The Canadian Journal of Program Evaluation Vol. 13 No. 2 Pages 75–87
ISSN 0834-1516

A CONFIRMATORY FACTOR ANALYTIC MODEL OF EVALUATION USE

Barbara Turnbull
Rutgers University
New Brunswick, New Jersey

Abstract: The purpose of this study was to develop and test a second-order factor analytic model of conceptual, instrumental, and symbolic use of evaluation information. The participants were 308 elementary and secondary school teachers who participated in the British Columbia School Accreditation Program. Structural equation modelling was used to test the fit of the model. The results confirmed a three-factor structure of conceptual, instrumental, and symbolic use. Cross-validation indicated that the model would likely replicate in other independent samples. The model in this study is unique in that it is a first attempt at confirming the factor structure of conceptual, instrumental, and symbolic use in a measure of evaluation use.

Résumé: Le but de cette recherche était de développer et de tester un modèle analytique de facteurs du second ordre de l'utilisation conceptuelle, instrumentale, et symbolique de l'information d'évaluation. Le groupe de participants comprenait 308 enseignants d'écoles primaires et secondaires qui ont participé au British Columbia School Accreditation Program. L'équation structurelle modèle a été utiliseé pour tester l'ensemble du modèle. Les résultats ont confirmé une structure de trois facteurs d'utilisation conceptuelle, instrumentale et symbolique, et la contrevalidation indiquait que le modèle pourrait probablement se repliquer dans d'autres prélèvements indépendants. Le modèle dans cette recherche est unique étant un premier essai pour confirmer la structure de facteurs de l'utilisation conceptuelle, instrumentale et symbolique dans une mesure d'évaluation.

Shadish, Cook, and Leviton (1991) describe a theory of evaluation use as having three components: (a) a description of types of use, (b) an explanation of time frames in which use occurs, and (c) an explanation of what can be done to increase use. The focus of this study is on the first component: types of evaluation use. Types of evaluation use are commonly described as conceptual, instrumental, and symbolic (e.g., Conner, 1981; Leviton & Hughes, 1981). These

terms have become widespread in evaluation literature. Conceptual use is understood as increased understanding and learning about a program as a result of an evaluation and is typically referred to as education-oriented use; instrumental use describes observable and specific changes to a program as the result of an evaluation and is referred to as action-oriented use; and symbolic use occurs when evaluation information is used to persuade others of a predetermined position and is referred to as persuasion-oriented use (Shadish et al., 1991).

Conceptual, instrumental, and symbolic have been used as a framework for most measures of evaluation use (e.g., Anderson, Ciarlo, & Brodie, 1981; Greene, 1988; Ramirez, 1985; Rinnie, 1993). For example, in a study on utilization and participatory evaluation, Greene (1988) developed a 10-item measure of Stakeholder Perceptions of the Usefulness of the Evaluation Process and Results. The measure consisted of two subscales and incorporated items relating to conceptual, instrumental, and symbolic use. The following items taken from Greene's measure (p. 347) illustrate each type of use: (a) conceptual use: "The evaluation results are mainly useful as confirmation of what was already known," (b) instrumental use: "The evaluation results are likely to help agency staff or board members make policy decisions," and (c) symbolic use: "Evaluation results are likely to be useful for persuading others to support the YE/DC program." Similarly, in a more recent unpublished study, Rinnie (1993) developed an evaluation use measure based on the work of Ramirez (1985). The items on Rinnie's measure were classified as cognitive use, behavioral use, and affective use. This classification is conceptually similar to that of conceptual use, instrumental use, and symbolic use. The items on Rinnie's measure are similar in structure to those developed by Greene. For example, a cognitive use item on Rinnie's measure was "Evaluation information confirms the perceived status of the program with no surprises" (p. 130); a behavioral use item was "Evaluation information resulted in new action decided upon and steps toward implementation"; and an affective item was "Evaluation information evoked concern for program status." The items developed by Greene and Rinnie are similar in that they both conform to the definitions of conceptual use as education-oriented, instrumental use as action-oriented, and symbolic use as persuasion-oriented. Other measures of evaluation use such as those developed by Ramirez (1985) and Anderson et al. (1981) are similar in nature to the measures developed by Greene and Rinnie and conform to the same structure of conceptual, instrumental, and symbolic use.

