Grouped Versus Randomized Format: An Investigation of Scale Convergent and Discriminant Validity Using LISREL Confirmatory Factor Analysis

Chester A. Schriesheim, University of Miami

Esther Solomon, Fordham University

Richard E. Kopelman, Baruch College, City University of New York

LISREL maximum likelihood confirmatory factor analyses (Jöreskog & Sörbom, 1984) were conducted to explore the effects of two questionnaire formats (grouping versus randomizing items) on the convergent and discriminant validity of two sets of questionnaire measures. The first set of measures consisted of satisfaction scales that had demonstrated acceptable psychometric properties in earlier studies; the second set of scales were job characteristics measures that had shown discriminant validity problems in previous research. Correlational data were collected from two groups of employed business administration students (N = 80 in each group) concurrently (Study 1) and at two points in time (Study 2). The results of the analyses showed that the grouped format was superior to the random format, particularly with respect to the weaker measures (the job characteristics scales). The results also illustrated and supported the usefulness of LISREL confirmatory factor analysis in studies of convergent and discriminant validity. *Index terms: confirmatory factor analysis, convergent validity, discriminant validity,* LISREL *analysis, questionnaire formats, scale validity.*

Questionnaires are widely used in applied psychological research and, of course, thousands of different instruments exist. The typical format presents items measuring different constructs to respondents in a randomized, unlabeled manner. Other item presentation formats are rarely used. There has been little research on whether format exerts meaningful effects on the psychometric adequacy of various measurement instruments, despite the fact that format is probably one of the easiest characteristics that might be changed on existing measures. Potentially, format changes could result in psychometric improvements without requiring the redevelopment of instruments or the expenditure of much time or effort.

Randomization of the order of questionnaire items is often assumed as necessary or desirable for measuring instruments, without fully considering other options or the possible consequences of randomization. Schriesheim and DeNisi (1980) have noted that randomizing items makes sense when it is necessary to disguise the true purpose of a measure to reduce the impact of various response biases. However, randomizing items could also be dysfunctional. If respondents are trying to provide accurate answers, randomization makes the task more difficult and may therefore impair the validity of responses. This suggests that grouping and labeling items measuring the same constructs may be advantageous; such a format may break monotony and lessen fatigue effects, for example, by presenting respondents with many short questionnaire subsections as opposed to several long subsections. It may also help convince respondents that many items are not simply duplicates, designed to "catch" them in inconsistencies (Schriesheim & DeNisi, 1980).

Although not extensive, some research bears on the effects of different item presentation formats.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 1, March 1989, pp. 19–32

0146-6216/89/010019-14$1.95

Baehr (1953) assessed the impact of grouping versus randomly distributing items measuring various job-related attitudes. Almost identical attitude profiles were obtained using both formats, suggesting no differences in measurement between the two modes.

Schriesheim and DeNisi (1980) and Schriesheim (1981a, 1981b) examined the effects of grouped and labeled versus randomized items on the reliability and validity of two leadership instruments. Grouping and labeling items resulted in impaired discriminant validity (Schriesheim & DeNisi, 1980) and heightened leniency response bias (Schriesheim, 1981a). Additionally, follow-up analyses suggested that the impaired discriminant validity was due to the heightened leniency effects (Schriesheim, 1981b).

Solomon and Kopelman (1984) examined the effects of grouping, grouping and labeling, and randomizing items on internal consistency reliabilities. Using four samples and measures of perceived job characteristics and personal satisfaction (job, life, family, and self), modestly to moderately higher internal consistency reliabilities were obtained for the grouped format and the grouped and labeled format than for the randomized format.

The results of these studies appear reasonably consistent: Grouping or grouping and labeling items measuring the same constructs may result in heightened response biases (such as leniency), producing higher reliabilities and also impairing discriminant validity (due to the effects of a common underlying response tendency). These studies suggest that the randomized format is superior for use in applied psychological research, because although reliability may be enhanced by grouping or grouping and labeling, systematic error may be increased as well.

The prior studies do not, however, provide adequate evidence for drawing firm and widely generalizable conclusions about the effects of item presentation format. The Baehr (1953) study involved only a limited comparison of grouped versus randomized item formats. The Solomon and Kopelman (1984) study used four different samples; thus differences in format effects could have been due, at least in part, to sample differences. Finally, the Schriesheim and DeNisi (1980) and Schriesheim (1981a, 1981b) studies all used the same small sample and examined format effects only for leadership scales. Because the leadership scales used in these studies suffer from substantial leniency effects (Schriesheim, Kinicki, & Schriesheim, 1979), generalization of the results to other, less response bias-prone research instruments is unclear.

Given that the existing empirical evidence merely suggests that the randomized item presentation format may be superior to grouped formats, the current investigation was undertaken. Two different studies were conducted to explore format effects on convergent and discriminant validity within the same group at one point in time (Study 1), and on convergent and discriminant validity within two randomly-assigned subgroups at two points in time (Study 2).

Study 1

Participants and Procedure

Participants in the first study were 80 students enrolled in business courses at an Eastern university. All participants had worked full- or part-time within the previous year, and their most recent job experiences were used as the referents in completing the study instruments. Participation was voluntary and respondents were guaranteed anonymity; assignment to the two conditions was on a random basis.

