might be included in our new measure. From this review we selected five dimensions of teaching performance: mastery of content (M), pedagogical organization (P), quality of feedback (F), quality of delivery (D), and learning outcomes (O). We then borrowed or generated 10 items for each dimension and administered the 50-item survey to 50 members of our faculty. We wanted faculty input to help us design an

Table 2
Mean Ratings From Content Adequacy Assessment for Study 1

Scale	*II*	*IM*	*IS*	*IC*
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

Note. Italicized items were rated significantly higher than other items on the appropriate dimension. The number associated with each item refers to the order in which the item appeared in the survey. II = idealized influence; IS = intellectual stimulation; IM = inspirational motivation; and IC = individualized consideration.

instrument that reflected the school's values and to obtain support for the new measure, as suggested by Arreola (1995). Using a 5-point Likert-type scale, respondents were instructed to rate those items they felt were most important for measuring teaching effectiveness at this school. Based on 36 responses, we retained the 5 most highly rated items on each dimension. It should be noted that an error was made in the construction of the final questionnaire used in this study and, as a result, the outcomes dimension

was comprised of 6 items, whereas the delivery dimension was composed of 4 items. As suggested in Hinkin (1998), more items than would be used in the final scale should be created for testing the psychometric properties of a measure. We were striving to generate a 15-item measure, with 3 items per effectiveness dimension.

Sample and Procedure

The sample consisted of 173 full-time undergraduate business students at a large northeastern university. Questionnaires were administered by several faculty members in their own classes and took approximately 15 minutes to complete. Explicit verbal and written instructions were provided prior to administration, and anonymity was assured.

Respondents rated each of the 25 teaching evaluation items on the extent to which they believed the items were consistent with each of the five teaching dimensions. Response choices ranged from 1 (*not at all*) to 5 (*completely*). The definition of one of the five effectiveness dimensions was presented at the top of each page of the questionnaire, followed by a list of all items. Two versions of the questionnaire were administered to avoid order effects, each with the definitions presented in a different order. The results revealed no differences in item means between the two versions.

Factor Analysis

As in Study 1, the first step was to calculate an item-by-item Q-correlation matrix. This matrix was then subjected to a principal components analysis. Five factors were extracted and then subjected to a Varimax rotation. These factors explained 72.9% of the total item variance and showed strong support for the proposed dimensionality. Twenty-one items would have clearly met the Ford et al. (1986) criteria for retention. However, for three items (O2, O4, and F4), the decision to retain or delete would have been quite subjective. One item (D4) did not load on the appropriate factor. Item loadings are presented in Table 3.

As in Study 1, the mean score for each item on each of the five teaching dimensions was calculated. Then, a one-way analysis of variance and Duncan's multiple range test was used to compare item means across the five dimensions.

The results from this analysis were consistent with those from the factor analysis, however, items O2, O4, D4, and F4 failed the multiple range test and were shown to not be statistically significantly different from at least one other item mean. Although these results correspond with those from the factor analysis, there is now a *statistical* basis for item retention or deletion, rather than a *judgmental* basis. Item means are presented in Table 4.

It should be noted that one of the benefits in conducting a pretest assessment of a measure's content adequacy is the ability to use small samples prior to a major data collection (Anderson & Gerbing, 1991; Schriesheim et al., 1993). To address this issue, we randomly selected two subsamples of 44 from the student population of 173. The analysis of variance previously described was repeated on each of the samples independently. The results were almost identical. For both samples, only one item (M1) failed to retain its distinctiveness.

Table 3
Item Loadings From the Factor Analysis of the Teaching Evaluation Items

Scale	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*	*Factor 5*
P4	.86634				
P3	.83158				
P1	.81529				
P5	.80957				
P2	.79979				
D4	.51083				.46989
O6		.89531			
O3		.88033			
O5		.88623			
O1		.83924			
O2		.60459			.50657
O4		.48474	.40172		
F2			.90584		
F5			.88556		
F3			.87746		
F1			.84554		
F4	.46444		.53830		
M3				.88038	
M5				.82804	
M2				.81925	
M4				.77370	
M1	.32281			.70550	.32181
D2					.88437
D1					.88120
D3					.79947

Note. Only item loadings of .30 or higher are listed. P = pedagogical organization; O = learning outcomes; F = quality of feedback; M = mastery of content; and D = quality of delivery.

Discussion

The first response that the reader might have is, "This is so simple!" We strongly concur. In a time when statistics are becoming more sophisticated and computer software more complicated, it is easy to become enamored with the elegance of a research design or the complexity of statistical analysis. When this happens, we can lose sight of the fundamentals of sound research principles. Without accurate measurement even advanced statistical techniques will not allow researchers to draw appropriate conclusions.