The similarity among items on measures of evaluation use provides evidence for content validity. Content validity is generally described as a "logical relationship of the items to predetermined content areas" (McMillan & Schumacher, 1989, p. 241). In the case of evaluation use measures, the predetermined content areas are conceptual, instrumental, and symbolic use. The logical relationship of the items of the content areas is evidenced through similar applications of education, action, and persuasion-oriented use. Greene (1988), Ramirez (1985), and Rinnie (1993) established content validity through (a) a comprehensive review of appropriate literature, (b) a pilot study, and (c) a review by content area experts to establish whether items adequately represented the content domains of conceptual, instrumental, and symbolic use. Therefore, the work by Greene (1988), Ramirez (1985), and Rinnie (1993) provided evidence for content validity; however, it did not provide evidence for construct validity.

Construct validity is generally explained as "the extent to which certain explanatory concepts (constructs) explain covariation in the responses to the items of the instrument" (Gable & Wolf, 1993, p. 105). As previously described, the content validity of the measures by Greene (1988), Ramirez (1985), and Rinnie (1993) was established through content experts' judgments as to whether the items adequately represented the content domains of conceptual, instrumental, and symbolic use. The similarity in the items across the three measures thereby provided evidence for the content validity of the items. Construct validity, however, is established not by personal judgment, but by determining whether the hypothesized content categories (or constructs) explain the covariance in the item responses. In studies of instrument development, confirmatory factor analysis is a common method used to establish evidence of construct validity. Confirmatory factor analysis (CFA) is a means of describing the relationships between judgmentally developed content categories and the empirically derived constructs. In CFA, a model is developed that describes the relationship between factors and the items that are hypothesized to relate to each factor. The process of confirming a predetermined factor structure is thereby seen as a means of establishing construct validity.

The measures developed by Greene (1988), Ramirez (1985), and Rinnie (1993) did not provide evidence of construct validity. Although Rinnie (1993) reported an internal consistency reliability of $r\ \alpha = 0.75$, no further exploratory or confirmatory analysis was reported. Greene (1988) reported that sample size and sampling issues pre-

vented reliability and validity analysis, and Ramirez (1985) used an open-ended response format and established reliability through consistency in coding but reported no construct validity information. The lack of evidence to support the construct validity of evaluation use was the impetus for this study. The purpose of this study, therefore, was to develop and test a second-order confirmatory factor analytic model of conceptual, instrumental, and symbolic use.

METHODOLOGY

Participants

The participants were 308 elementary and secondary school teachers who participated in the British Columbia (B.C.) School Accreditation Program. B.C. School Accreditation is a provincial school evaluation program sponsored by the Ministry of Education (British Columbia Ministry of Education, 1996) and was chosen because it met the criteria necessary to test the hypotheses in the current study. The process of B.C. School Accreditation is such that all three types of evaluation use are likely to be experienced. For example, teachers are likely to learn about their school as a result of the evaluation (conceptual use), changes are usually made to school practices (instrumental use), and it is common for the accreditation process to increase school advocacy among teachers (symbolic use).

Questionnaire Design

The item pool for the questionnaire was based on previously published and unpublished measures of evaluation use (i.e., Greene, 1988; Ramirez, 1985; Rinnie, 1993). The items included in the model are explained in Table 1 along with the response mean, skewness, and kurtosis. For each item, respondents were asked to rate their perception of evaluation use on a six-point scale (strongly disagree, moderately disagree, disagree slightly, agree slightly, moderately agree, strongly agree). Content validity for the items was established by reviewing the initial draft items for ambiguity and wording by teachers who had and had not previously participated in the Accreditation Program. Revisions were made before the questionnaire was pilot-tested on 28 teachers from one elementary and two secondary schools. Final revisions were made to the items based on the pilot study data.