Study 1 compared responses to grouped and randomized measures obtained from the same individuals, thereby controlling for respondent effects. In this study, scale formats were treated as methods. In half the cases (N = 40), the first part of the questionnaire included randomized satisfaction scales and grouped job characteristics scales; the second part consisted of grouped satisfaction scales and randomized job characteristics scales. In the other half of the cases (N = 40), the first section consisted of grouped satisfaction scales and randomized job characteristics scales; the second section consisted of randomized satisfaction scales and grouped job characteristics scales. (Within these two groups, the instruments were also counterbalanced with respect to presentation order—first the satisfaction scales and then the job characteristics scales, and vice-versa.)

Measures

Job characteristics scales. Two sets of measures were used. The first consisted of five job characteristics scales from the Job Diagnostic Survey (JDS; Hackman & Oldham, 1975). These scales have had notable discriminant validity problems in previous research (Dunham, Aldag, & Brief, 1977; Pierce & Dunham, 1978); including them permitted the examination of format effects for scales with known psychometric problems. The five JDS scales measure the perceived "core job dimensions" of Autonomy (A), Variety (V), Significance (S), Identity (I), and Feedback from the job (F). In the grouped version, items relating to each job dimension were grouped together and labeled; in the randomized version, the 15 items were randomly distributed, so that no indication was given of the dimension each was measuring.

Satisfaction measures. The second set of measures consisted of four satisfaction scales examining Life, Self, Job, and Family satisfaction. All scales used a five-point (Likert) response format, with alternatives ranging from "strongly agree" to "strongly disagree". These scales were selected for investigation because they were expected not to suffer from discriminant validity problems (unlike the JDS scales), as they came from fairly diverse content domains and had not previously shown psychometric deficiencies.

Life satisfaction (LS) was measured by two items similar to those used by the Survey Research Center's quality-of-employment surveys (e.g., "In general, I am satisfied with the way I spend my time these days"; Robinson & Shaver, 1973), plus two additional items which were developed in prior research (Kopelman, Greenhaus, & Connolly, 1983). Self-satisfaction (SS) was measured by two items, one identical to an item from the Rosenberg (1965) scale and one similar to an item from that scale. Job satisfaction (JS) was measured by the three-item JDS general satisfaction scale (Hackman & Oldham, 1975).

Family satisfaction (FS) was measured by three items which basically restated the general job satisfaction scale, substituting the word "family" for "job" (e.g., "I frequently think I would like to change my family situation"). In the grouped version of the questionnaire, items pertaining to each of the four satisfaction areas were grouped by area and labeled; this was not done in the randomized version.

Data Analysis

Internal consistency reliability. Coefficient alpha internal consistency reliabilities were first computed for all the measures of Study 1; no statistically significant differences were found between the grouped and randomized scale reliabilities. Of the 18 computed coefficients, the 3 lowest were .59, .65, and .69; all others exceeded .70. Thus, while 3 coefficients were lower than desired, the scales were generally of adequate reliability for meaningful data analyses (Nunnally, 1978).

Convergent and discriminant validity. The multitrait-multimethod matrix framework first suggested by Campbell and Fiske (1959) for examining the validity of an instrument was employed. However, rather than using Campbell and Fiske's largely nonstatistical assessment criteria, the analyses were performed by LISREL VI maximum likelihood confirmatory factor analysis (Jöreskog & Sörbom, 1984), employing the procedures outlined by Widaman (1985) and recommended by Schmitt and Stults (1986). Confirmatory factor analysis was employed because it allows the examination of the grouped and random scales separately. This is not possible in rival approaches, such as Kavanagh, MacKinney, and Wollins' (1971) analysis-of-variance approach (Schmitt & Stults, 1986).

First, a full matrix model (Model 3C in Widaman's taxonomy) was fit to the multitrait-multimethod correlation matrices,[1] using Rindskopf's (1983) parameterization technique (Widaman, 1985,

[1]To conserve space, the input correlation matrices are not presented here; they are given in a preliminary analysis of these data (Schriesheim, Kopelman, & Solomon, in press) which used weaker and more traditional analyses (and reached different conclusions than those presented here).

pp. 10–11). This full model consisted of three independent sets of factor components: trait, method, and unique factors. Each measured variable was specified as having a loading on one trait, one method, and one unique factor, with no inappropriate or cross-loadings (see, e.g., Table 1). Following convention, the trait factors were allowed to be correlated with each other but not with the method or unique factors, and the methods were allowed to be correlated only between themselves. The unique (random error) factors were specified as uncorrelated among themselves or with any other factors.

The goodness of fit of this full model to the data was assessed by a chi-square likelihood test (Jöreskog & Sörbom, 1984) and by Bentler and Bonett's (1980) rho and delta indices. Rho (ρ) is a relative index of the off-diagonal covariation among observed variables which is explained by a model; delta (Δ) is an absolute measure representing the proportion of off-diagonal covariance independent of degrees of freedom (see Schmitt & Stults, 1986, pp. 13–14). Following Widaman (1985) and Schmitt and Stults (1986), the full model was considered acceptable if the chi-square probability exceeded .05 and both ρ and Δ were .90 or greater.

