Factorial Invariance in Student Ratings of Instruction

Isaac I. Bejar
Educational Testing Service

Kenneth O. Doyle
University of Minnesota

The factorial invariance of student ratings of instruction across three curricular areas was investigated by means of maximum likelihood factor analysis. The results indicate that a one-factor model was not completely adequate from a statistical point of view. Nevertheless, a single factor was accepted as reasonable from a practical point of view. It was concluded that the single factor was invariant across three curricular groups. The reliability of the single factor was essentially the same in the three groups, but in every case it was very high. Some of the theoretical and practical implications of the study were discussed.

The invariance, or the generalizability, of factor structures underlying student ratings of instruction is of considerable practical and theoretical interest. Practically, student ratings are often interpreted comparatively or normatively, and meaningful interpretation of the ratings requires that they "measure the same thing" in the different contexts in which they are compared. Theoretically, the generalizability of factor structures has implications for the study of differences in students, in instructional characteristics, and in courses across different populations or segments of populations, and for the design of reliability and validity studies. Invariant factor structures would suggest a consistency of human behavior that is interesting in itself and that reduces the need to consider each population segment separately in attempting to devise a theory of instruction or in appraising the quality of measures for the evaluation of instruction.

Prior studies have raised the issue of factorial invariance in student ratings (e.g., Bejar & Doyle, 1974). However, the problem is a general one and may be found throughout the social sciences, ranging from personality measurement (e.g., Bejar, 1977) to cross-cultural research (e.g., Irvine & Sanders, 1972). The purpose of this paper is to examine the extent to which factor structures obtained from student ratings of instruction are invariant across courses from three curriculum areas.

There are a number of different levels at which factorial invariance can be investigated. At one extreme, factorial invariance refers only to the existence of the same number of common factors in each population under consideration (e.g., the congruence of Teaching Assistant and Professor factors in Whitely & Doyle, 1979). At the other extreme, all aspects of the model are examined, including the number of factors, the loadings on each factor, the uniqueness of each variable, and the covariation among the factors. An intermediate model would examine invariance up to the uniqueness of the variables. If this model were found to be correct, it would im-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 3, Summer 1981, pp. 307–312

0146-6216/81/030307-06$1.30

ply that all situational and other random disturbances are present to the same extent relative to the true variance. This would in turn imply the invariance of scale reliabilities across all populations and the invariance of the metric of the underlying factors. An important implication of these two results is that the "numbers" obtained are in all cases directly comparable. This investigation is concerned with assessing the fit of the less stringent model, where only the numbers of common factors is invariant (Model I), versus the fit of the intermediate model, which specifies invariance up to the uniqueness of the scales (Model II).

Method

Instrument

The *Student Opinion Survey* (Doyle, 1977) was used to collect the ratings. Because items dealing with tests and reading materials were not appropriate for all courses in the sample, those items were deleted from the analysis. The items used in this study appear in Table 2.

Sample and Design

The data consist of ratings from students in 177 courses in a wide range of upper and lower division courses in the College of Liberal Arts, University of Minnesota. In the context of a larger study, a random sample of 369 courses was drawn, stratified by department and course levels. Some 274 instructors agreed to participate, but only 177 (47%) provided usable data sets. Although a visual comparison of the list of participating courses to the list of sampled courses revealed no substantial differences in discipline or level, it is likely that some sampling bias of unknown direction and degree existed in these data.

On the basis of the College of Liberal Arts' specifications for undergraduate distribution requirements, 67 of the courses were classified as "Communication/Language/Symbolic Systems," comprising such fields as German and speech communications; 63 were classified as "Man and Society," including sociology and economics; and 47, as "Artistic Expression," including music and theatre arts.

Analysis and Results

As a first step in the analysis, the mean ratings for each instructor within each of the three curriculum groups was obtained. The variance-covariance matrix for mean ratings was computed for each group of instructors.

To assess the fit of Model I, each covariance matrix was converted to a correlation matrix. The eigenvalues associated with the three matrices are shown in Figure 1, which shows a predominant single factor. Each correlation matrix was then analyzed by using a maximum likelihood approach (Jöreskog, 1969) specifying up to three common factors. The χ^2 fit statistic for each solution was obtained and the Tucker-Lewis (1973) index of fit was computed. These results are shown in Table 1.