The Second-Order Confirmatory Factor Analytic Model

Figure 1 depicts the second-order factor analytic model of evaluation use tested in this study and depicts typical drawing conventions employed in structural equation diagrams. Ovals represent latent variables, squares represent observed variables, and straight lines with arrows depict relationships. The equations for the second-order factor model shown in Figure 1 were:

$$\eta = \Gamma \xi + \zeta \quad (1)$$

$$y = \Lambda_y \eta + \varepsilon \quad (2)$$

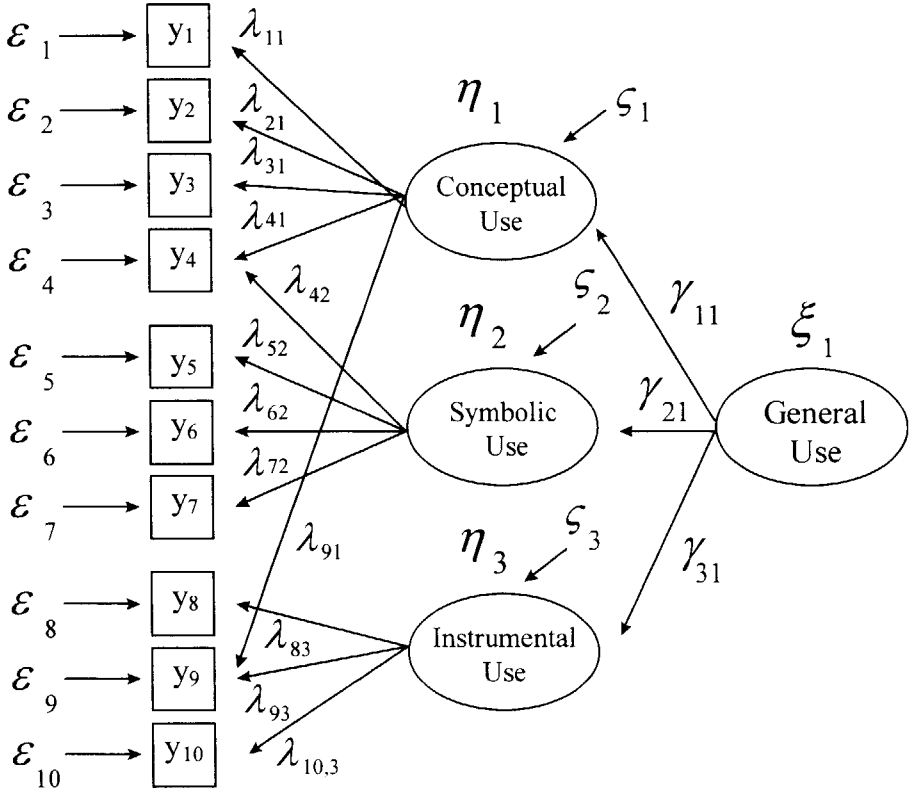
Table 1
Univariate Summary Statistics for Observed Variables

Question	Mean	SD	Skewness	Kurtosis
Conceptual Use (η_1)				
Y1 Because of the accreditation, I learned useful things about my school.	4.44	1.42	-0.95	0.22
Y2 The accreditation process gave me a better understanding of the strengths of our school.	4.68	1.26	-1.33	1.66
Y3 The accreditation activities provided me with necessary knowledge to contribute to the development of our school growth plan.	4.43	1.35	-0.87	0.29
Y4 The accreditation process gave me a better understanding of the weaknesses of our school.	4.26	1.33	-0.92	0.37
Symbolic Use (η_2)				
The accreditation information:				
Y5 is useful for persuading parents to support what we are doing in our school.	4.48	1.11	-0.77	0.88
Y6 convinced staff members of need for maintaining existing school practices.	4.43	1.18	-0.67	0.26
Y7 is useful in gaining community support for what we are doing in our school.	4.36	1.94	-0.88	0.95
Instrumental Use (η_3)				
The accreditation information:				
Y8 resulted in no action to make specific changes in our school practices.	2.06	1.23	1.19	0.74
Y9 resulted in specific suggestions for how to improve the school.	5.11	0.96	-1.45	3.09
Y10 resulted in implementation of new ideas for how to improve school practices.	4.74	1.08	-0.94	1.07