To test for convergent and discriminant validity and method bias effects, the full model described above was then compared to three additional models by first estimating those models and then calculating the differences in χ^2, ρ, and Δ between each rival model and the full model. The first rival model consisted of only method and unique factors—with no trait factors (Widaman's Model 1C)—and assessed the lack of convergent validity. The second rival model (Model 2C) was identical to the initial full model except that all trait intercorrelations were constrained to 1; this model assessed the lack of discriminant validity. Finally, the third rival model (Model 3A) included no method factors and assessed the presence of method bias effects. Following Widaman (1985), χ^2 differences with probabilities in excess of .05 or differences of .01 or greater in ρ or Δ were considered to indicate that a rival model provided a more meaningful fit to the data than did the initial model.

Results

Satisfaction measures. Tables 1 and 2 present the LISREL-estimated full model results for the satisfaction and job characteristics scales, respectively; Table 3 presents the analyses comparing these models to those indicating a lack of convergent and discriminant validity and the presence of method bias.

As shown in Tables 1 and 3, the confirmatory analysis of the satisfaction scales indicates an excellent fit of the initial full model to the data ($\chi^2 = 8.86$, $p = .12$; $\rho = .94$; $\Delta = .98$), and indicates that meaningful convergent and discriminant validity and method bias effects are present (all ρ and Δ differences were $-.07$ or less). Furthermore, examination of the trait intercorrelations (Table 1) does not suggest any need to eliminate or combine trait factors (i.e., to hypothesize and estimate other initial structural models).

Averaging and squaring the factor loadings shown in Table 1 (to examine variance attributable to trait, method, and unique factors) shows that for the eight satisfaction measures, trait factor variance accounts for 75% of the total variance. Furthermore, method factor variance averages only 12% and error factor variance only 14% (the sum exceeds 100% due to rounding error). The grouped scales have a higher average percentage of trait variance than do the random scales (78% vs. 71%, respectively), and less method variance (7% vs. 18%). However, the grouped scales show slightly more error (16% vs. 12%). Overall, this analysis suggests some advantage for the grouped format for the satisfaction scales, as they have greater trait and less method variance than do the random scales.

Job characteristics scales. Tables 2 and 3 present the confirmatory analyses relevant to the job characteristics measures. Again, these results indicate an excellent fit of the initial full model to the data ($\chi^2 = 5.02$, $p = .99$; $\rho = 1.09$; $\Delta = .99$), and indicate that meaningful convergent and discriminant validity and method bias effects are present (all ρ and Δ differences were $-.10$ or less).

The trait factor intercorrelations (see Table 2) show that it was necessary in this analysis to con-

Table 1
Parameters Estimated in the Confirmatory Analysis of the Study 1 Satisfaction Scales With Four Trait and Two Method Factors

	Trait Factors				Method Factors	
	LS	JS	FS	SS	Grouped	Random
Trait and Method Factor Loadings						
Grouped LS	1.00	0.0*	0.0*	0.0*	-.05	0.0*
JS	0.0*	.78	0.0*	0.0*	.39	0.0*
FS	0.0*	0.0*	.97	0.0*	.25	0.0*
SS	0.0*	0.0*	0.0*	.77	.27	0.0*
Random LS	.76	0.0*	0.0*	0.0*	0.0*	.66
JS	0.0*	1.00	0.0*	0.0*	0.0*	.07
FS	0.0*	0.0*	.73	0.0*	0.0*	.27
SS	0.0*	0.0*	0.0*	.85	0.0*	.46
Intercorrelation of Trait and Method Factors						
LS	1.0*					
JS	.45	1.0*				
FS	.38	.14	1.0*			
SS	.67	.55	.26	1.0*		
Grouped	0.0*	0.0*	0.0*	0.0*	1.0*	
Random	0.0*	0.0*	0.0*	0.0*	.41	1.0*
Random Errors Associated With Each Measured Variable						
Grouped	0.00	.53	0.00	.60		
Random	0.00	0.00	.58	.35		

*Fixed values.

strain the Variety-Feedback trait factor intercorrelation at .99 to obtain parameter estimates within bounds (i.e., with absolute values less than or equal to 1),[2] and that the Autonomy-Feedback intercorrelation was estimated to be 1. Although this suggests that several other models might be examined, further analyses indicated that they did not provide meaningfully better fit to the data. Constraining the Variety-Feedback parameter at 1 did not produce a better fit. Poorer fits were obtained by eliminating the Feedback trait factor and examining three new models with both the grouped and random Feedback measures loading on (1) the Variety trait factor, (2) the Autonomy trait factor, and (3) both the Variety and Autonomy trait factors. This suggests that the original full model was an acceptable portrayal of the job characteristics data.

Analysis of the results in Table 2 shows that the job characteristics scales are of poorer psychometric quality than are the satisfaction scales. Here, trait factor variance accounts for 43% of the total variance in the 10 job characteristics measures, while method factor variance accounts for 35% and error explains 23%. This poor showing relative to the satisfaction scales is due largely to the random job scales rather than the grouped scales. The grouped scales average 51% trait variance, as compared to only 35% for the random scales. Both have approximately equal method variance (36% for the grouped and 34% for the random), while the grouped job scales have only 16% error variance, as compared to 30% for the random scales. This pattern is similar to—but stronger than—that found for the satisfaction scales, indicating a considerable advantage for the grouped job characteristics scales over the random scales.

[2]The footnotes to Tables 3, 6, and 8 show all parameters in the analyses that were constrained so as to produce within-bounds estimates (Widaman, 1985); other than in this instance (the .99 Variety-Feedback correlation), they all involved the method correlations of the rival models (which were all constrained at 1; other values uniformly produced poorer model fits).