The purpose of this article was to build on the work of Anderson and Gerbing (1991) and Schriesheim et al. (1993) by presenting a process for quantitatively assessing the content adequacy of a measure. Several claims were made about the benefits of the proposed procedure that merit some discussion. First, the use of small sample sizes is advantageous both because of convenience and also for statistical purposes. Based on the results of the current study, a sample size of 50 would appear to be adequate for this type of analysis. With respect to the use of students, Schriesheim et al. (1993) point

Table 4
Mean Ratings From Content Adequacy Assessment for Study 2

Scale	*Mastery*	*Outcome*	*Pedagogy*	*Feedback*	*Delivery*
O1	2.87	*4.50*	2.91	2.17	2.79
D1	2.49	2.62	2.26	1.75	*4.69*
P1	2.14	2.15	*4.34*	2.49	2.35
M1	*3.97*	2.81	3.51	1.89	3.66
O2	2.62	3.71	2.14	2.12	3.65
D2	2.67	2.30	2.36	1.84	*4.33*
M2	*4.17*	2.99	2.94	1.69	3.08
P2	2.44	2.23	*4.19*	2.83	2.28
O3	2.76	*4.49*	2.61	2.16	2.60
O4	2.24	3.23	2.35	3.04	2.64
P3	2.67	2.37	*4.65*	2.03	2.77
M3	*4.43*	2.87	2.81	1.82	2.85
D3	3.12	2.32	2.38	1.91	*4.20*
F1	1.92	1.89	1.96	*3.85*	1.83
D4	3.03	2.87	3.93	2.32	3.80
O5	2.69	*4.61*	2.41	2.35	2.44
M4	*4.18*	3.49	2.81	2.03	3.32
F2	1.78	1.86	1.95	*4.40*	1.84
P4	2.50	2.33	*4.66*	2.21	2.67
F3	2.22	2.45	2.40	*4.64*	1.93
P5	2.80	2.79	*4.72*	2.38	3.39
F4	2.42	2.37	3.31	3.31	2.29
O6	2.76	*4.55*	2.31	2.34	2.53
F5	2.27	2.57	2.14	*4.61*	1.99
M5	*4.81*	2.60	2.57	1.95	2.77

Note. Italicized items were rated significantly higher than other items on the appropriate dimension. The number associated with each item refers to the order in which the item appeared in the questionnaire. P = pedagogical organization; O = learning outcomes; F = quality of feedback; M = mastery of content; and D = quality of delivery.

out that this type of judging process requires only that respondents are not biased and possess sufficient intellectual ability to perform the item rating tasks. As such, university students are very appropriate for completing this task. We would point out, however, the importance of explicit instructions to assure that respondents understand the nature of the task. The elimination of the use of subjective judgment for item retention is perhaps the most important contribution of this analysis. The use of statistical criteria can assist researchers in making important decisions when developing or evaluating measures. Finally, because of its simplicity, it is likely that this type of analysis will be more appealing to, and hopefully used by, researchers. The immediate access to respondents and straightforward analytic procedure should encourage researchers to conduct content adequacy assessments.

As noted by Hinkin (1995), many measures have been used in survey research that later are found to be flawed, rendering the results of this research questionable. Schriesheim et al. (1993) argued that "the demonstration of instrument content adequacy be demanded as an initial step toward construct validation by all studies which use new, modified, or previously unexamined measures" (p. 385). It is hoped that the procedure presented in this article will make it easier for researchers to satisfy this demand.

References

Agresti, A., & Agresti, B. F. (1979). *Statistical methods for the social sciences*. San Francisco: Dellen.

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.

Arreola, R. A. (1995). *Developing a comprehensive faculty evaluation system*. Bolton, MA: Anker.

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

Bass, B. M., & Avolio, B. J. (1990). *Multifactor leadership questionnaire*. Palo Alto, CA: Consulting Psychologists Press.

Cook, J. D., Hepworth, S. J., Wall, 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.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley.

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

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104-121.

Kerlinger, F. N. (1986). *Foundations of behavioral research*. New York: Holt, Rhinehart, & Winston.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

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

Runkel, P. J., & McGrath, J. E. (1984). *Research on human behavior: A systematic guide to method*. New York: Holt, Rinehart, & Winston.

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

Schriesheim, C. A., Hinkin, T. R., & Podsakoff, P. M. (1991). Can ipsative and single-item measures produce erroneous results in field studies of the five French and Raven bases of power? An empirical investigation. *Journal of Applied Psychology*, 106-114.

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 and survey-type instruments. *Journal of Management, 19*, 385-417.

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

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). *Manual for the Minnesota Satisfaction Questionnaire*. Minneapolis, MN: Industrial Relations Center, University of Minnesota.

Timothy R. Hinkin is an associate professor of management, research fellow, and director of undergraduate studies at Cornell University's School of Hotel Administration. He received his Ph.D. in organizational behavior from the University of Florida. His research interests include leadership, organizational performance, quality management, and training effectiveness.

J. Bruce Tracey is an associate professor of management and research fellow at Cornell University's School of Hotel Administration. He received his Ph.D. in organizational studies from the State University of New York at Albany. His research interests include leadership, training effectiveness, and organizational culture.