Note: SD = Standard deviation.

The relations between the first and second-order factors are given by Equation 1, and the relations between the observed variables and the latent first-order variables are given by Equation 2. The variable labeled General Evaluation Use (ξ_1) in Figure 1 was an exogenous second-order construct. General Evaluation Use has three latent indicators: (a) Conceptual Use (η_1), which had five observed variables (y_1, y_2, y_3, y_4, y_9); (b) Symbolic Use (η_2), which had four observed variables (y_4, y_5, y_6, y_7); and (c) Instrumental Use (η_2), which had three observed variables (y_8, y_9, y_{10}). The Λ_y matrix was ten by three, with one indicator in each construct fixed to one so the measurement for each latent variable would be the same as the observed variables. The model was tested with maximum likelihood confirmatory factor analysis (CFA) methods used in LISREL8

Figure 1
Second-order Factor Analytic Model



(Joreskog & Sorbom, 1993). Although a number of goodness-of-fit indices are computed through LISREL8, for the purposes of this study only four indices are used: χ^2 (chi-square); Goodness of Fit Index (GFI); Adjusted Goodness of Fit Index (AGFI); and Root Mean Square Residual (RMSR).

RESULTS

Preliminary Analysis

Missing observations ranged from 0.4% to 0.6%, which is considered low enough to have minimal effect on the results (Little, 1987). Missing data for each observed variable were replaced with the variable mean. Multivariate normality was examined through PRELIS (a data screening program included in LISREL8), which provides Mardia's (1970) measure of multivariate kurtosis. The relative multivariate kurtosis for current sample was considered to be acceptable at 0.13. Data were screened for multivariate outliers using SPSS to calculate the Mahalanobis distance (D^2) for each case. An outlier was indicated by a D^2 that was significant at $p < .001$ level. Mahalanobis distance was calculated as chi-square with degrees of freedom equal to the number of observed variables in the hypothesized model (Tabachnick & Fidell, 1989). For the current model, which has 10 observed variables, the chi-square critical value at $p < .001$ was 29.6. D^2 values for 13 cases exceeded the chi-square critical value and were identified as outliers. However, a comparison indicated that skewness, kurtosis, and means were similar between outlier cases and the total sample. Therefore, it was decided that all were accurate observations and none were deleted from the analysis. For the total sample (n = 308), the ratio of cases to observed variables was 30:1, the ratio of cases to estimated parameters was 12:1, and the ratio of cases to latent variables was 77:1.

Analysis of Model Fit

The modelling process involved fitting each set of observed variables to its hypothesized latent construct. The process began with 23 items (nine Conceptual Use items, six Symbolic Use items, and eight Instrumental Use items) that were fitted to three latent constructs of Conceptual, Instrumental, and Symbolic Use. Of the 23 items, those with the highest loadings and the best fit on each latent construct were chosen as the set of marker variables for that particular latent construct (cf. Gable & Wolf, 1993). The process of determining marker

variables was repeated for each latent variable. After each latent construct was fit with a set of marker variables, the model was tested as a whole. The overall model fit was χ^2 (32, n = 308) = 73.50, p = .000, Goodness of Fit Index (GFI) = 0.96, Adjusted Goodness of Fit Index (AGFI) = 0.93, and Root Mean Square Residual (RMSR) = 0.06. This fit was considered unacceptable because the chi-square estimate was significant and thereby the null hypothesis that the data do not differ significantly from the model could not be rejected.