Table 2
Parameters Estimated in the Confirmatory Analysis of the Study 1
Job Characteristics Scales With Five Trait and Two Method Factors

		Trait Factors					Method Factors	
		V	I	S	A	F	Grouped	Random
Trait and Method Factor Loadings								
Grouped	V	.74	0.0*	0.0*	0.0*	0.0*	.53	0.0*
	I	0.0*	.90	0.0*	0.0*	0.0*	.42	0.0*
	S	0.0*	0.0*	.89	0.0*	0.0*	.44	0.0*
	A	0.0*	0.0*	0.0*	.61	0.0*	.70	0.0*
	F	0.0*	0.0*	0.0*	0.0*	−.13	.75	0.0*
Random	V	.69	0.0*	0.0*	0.0*	0.0*	0.0*	.66
	I	0.0*	.66	0.0*	0.0*	0.0*	0.0*	.40
	S	0.0*	0.0*	.50	0.0*	0.0*	0.0*	.58
	A	0.0*	0.0*	0.0*	.70	0.0*	0.0*	.55
	F	0.0*	0.0*	0.0*	0.0*	−.33	0.0*	.75
Intercorrelation of Trait and Method Factors								
	V	1.0*						
	I	.22	1.0*					
	S	.06	−.11	1.0*				
	A	.46	.20	.09	1.0*			
	F	.99*	−.24	.56	1.00	1.0*		
	Grouped	0.0*	0.0*	0.0*	0.0*	0.0*	1.0*	
	Random	0.0*	0.0*	0.0*	0.0*	0.0*	.80	1.0*
Random Errors Associated With Each Measured Variable								
	Grouped	.45	0.00	0.00	.40	.64		
	Random	.37	.65	.63	.49	.55		

*Fixed values.

Discussion

The results presented above show that grouping may have a beneficial effect on psychometric adequacy, particularly for measures with a history of convergent and discriminant validity problems (Dunham et al., 1977; Pierce & Dunham, 1978). However, because the findings of a single study may be due to various artifacts, a second study was undertaken.

Study 2

As Schneider and Dachler (1978) have noted, "when a measure is employed in a time-based mode, stability ('retest reliability'), not internal consistency, is the critical index of reliability" (p. 650). For this reason, and because it is desirable to replicate confirmatory analysis results (Schmitt & Stults, 1986), Study 2 explored format effects over time.

Schneider and Dachler (1978) conceptualized two administrations of an instrument (separated by time) as two measurement methods or sources; therefore, a standard multitrait-multimethod matrix analysis can be used to examine test-retest reliability as "convergent stability" ("convergent validity" in Campbell and Fiske's terminology) by assessing overall respondent agreement across administrations of an instrument and its scales. Discriminant validity can be assessed by examining within- and across-administration scale intercorrelations, and method or source bias can be assessed on the basis of respondent × administration variance over scales. The principal advantage of treating test-retest administrations in this manner is that it allows an assessment of the relative magnitudes of score stability and scale independence (discrimination) in comparison with source or method bias. This, of course, is useful as another way of assessing the merits of grouping or randomizing questionnaire items on psychometric quality.

Table 3
Goodness-of-Fit Indices for the Study 1 Confirmatory Models

	Indices of Model Fit				
	Likelihood Ratio Test				
Model	χ^2	df	p	ρ	Δ
Satisfaction Scales					
Full Model (Model 3C; see Table 1)	8.86	5	.12	.94	.98
No Trait Factors (Model 1C)	165.04	20[a]	.00	.41	.56
Perfectly Correlated Traits (Model 2C)	165.04	12[a]	.00	−.04	.56
No Method Factors (Model 3A)	36.90	14	.01	.87	.90
Job Characteristics Scales					
Full Model (Model 3C; see Table 2)	5.02	15[b]	.99	1.09	.99
No Trait Factors (Model 1C)	176.41	35[a]	.00	.47	.54
Perfectly Correlated Traits (Model 2C)	128.52	24	.00	.42	.67
No Method Factors (Model 3A)	43.78	25	.02	.90	.89

	Comparative Indices of Model Differences				
	Difference in			Difference in	
Model Comparison	χ^2	df	p	ρ	Δ
Satisfaction Scales					
Convergent Validity (Model 3C vs. 1C)	156.18	15	.00	−.53	−.42
Discriminant Validity (Model 3C vs. 2C)	156.18	7	.00	−.98	−.42
Method Bias (Model 3C vs. 3A)	28.04	9	.01	−.07	−.08
Job Characteristics Scales					
Convergent Validity (Model 3C vs. 1C)	171.39	20	.00	−.62	−.45
Discriminant Validity (Model 3C vs. 2C)	123.50	9	.00	−.67	−.32
Method Bias (Model 3C vs. 3A)	38.76	10	.01	−.19	−.10

[a]For this analysis, the correlation between methods was constrained at 1.0.
[b]For this analysis, the Variety-Feedback factor correlation was constrained at .99 (see Table 2).

Participants and Procedure

The participants in the second study were 80 different business students from the same Eastern university as the first study. The questionnaire referents, anonymity guarantees, and administration procedures were identical, and the respondents were randomly assigned to the two conditions.