Figure 1
Eigenvalues of the Correlation Matrix for the 11 Scales for Each Curriculum Group

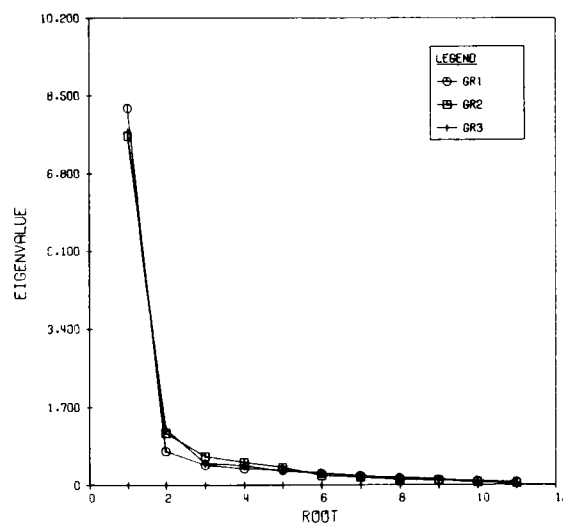

Table 1
Chi-Square Statistics and Tucker-Lewis (T-L) Indices for the Number of Common Factors in Three Curriculum Groups

			Number of common factors			
Curriculum Group	N		0	1	2	3
Communication/Languages/ Symbolic Systems/	67	χ^2	3913	110	44	27
		df	55	44	34	25
		T-L	--	.96	.98	.98
Man and Society	63	χ^2	3765	221	87	37
		df	55	44	34	25
		T-L	--	.93	.96	.98
Artistic Expression	47	χ^2	2785	139	75	37
		df	55	44	34	25
		T-L	--	.94	.96	.97
Total	177	χ^2	10463	470	206	101
		df	165	132	106	75
		T-L	--	.94	.97	.98

There is a strong suggestion in Table 1 that, despite the predominant single factor shown in Figure 1, a single factor is not sufficient to account for all the correlation among the scales within each group. First, the χ^2 associated with the one-factor solution was significant for each curriculum group, which suggests that the null hypothesis of a single factor was not statistically justified. Secondly, there was a statistically significant improvement in fit by considering two common factors. This can be seen by comparing the reduction in χ^2 relative to the reduction in degrees of freedom within each group. That is, for the first curriculum group the χ^2 dropped from 110 to 44, a reduction of 66, while the degrees of freedom were reduced by 10. The statistical significance of this reduction can be assessed by determining the probability of a χ^2 value of 66 with 10 degrees of freedom. In all three curriculum groups the improvement in fit was significant.

To explore the possible usefulness of a two-factor solution, two factors were extracted by the principal axis method and rotated to a varimax criterion. (For this analysis the covariance matrices were converted to correlation matrices.) The results can be seen in Table 2. The obvious pattern that emerges is that Scales 1 through 5 and 11 formed one factor, while Scales 6 through 10 defined the second factor. The first factor refers to teacher characteristics, while the second refers to the "mechanics" of the course. However, the scales from one factor loaded substantially on the second factor. This, coupled with the comparatively smaller magnitude and contribution of the second factor, persuaded the

Table 2
Rotated Factor-Loading Matrices for the Two-Factor Solution

	Curriculum Group					
	Communication		Man and Society		Artistic Express.	
Scale	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
1	.45	.79	.36	.83	.90	.33
2	.67	.62	.76	.51	.61	.63
3	.44	.83	.41	.84	.84	.33
4	.39	.74	.37	.78	.68	.26
5	.48	.63	.44	.63	.76	.24
6	.81	.41	.79	.37	.55	.63
7	.65	.36	.55	.21	.44	.70
8	.70	.40	.70	.33	.10	.85
9	.82	.49	.66	.39	.55	.77
10	.73	.53	.82	.44	.65	.61
11	.48	.78	.37	.90	.85	.41

authors to accept the one-factor model as an adequate representation of the data.

The fit of the data to Model II requires invariance up to the uniqueness of the variables. It requires that not only the same number of dimensions but the same dimensions account for the covariation within the population. Program SIFASP (Van Thillo & Jöreskog, 1970) was used to assess the fit of the single common factor solution to the second model.

The variance-covariance matrix among scales in each curriculum group was scaled (by SIFASP) in such a way that the sum of three covariance matrices weighted by their respective sample size would yield a matrix with 1's in the diagonal. A one-factor model was fitted simultaneously to the three scaled variance-covariance matrices, with the restriction that the loadings and unique variances be the same in the three populations. The χ^2 from this analysis was 577, with 173 degrees of freedom; the Tucker-Lewis index was .95.

The fit of the one-factor solution to Model II versus Model I can be appraised by the ratio $(\chi_2^2 - \chi_1^2) / (df_2 - df_1)$, where χ_2^2 is the χ^2 obtained from fitting Model II; χ_1^2 is the sum of the χ^2s across curriculum areas under Model I; and df_2 and df_1 stand for the corresponding degrees of freedom. The more closely this ratio approximates unity, the greater is the similarity between the fits to the two models. The difference in χ^2 was $577 - 470 = 107$; the difference in degrees of freedom was $173 - 132 = 41$. The ratio of differences, $107/41 = 2.6$, is moderately large but does not rule out the second model. The more important Tucker-Lewis index is, for the simultaneous solution, virtually the same as it was for the unconstrained single factor hypothesis (i.e., .95 versus .94), suggesting that the additional constraints of equality of loadings and unique variances did not erode the fit.