Therefore, two refinements were made to the model. Variable Y4 was allowed to load on Symbolic Use (η_1), and variable Y9 was allowed to load on Conceptual Use (η_2). The decision to add these particular paths was based on a thorough understanding of the variables as well as inspection of the residuals and modification indices. The overall fit of the model with the two additional paths was considered acceptable: χ^2 (30, n = 308) = 41.24, p = .08, GFI = 0.98, AGFI = 0.95, and RMSR = 0.03. The chi-square was not significant, all paths were statistically significant at $p < .05$, and the GFI and AGFI were all at acceptable levels. As well, the RMSR was low, which indicated small residuals between the implied Σ and observed S covariance matrices. Furthermore, a substantial portion of the variation in the three-factor first-order model was accounted for by the second-order general use factor as indicated by the large standardized loadings: γ_{11} = 0.89 for Conceptual Use, γ_{21} = 0.82 for Symbolic Use, and γ_{31} = 0.79 for Instrumental Use. Therefore, the null hypothesis that the data do not differ significantly from the model was accepted. From this, it was posited that the model shown in Figure 1 adequately represents the data. Parameter estimates, standard errors, t-values, and squared multiple correlations are reported in Table 2, and the covariance matrix for the second-order factor analytic model is reported in Table 3.

Cross-Validation

The single sample cross-validation strategy used in this study was based on Joreskog and Sorbom (1988). In single sample cross-validation, the total sample is usually divided into random samples or samples based on logical categorical variables. Establishing invariance across samples is seen as a means of substantiating the generalizability of the model. In the current study, it seemed appropriate that the sample be divided into elementary and secondary teacher samples and random samples. It was hypothesized that the model would fit equally well across both sets of samples. The proc-

ess described by Joreskog and Sorbom (1988) involved a series of increasingly restrictive invariance tests. The first and least constraining test was that both groups have the same factor pattern of

Table 2
Parameter Estimates for the Second-Order Factor Analytic Model

Lambda (λ)	Estimate	Standardized Estimate	Standard Error	*t*-value	SMC[a]
1,1	1.000*	1.218	—	—	0.737
2,1	0.798	0.972	0.049	16.419	0.600
3,1	0.964	1.174	0.049	19.762	0.765
4,1	1.116	1.359	0.090	12.463	0.666
9,1	-0.249	-0.303	0.084	-2.967	—
4,2	-0.500	-0.429	0.119	-4.197	—
5,2	1.000*	0.858	—	—	0.600
6,2	9.17	0.787	0.078	11.689	0.444
7,2	1.238	1.062	0.082	15.083	0.791
8,3	-0.760	-0.692	0.072	-10.579	0.317
9,3	1.191	1.086	0.142	8.418	0.870
10,3	1.000*	0.911	—	—	0.719
Gamma (γ)					
1,1	1.089	0.894	0.074	14.630	0.799
2,1	0.705	0.822	0.060	11.829	0.675
3,1	0.719	0.789	0.058	12.409	0.622

Note. [a]SMC = Squared multiple correlation, an estimate of lower bound reliability for each item. *t*-values > 2.0 are significant at $p < .05$. *Indicates that parameter value was set to 1.0. Dashes indicate value wasn't estimated.