In this study, format effects on scale score stabilities were examined by treating instrument administration times as methods. Participants were twice administered the same set of randomized or grouped questionnaire versions at 5- to 8-week intervals. Half of the participants ($N = 40$) received a randomized satisfactions instrument and a grouped job characteristics instrument for both administrations, while the other half ($N = 40$) received a grouped satisfactions instrument and a randomized job characteristics instrument at both administrations (within these two groups, the instruments were also counterbalanced with respect to presentation order).

Measures and Analysis

The measures used in Study 2 were identical to those employed in the first study. The data-analytic methods used were also identical, except as noted below.

Results

Internal consistency reliabilities. The coefficient alpha reliability estimates obtained for Study 2 were very similar to those of Study 1. Again, no reliability differences were statistically significant, so that neither format appeared preferable with respect to internal consistency reliability.

Satisfaction scale results. With respect to convergent and discriminant validity, the general results of Study 1 were further supported by the confirmatory analyses of Study 2. Tables 4 through 6 present these results for the satisfaction measures, while Tables 7 through 9 present results for the job characteristics scales.

As shown in Table 6, the goodness of fit of the initial model to the grouped satisfaction scales is quite satisfactory ($\chi^2 = 7.21$, $p = .21$; $\rho = .92$; $\Delta = .96$), as is that for the random measures ($\chi^2 = 2.11$, $p = .83$; $\rho = 1.10$; $\Delta = .99$). The intercorrelations among the grouped satisfaction trait factors shown in Table 4 do not suggest the necessity of trying other initial models, nor do those shown in Table 5 for the random satisfaction scales (testing several alternative initial models supported this conclusion). Testing the initial grouped and random models (Tables 4 and 5) against the three rival models shows meaningful convergent and discriminant validity, as well as meaningful method bias effects (see Table 6).

The factor loadings shown in Table 4 indicate that 73% of the grouped satisfaction scale variance is accounted for by the trait factors, but only 13% is attributable to method factors and 15% to random error. Furthermore, the average Fisher z-transformed intercorrelation (McNemar, 1969) between trait factors is only .44 (the highest $r = .84$), indicating reasonable discriminant validity. Almost identical results were obtained for the randomized satisfaction scales. Here (see Table 5), 72% of the total variance is attributable to the trait factors, with only 15% due to methods and 14% to error. The average trait factor r in this analysis is identical to that for the grouped scales (.44), and the highest r is a more modest .73. These results, then, suggest good measurement properties for the satisfaction measures as a set and no advantage for either the grouped or random formats.

Job characteristics scale results. The results for the job characteristics measures, as shown in Tables 7 through 9, are quite different from those for satisfaction. Here, the initial model fits the

Table 4
Parameters Estimated in the Confirmatory Analysis of the Study 2 Grouped Satisfaction Scales With Four Trait and Two Method Factors

		Trait Factors				Method Factors	
		LS	JS	FS	SS	Adm. 1	Adm. 2
Trait and Method Factor Loadings							
Admin. 1	LS	.88	0.0*	0.0*	0.0*	−.20	0.0*
	JS	0.0*	.96	0.0*	0.0*	.21	0.0*
	FS	0.0*	0.0*	.91	0.0*	−.01	0.0*
	SS	0.0*	0.0*	0.0*	.84	.55	0.0*
Admin. 2	LS	.76	0.0*	0.0*	0.0*	0.0*	.52
	JS	0.0*	.78	0.0*	0.0*	0.0*	−.09
	FS	0.0*	0.0*	.89	0.0*	0.0*	.59
	SS	0.0*	0.0*	0.0*	.80	0.0*	.13
Intercorrelation of Trait and Method Factors							
	LS	1.0*					
	JS	.52	1.0*				
	FS	.40	.40	1.0*			
	SS	.84	.13	.33	1.0*		
	Admin. 1	0.0*	0.0*	0.0*	0.0*	1.0*	
	Admin. 2	0.0*	0.0*	0.0*	0.0*	−.58	1.0*
Random Errors Associated With Each Measured Variable							
	Admin. 1	.45	0.00	.41	0.00		
	Admin. 2	.41	.61	0.00	.59		

*Fixed values.

Table 5
Parameters Estimated in the Confirmatory Analysis of the Study 2 Random Satisfaction Scales With Four Trait and Two Method Factors

		Trait Factors				Method Factors	
		LS	JS	FS	SS	Adm. 1	Adm. 2
Trait and Method Factor Loadings							
Admin. 1	LS	.72	0.0*	0.0*	0.0*	.44	0.0*
	JS	0.0*	.94	0.0*	0.0*	.12	0.0*
	FS	0.0*	0.0*	1.00	0.0*	.01	0.0*
	SS	0.0*	0.0*	0.0*	.83	.27	0.0*
Admin. 2	LS	.68	0.0*	0.0*	0.0*	0.0*	.64
	JS	0.0*	.83	0.0*	0.0*	0.0*	.15
	FS	0.0*	0.0*	.72	0.0*	0.0*	.75
	SS	0.0*	0.0*	0.0*	1.00	0.0*	.03
Intercorrelation of Trait and Method Factors							
	LS	1.0*					
	JS	.73	1.0*				
	FS	.36	.20	1.0*			
	SS	.65	.30	.38	1.0*		
	Admin. 1	0.0*	0.0*	0.0*	0.0*	1.0*	
	Admin. 2	0.0*	0.0*	0.0*	0.0*	.55	1.0*
Random Errors Associated With Each Measured Variable							
	Admin. 1	.55	.33	0.00	.50		
	Admin. 2	.44	.56	0.00	0.00		

*Fixed values.

grouped job scales well ($\chi^2 = 15.92$, $p = .32$; $\rho = .97$; $\Delta = .94$), and the pattern of trait factor intercorrelations is very moderate (mean = .17; highest = .67). This does not suggest the necessity of searching for better-fitting initial models and, as a consequence, none was examined.