Table 3 shows the loadings based on the solution for an invariant single factor (λ) as well as the unique variance estimates (ψ) for each scale.

Table 3
Simultaneous Solution for One-Factor model

Variable	Simultaneous Solution* ψ	λ
The instructor clearly presented subject	.25	.85
Was approachable	.19	.88
Got me interested in subject	.22	.86
Raised challenging questions	.40	.75
Expanded course material	.41	.75
Procedures for grades appropriate	.30	.81
Amount of work appropriate	.54	.67
Feedback readily available	.50	.70
Help available when needed	.19	.88
Responsibility clearly defined	.22	.86
Overall teaching ability	.21	.87

*λ refers to the estimated loading of each scale on the single factor, ψ refers to the estimated unique variance for each scale.

It should be noted that the single factor was not dominated by either of the two factors identified in the exploratory analysis.

The reliability of ratings from the 11 items that comprised the invariant factor can be computed with the loadings from the simultaneous solution by means of a formula described in Rock, Werts, and Flaugher (1978):

$$R = (\Sigma\lambda_i)^2\phi_g / [(\Sigma\lambda_i)^2\phi_g + \Sigma\psi_i] \quad [1]$$

where λ_i and ψ_i are the loading and unique variance estimates for the i^{th} scale and ϕ_g is the variance of the factor for the g^{th} group. For these data the computed reliabilities were .959, .957, and 958 for Curriculum Groups 1 through 3, respectively.

Discussion

The covariation of the 11 items used in this study is explainable by a single dimension, which appears to be invariant across populations of instructors from three divergent curriculum areas. Moreover, the reliability of the ratings on this dimension was very high.

The practical implication of these findings is that ratings with this instrument are not affected by bias that would arise from factor structures that varied across these curriculum groups. The scales "mean the same thing" in the different contexts, and the ratings are very reliable. Thus, the results of a previous study (Bejar & Doyle, 1978), which noted significant differences among curricular groups, can be interpreted with more confidence; based on the present analysis, it can be concluded that the rating instrument taps the same dimensions in the three groups.

The theoretical implication is that there is consistency in raters' use of these items across curriculum areas and consistency in the patterns of instructor behavior. These kinds of consistency may facilitate the development of a general theory of rating and theory of instruction. Because they indicate consistency across populations, they also expand the generalizability of validity studies of ratings and teaching, making it less necessary to replicate validations in these different populations.

Further research needs to address the limits of this generalizability. The ratings are consistent across groups within a single, albeit heterogeneous, college; but would similar invariance be found across more disparate groups? Further research might also be helpful for partitioning out the "components of invariance": that portion of the observed consistency rising from factors relevant to the purpose of the measurement, i.e., instructor behaviors and characteristics, and that portion from irrelevant sources, i.e., constant or systematic errors such as halo effect (Brown, 1976, chap. 4; Doyle, 1975, chap. 3).

References

Bejar, I. I. An application of the continuous response level model to personality data. *Applied Psychological Measurement,* 1977, *1,* 509-521.

Bejar, I. I., & Doyle, K. O. Relationship of curriculum area and course format with student ratings of instruction. *American Educational Research Journal,* 1978, *15,* 483-487.

Bejar, I. I., & Doyle, K. O. *Generalizability of factor structures underlying student ratings of instruction.* Paper presented at the 1974 annual meeting of American Educational Research Association, Chicago, September 1974. (ERIC Document Reproduction Service No. ED 033945)

Brown, F. G. *Principles of educational and psychological testing* (2nd ed.). New York: Holt, Rinehart, & Winston, 1976.

Doyle, K. O. Development of the Student Opinion Survey. *Educational and Psychological Measurement,* 1977, *37,* 439-443.

Doyle, K. O. *Student evaluation of instruction.* Lexington MA: Heath, 1975.

Irvine, S. H., & Sanders, J. T. Logic language and method in construct identification across cultures. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptation.* The Hague, Netherlands: Mouton, 1972.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika,* 1969, *34,* 183.

Rock, D. A., Werts, C. E., & Flaugher R. L. The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research,* 1978, *13,* 403-418.

Tucker, L. R., & Lewis, C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika,* 1973, *38,* 1-10.

Van Thillo, M., & Jöreskog, K. G. *SIFASP—A general computer program for simultaneous factor analysis in several populations* (Research Bulletin 70-62). Princeton NJ: Educational Testing Service, 1970.

Whiteley, S. F., & Doyle, K. O. Validity of generalizability of student ratings from between-class and within-class data. *Journal of Educational Psychology,* 1979, *71,* 117-124.

Author's Address

Send requests for reprints or further information to Isaac I. Bejar, Educational Testing Service, Princeton NJ 08541.