Table 3
Covariance Matrix for Factor Analytic Model

	1	2	3	4	5	6	7	8	9	10
1	2.012									
2	1.192	1.575								
3	1.435	1.139	1.823							
4	1.254	1.029	1.224	1.765						
5	0.715	0.537	0.709	0.528	1.227					
6	0.676	0.560	0.715	0.338	0.686	1.394				
7	0.958	0.759	0.966	0.615	0.920	0.813	1.427			
8	-0.603	-0.489	-0.486	-0.432	-0.444	-0.392	-0.511	1.514		
9	0.571	0.463	0.516	0.436	0.451	0.430	0.482	-0.608	0.927	
10	0.830	0.614	0.702	0.670	0.555	0.482	0.584	-0.633	0.794	1.156

fixed and free parameters ($\Lambda_{Pattern}$). If the model did not have satisfactory fit at this level there would be no point testing the subsequent hypotheses. However, if the model had adequate fit then the next test (H_Γ) would be to constrain the loadings of the first-order factors on the second-order factor to be the same. For the model in this study, this meant that $\gamma_{11}^{(1)} = \gamma_{11}^{(2)}$, $\gamma_{21}^{(1)} = \gamma_{21}^{(2)}$, $\gamma_{31}^{(1)} = \gamma_{31}^{(2)}$. The next test ($H_{\Gamma\Lambda y}$) constrained the factor loadings (λ_i) for the observed variables (Yi) to be equal across samples. The final and most restrictive test ($H_{\Gamma\Lambda y\theta\varepsilon}$) added the constraint of equal errors (ε_i) across samples.

Table 4 includes the chi-square (χ^2), *p*-value, degrees of freedom, GFI, and RMSR for each of the hierarchical invariance tests. The chi-square for the elementary/secondary groups for the first test ($H_{Pattern}$) was χ^2 (60, n = 173/135) = 61.07, $p > .05$, which indicated good model fit. The GFIs were greater than 0.95 and the RMSRs were acceptable at 0.05. This result suggested that the factor pattern for the model was invariant across samples. The test for invariance of gamma (H_Γ) also had a non-significant chi-square, χ^2 (63, n = 173/135) = 71.27, $p > .05$, and the GFIs were also greater than .95. In this test, the RMSRs were higher than the generally accepted level of 0.05. The third invariance test, which added the constraint of factor loadings ($H_{\Gamma\Lambda y}$) was also not significant:

Table 4
Invariance of Factor Structures

Group	χ^2	*p*-value	*df*	GFI (a) / (b)	RMSR (a) / (b)
(a) Elementary (n = 173)					
(b) Secondary (n = 135)					
$H_{Pattern}$	61.07	.44	60	.96/.96	.05/.05
H_Γ	71.27	.22	63	.95/.95	.18/.14
$H_{\Gamma\Lambda y}$	92.44	.26	72	.93/.96	.26/.19
$H_{\Gamma\Lambda y\theta\varepsilon}$	110.99	.00	82	.92/.95	.25/.19
(a) 1–10 years teaching (n = 119)					
(b) 11–25 years teaching (n = 189)					
$H_{Pattern}$	98.77	.001	60	.95/.93	.06/.05
H_Γ	99.65	.002	63	.95/.93	.06/.06
$H_{\Gamma\Lambda y}$	113.55	.001	72	.95/.92	.07/.08
$H_{\Gamma\Lambda y\theta\varepsilon}$	141.95	.000	82	.94/.90	.07/.09

Note. χ^2 = Chi-square; GFI = Goodness of fit index. RMSR = Root mean square residual.

χ^2 (72, n = 173/135) = 92.44, p > .05, and again the GFI were acceptable, but the RMSRs were high. The final test of invariance ($H_{\Gamma\Lambda y\theta\varepsilon}$), which added the constraint of equal errors across samples, was significant, χ^2 (82, n = 173/135) = 110.99, p < .05. Again, the GFIs were acceptable and the RMSRs were high. Although the invariance tests for the elementary/secondary samples showed a pattern of high RMSRs, the not significant chi-square estimates and the acceptably high GFI estimates suggested that the model would likely replicate in another independent sample.