Testing the initial grouped job scales model (Table 7) against the three rival models (Table 8) shows them to have meaningful convergent and discriminant validity, as well as meaningful method bias. The trait factors account for 48% of the total grouped job scale variance, with the method factors explaining 35% and random error 17%. Although these are less favorable results than those for both the grouped and random satisfaction measures, they are very close to the variance percentages estimated for the Study 1 analysis of the grouped job scales (the corresponding results were 51%, 34%, and 16%). Again, these data suggest that the satisfaction scales are better measures than the job characteristics scales.

The results for the random job characteristics scales further support this conclusion, as well as the possibility that grouping may improve the psychometric quality of weaker measures. The initial model does not fit the random job scales very well. In fact, the statistics presented in Table 8 show a rather poor fit ($\chi^2 = 171.26$, $p = .00$; $\rho = -.08$; $\Delta = .67$). In an attempt to find a satisfactory initial model, more than 100 additional separate LISREL analyses were undertaken. These were guided by the results from exploratory factor and component analyses, beginning with models that were closest to the initial model, and progressing to simpler ones (involving various combinations of five, four, three, and two trait factors, with two, one, and zero method factors).

These exploratory factor analyses used both the principal factor (with R^2 communality estimates) and principal components methods, with varimax and promax (Hendrickson & White, 1964) rotations as suggested by Gorsuch (1974), Harman (1976), and Rummel (1970) for complex data. The principal factor analyses produced ultra-Heywood (1931) solutions when attempts were made to extract more than three factors. The factor eigenval-

Table 6
Goodness-of-Fit Indices for the Study 2 Satisfaction Scales Confirmatory Models

Model	Indices of Model Fit: Likelihood Ratio Test χ^2	df	p	ρ	Δ
Grouped					
Full Model (Model 3C; see Table 4)	7.21	5	.21	.92	.96
No Trait Factors (Model 1C)	99.71	19	.00	.24	.46
Perfectly Correlated Traits (Model 2C)	99.78	12[a]	.00	−.30	.46
No Method Factors (Model 3A)	34.02	14	.01	.74	.82
Random					
Full Model (Model 3C; see Table 5)	2.11	5	.83	1.10	.99
No Trait Factors (Model 1C)	99.33	20[a]	.00	.32	.48
Perfectly Correlated Traits (Model 2C)	75.82	11	.00	−.02	.60
No Method Factors (Model 3A)	26.19	14	.03	.85	.86

Model Comparison	Comparative Indices of Model Differences: Difference in χ^2	Difference in df	p	Difference in ρ	Difference in Δ
Grouped					
Convergent Validity (Model 3C vs. 1C)	92.50	14	.00	−.68	−.50
Discriminant Validity (Model 3C vs. 2C)	92.57	7	.00	−1.22	−.50
Method Bias (Model 3C vs. 3A)	26.81	9	.01	−.18	−.14
Random					
Convergent Validity (Model 3C vs. 1C)	97.22	15	.00	−.78	−.51
Discriminant Validity (Model 3C vs. 2C)	73.71	6	.00	−1.12	−.39
Method Bias (Model 3C vs. 3A)	24.08	9	.01	−.25	−.13

[a]For this analysis, the correlation between methods was constrained at 1.0.

ues (5.63, 1.36, 1.01, .59, .48, .42, .23, .14, .13, and .01) and a scree plot (Cattell, 1966) also suggested the presence of only three factors.

Because none of these analyses produced both fully acceptable within-bounds estimates and a satisfactory fit to the data, various additional "patch-up" designs were tried, including models with five, four, three, and two trait factors. These additional designs included

1. Kenny's (1976) method that eliminated the method factors and allowed two sets of correlated within-method error;
2. A modification of the Kenny (1976) procedure with error also allowed to be equally correlated across the unique factors;
3. Marsh and Hocevar's (1983) procedure that sets uniqueness factor loadings equal to communalities estimated from principal components analyses (two-, three-, four-, and five-factor communality models were examined); and
4. Combinations of the third approach with the first two approaches.

None of these analyses produced satisfactory results. Because an acceptable initial confirmatory model could not be identified, Table 9 instead presents the results of a promax-rotated three-factor principal components analysis of these data, as well as a five-factor solution (for these analyses, varimax rotations produced virtually identical results). As indicated above, the three-factor results better portray the data, but the five-factor results are shown because of the a priori dimensionality of the job characteristics scales (principal components results

Table 7
Parameters Estimated in the Confirmatory Analysis of the Study 2 Grouped Job Characteristics Scales With Five Trait and Two Method Factors

		Trait Factors					Method Factors	
		V	I	S	A	F	Adm. 1	Adm. 2
Trait and Method Factor Loadings								
Admin. 1	V	.81	0.0*	0.0*	0.0*	0.0*	.42	0.0*
	I	0.0*	.62	0.0*	0.0*	0.0*	.78	0.0*
	S	0.0*	0.0*	.70	0.0*	0.0*	.61	0.0*
	A	0.0*	0.0*	0.0*	.36	0.0*	.77	0.0*
	F	0.0*	0.0*	0.0*	0.0*	.57	.59	0.0*
Admin. 2	V	.70	0.0*	0.0*	0.0*	0.0*	0.0*	.47
	I	0.0*	.23	0.0*	0.0*	0.0*	0.0*	.70
	S	0.0*	0.0*	.87	0.0*	0.0*	0.0*	.49
	A	0.0*	0.0*	0.0*	.81	0.0*	0.0*	.59
	F	0.0*	0.0*	0.0*	0.0*	.88	0.0*	.39
Intercorrelation of Trait and Method Factors								
	V	1.0*						
	I	.19	1.0*					
	S	.34	−.12	1.0*				
	A	.67	−.10	.04	1.0*			
	F	.28	.24	.23	−.06	1.0*		
Admin. 1		0.0*	0.0*	0.0*	0.0*	0.0*	1.0*	
Admin. 2		0.0*	0.0*	0.0*	0.0*	0.0*	.90	1.0*
Random Errors Associated With Each Measured Variable								
Admin. 1		.39	0.00	.33	.52	.57		
Admin. 2		.53	.67	.08	0.00	.23		

*Fixed values.

are shown for comparability between the three- and five-factor results because of the problems encountered with the principal factor analyses).[3]

Table 9 presents the oblique-rotated pattern loadings and the factor intercorrelations for the three- and five-factor solutions (the loadings in excess of 1.0 are due to their obliqueness). As shown in Table 9, a clear lack of discriminant validity occurs among the Variety, Autonomy, and Feedback scales for the three-factor solution. In the three-factor solutions, these scales (at both administrations) have high (.50 or greater) loadings on the same factor (I); the Administration 1 Feedback scale also has a high loading on Factor II, and the Administration 2 Variety scale has a moderate Factor III loading. The Identity scales fare much better, with both administrations' measures having only large loadings on Factor III (and none on the other factors); although the Administration 2 Variety scale also loads on this factor (III), the loading is moderate. A similar pattern is evident with the two administrations of the Significance scales, as these have their sole meaningful loadings on Factor II and only the Administration 2 Feedback scale loads highly on this factor as well.

The five-factor results shown in Table 9 appear in some ways even poorer. Not only are confounded factors obtained, but several of the scales do not meaningfully load on the same factors across the two administrations (the Variety and Identity scales, in particular). These results, then, taken in conjunction with those for both the grouped job scales and both sets of satisfaction scales, indicate clearly that the grouped format generally improved

[3]More detailed information concerning the various attempted analyses is available from the authors on request.

Table 8
Goodness-of-Fit Indices for the Study 2 Job Scales Confirmatory Models

Model	Indices of Model Fit: Likelihood Ratio Test χ^2	df	*p*	ρ	Δ
Grouped					
Full Model (Model 3C; see Table 7)	15.92	14	.32	.97	.94
No Trait Factors (Model 1C)	130.05	35[a]	.00	.48	.54
Perfectly Correlated Traits (Model 2C)	130.05	25[a]	.00	.20	.54
No Method Factors (Model 3A)	38.93	25	.04	.89	.86
Random					
Full Model (Model 3C; not shown)	171.26	14	.00	−.08	.67
No Trait Factors (Model 1C)	302.06	35[a]	.00	.27	.41
Perfectly Correlated Traits (Model 2C)	256.10	24	.00	.07	.50
No Method Factors (Model 3A)	233.21	25	.00	.20	.55

Model Comparison	Comparative Indices of Model Differences: Difference in χ^2	Difference in df	*p*	Difference in ρ	Difference in Δ
Grouped					
Convergent Validity (Model 3C vs. 1C)	114.13	21	.00	−.49	−.40
Discriminant Validity (Model 3C vs. 2C)	114.13	11	.00	−.77	−.40
Method Bias (Model 3C vs. 3A)	23.01	11	.02	−.08	−.08
Random					
Convergent Validity (Model 3C vs. 1C)	130.81	21	.00	.35	−.25
Discriminant Validity (Model 3C vs. 2C)	84.84	10	.00	.15	−.17
Method Bias (Model 3C vs. 3A)	61.95	11	.00	.28	−.12

[a]For this analysis, the correlation between methods was constrained at 1.0.

Table 9
Promax-Rotated Principal Components Analyses
for Study 2 Random Job Characteristics Scales

		Three-Factor Solution				Five-Factor Solution					
		I	II	III	h^2	I	II	III	IV	V	h^2
Rotated Pattern Matrix Loadings											
Admin. 1	V	.62	.12	.36	.82	.24	.14	.42	.36	.08	.84
	I	−.21	.25	.83	.73	−.03	−.05	−.10	.25	.93	.97
	S	−.06	.98	−.01	.91	−.04	1.04	−.06	−.05	−.02	.93
	A	.84	.16	−.12	.79	.92	.03	.11	−.13	−.08	.89
	F	.54	.55	−.23	.76	.15	.30	.54	−.36	.26	.90
Admin. 2	V	.55	.07	.47	.81	.45	.11	.15	.44	.06	.82
	I	.10	−.20	.93	.85	−.05	0.00	.05	.90	.15	.91
	S	.08	.77	.22	.83	.06	.89	−.04	.23	−.08	.90
	A	.75	.07	.18	.78	1.05	−.05	−.12	.08	.02	.98
	F	.95	−.16	−.08	.72	−.04	−.10	1.04	.18	−.15	.94
Factor Intercorrelations											
	I	1.00				1.00					
	II	.52	1.00			.58	1.00				
	III	.41	.34	1.00		.64	.51	1.00			
	IV					.34	.20	.20	1.00		
	V					.42	.48	.32	.33	1.00	

measurement quality in Study 2, particularly for the psychometrically weaker job characteristics scales. Although difficulties were encountered in analyzing the random job scales in Study 2, the overall pattern of results seems sufficiently clear to draw this conclusion.