The chi-square for the years of teaching groups for the first test ($H_{Pattern}$) was χ^2 (60, n = 119/189) = 98.77, p < .05, which indicated poor model fit. However, the GFIs and RMSRs were acceptable. This pattern was evident in all of the subsequent tests of invariance for the years of teaching groups. Therefore, it appeared that the model adequately represented the data for the elementary/secondary groups, but it did not represent the data for the years of teaching groups. The variance of the model across years of teaching groups is an interesting finding; however, cross-validation in a second sample for both elementary/secondary and years of teaching groups is advised before any further inferences are made regarding differences in perceptions of use and years of teaching experience.

CONCLUSIONS AND DISCUSSION

For nearly two decades, researchers have developed measures of evaluation use based on the concepts of conceptual, symbolic, and instrumental use. The purpose of this study was to review the constructs of conceptual, instrumental, and symbolic use as used in prior measures, and to conduct a CFA of a model of evaluation use. Confirmation of the factor structure was deemed important because it would provide construct validity evidence for a measure of evaluation use. The results confirmed the three-factor structure of conceptual, instrumental, and symbolic use and thereby provide evidence of construct validity for a measure of evaluation use. There is, however, a limitation to the argument of a second-order factor of General Use (see Figure 1). The model depicted in Figure 1 is an example of a second-order factor that predicts a just-identified structural model. In order for the model to be over-identified, an additional first-order latent factor was needed. An additional latent variable would have provided more accurate measures of fit for the second-order factor and a more convincing argument for the existence of the second-order General Use factor. However, based on the cur-

rent theory of evaluation use, there are only three types of evaluation use, so it was not possible to add a fourth latent variable to the model. Therefore, until a fourth category of use is introduced, the model will remain just identified.

REFERENCES

Anderson, C.D., Ciarlo, J.A., & Brodie, S.F. (1981). Measuring utilization for mental health program consultation. In J.A. Ciarlo (Ed.), *Utilization evaluation: Concepts and measurement techniques* (pp. 97–124). Beverly Hills, CA: Sage Publication.

British Columbia Ministry of Education, Skills and Training. (1996). *British Columbia school accreditation: Guide for schools.* Victoria, BC: Author.

Conner, R.F. (1981). Measuring evaluation utilization: A critique of different techniques. In J.A. Ciarlo (Ed.), *Utilization evaluation: Concepts and measurement techniques* (pp. 59–76). Beverly Hills, CA: Sage Publication.

Gable, R.K., & Wolf, J.W. (1993). *Instrumental development in the affective domain: Measuring attitudes and values in corporate and school settings* (2nd ed.). Boston: Kluwer Academic Publishers.

Greene, J.G. (1988). Communication of results in utilization on participatory program evaluation. *Evaluation and Program Planning, 11,* 341–351.

Joreskog, K.G., & Sorbom, D. (1988). *LISREL7: A guide to the program and applications* (2nd ed.). Chicago: SPSS Inc.

Joreskog, K.G., & Sorbom, D. (1993). *LISREL8 user's reference guide.* Chicago: SPSS Inc.

Leviton, L.C., & Hughes, E.F.X. (1981). Research on utilization of evaluations: A review and synthesis. *Evaluation Review, 5,* 525–548.

Little, R.J. (1987). *Statistical analysis with missing data.* Toronto: Wiley & Sons.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

McMillan, J.H., & Schumacher, S. (1989). *Research in education* (2nd ed.). Glenview, IL: Scott, Foresman, & Company.

Ramirez, N.A. (1985). *Development of an evaluation feedback process and an evaluation utilization assessment instrument* [CD-ROM]. Abstract from: ProQuest File: Dissertation Abstracts Item: 8522749.

Rinnie, C. (1993). *The impact of anxiety as a mediating variable on health educators' utilization of evaluation results* [CD-ROM]. Abstract from: ProQuest File: Dissertation Abstracts Item: 9334464.

Shadish, W.R., Cook, T.D., & Leviton, L.C. (1991). *Foundations of program evaluation: Theories of practice.* Newbury Park, CA: Sage.

Tabachnick, B.G., & Fidell, L.S. (1989). *Using multivariate statistics.* Northridge, CA: HarperCollins.