Discussion and Conclusions

Previous research has suggested that grouping items may improve scale internal consistency reliabilities while impairing scale discriminant validity. The results of the current two studies do not support these conclusions.

The present evidence suggests that grouping items does not significantly enhance internal consistency reliabilities, nor does it necessarily result in impaired discriminant validity. In fact, the confirmatory factor analyses suggested that grouping may enhance the convergent validities of measures without impairing discriminant validity or seriously increasing method bias effects. Overall, these results suggested a reasonable improvement in psychometric quality (both convergent and discriminant validity) for scales with previously documented problems (the job characteristics scales).

Because the present evidence suggests that grouping items may improve the psychometric properties of weak measures, further studies of format effects should be conducted. In particular, research might profitably explore format effects in different measurement domains (other than satisfaction and job characteristics) and under different conditions, perhaps for scales with varying levels of psychometric adequacy.

The current research also supports the assertions of Schmitt and Stults (1986), Widaman (1985), and others: that confirmatory factor analysis is a sensitive analytic technique for this type of research, and that future research in this domain would be well advised to use this approach (perhaps as the main data-analytic technique). However, regardless of the methodology employed, researchers should not be content with the current psychometric properties of their instruments; they should seek to improve all measures by whatever means possible (Schriesheim & DeNisi, 1980). Because improvements in psychometric adequacy appear possible from use of the grouped questionnaire format, future research exploring its usefulness in other domains seems both necessary and desirable.

References

Baehr, M. E. (1953). A simplified procedure for the measurement of employee attitudes. *Journal of Applied Psychology, 37,* 163–167.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

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

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

Dunham, R. B., Aldag, R. J., & Brief, A. P. (1977). Dimensionality of task design as measured by the job diagnostic survey. *Academy of Management Journal, 20,* 209–223.

Gorsuch, R. L. (1974). *Factor analysis.* Philadelphia: Saunders.

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60,* 159–179.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology, 17,* 65–70.

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London, 134,* 486–501.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods.* Mooresville IN: Scientific Software.

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

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12,* 247–252.

Kopelman, R. E., Greenhaus, J. H., & Connolly, T. F. (1983). A model of work, family, and inter-role conflict: A construct validation study. *Organizational Behavior and Human Performance, 32,* 198–215.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analyses of multitrait-multimethod matrices. *Journal of Educational Measurement, 20,* 231–248.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

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

Pierce, J. L., & Dunham, R. B. (1978). The measurement of perceived job characteristics: The job diagnostic survey versus the job characteristics inventory. *Academy of Management Journal, 21,* 123–128.

Rindskopf, D. (1983). Parameterizing inequality constraints on unique variances in linear structural models. *Psychometrika, 48,* 73–83.

Robinson, J. P., & Shaver, P. R. (1973). *Measures of social psychological attitudes* (rev. ed.). Ann Arbor: University of Michigan, Institute for Social Research, Survey Research Center.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton NJ: Princeton University Press.

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

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

Schneider, B., & Dachler, H. P. (1978). A note on the stability of the job descriptive index. *Journal of Applied Psychology, 63,* 650–653.

Schriesheim, C. A. (1981a). The effect of grouping or randomizing items on leniency response bias. *Educational and Psychological Measurement, 41,* 401–411.

Schriesheim, C. A. (1981b). Leniency effects on convergent and discriminant validity for grouped questionnaire items: A further investigation. *Educational and Psychological Measurement, 41,* 1093–1099.

Schriesheim, C. A., & DeNisi, A. S. (1980). Item presentation as an influence on questionnaire validity: A field experiment. *Educational and Psychological Measurement, 40,* 175–182.

Schriesheim, C. A., Kinicki, A. J., & Schriesheim, J. F. (1979). The effect of leniency on leader behavior descriptions. *Organizational Behavior and Human Performance, 23,* 1–29.

Schriesheim, C. A., Kopelman, R. E., & Solomon, E. (in press). The effect of grouped versus randomized questionnaire format on scale reliability and validity: A three-study investigation. *Educational and Psychological Measurement*.

Solomon, E., & Kopelman, R. E. (1984). Questionnaire format and scale reliability: An examination of three modes of item presentation. *Psychological Reports, 54,* 447–452.

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

Acknowledgments

This research was supported by grants from the Research Council and Social Science Committee of the University of Miami and by the Corporate Affiliate Research Program of the School of Business Administration, University of Miami. The authors thank the Editor and two anonymous reviewers of earlier drafts of this paper for their helpful comments, David Rindskopf for his advice and suggestions on some of the LISREL analyses, and Linda A. Tetrault for her assistance in performing many of the analyses.

Author's Address

Send requests for reprints or further information to Chester A. Schriesheim, Management Department, School of Business Administration, University of Miami, 414 Jenkins Building, Coral Gables FL 33124-9145, U.S.A.