2007

Experimental Analysis of Question Wording in an Instrument Measuring Teachers' Attitudes Toward Inclusive Education

Robert F. Dedrick
University of South Florida

Kofi Marfo
Aga Khan University, kofi.marfo@aku.edu

Deborah M. Harris
University of South Florida

Recommended Citation

Dedrick, R. F., Marfo, K., Harris, D. M. (2007). Experimental Analysis of Question Wording in an Instrument Measuring Teachers' Attitudes Toward Inclusive Education. *Educational and Psychological Measurement, 67*(1), 116-131.
Available at: https://ecommons.aku.edu/eastafrica_ihd/15

Educational and Psychological Measurement
Volume 67 Number 1
February 2007 116-131

10.1177/0013164406292034
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Experimental Analysis of Question Wording in an Instrument Measuring Teachers' Attitudes Toward Inclusive Education

Robert F. Dedrick
Kofi Marfo
Deborah M. Harris
University of South Florida

An experimental study (n = 288 general and special education teachers) examining the effects of altering the referent ("students with mild disabilities," "students with severe disabilities," or "students with disabilities") on a four-item scale (Negative Effect of Inclusion) indicated that wording changes had little effect on the scale's psychometric properties (e.g., factor pattern coefficients). Changes did result in a shift in the mean level of the attitude scale. Regression coefficients between the scale and type of teacher, total years of teaching experience, years of experience at current school, and training in inclusion were not significantly altered by changing the referent. Gender was the only predictor that exhibited a lack of invariance in its regression coefficients across questionnaire forms that differed in referent. For most of the bivariate relationships examined in this study, the same conclusions would be drawn no matter which of the three referents were used.

Keywords: *attitudes toward inclusion; question wording; measurement invariance; multigroup confirmatory factor analysis*

An inclusive education philosophy is widespread in schools across the country, although there is significant variability in the types of inclusive service delivery models used and the numbers of children with disabilities served primarily in general education classrooms. Notwithstanding this variability, many researchers and policy

Authors' Note: This study was supported by Grant H324D990034 from the Office of Special Education Programs, U.S. Department of Education, to Deborah M. Harris and Kofi Marfo. We are grateful to the National Technical Advisory Board (Matthew Brown, Lynne Cook, Marilyn Friend, and Deborah Voltz) and our research team (Brenda Curtwright, Tricia Spears, Deidre Justice, Maureen Artalona, Kim Leopold, Jessica Podolski, Danielle Bianco, and Isaac Bristol). Special thanks to Donnie Evans. Correspondence concerning this article should be addressed to Robert F. Dedrick, University of South Florida, Department of Measurement and Research, EDU 162, Tampa, FL 33620; e-mail: dedrick@tempest.coedu.usf.edu.

makers have acknowledged that teachers' beliefs and attitudes toward inclusion are critical to the success of inclusive education. With this acknowledgement has come a proliferation of research on teachers' attitudes toward inclusion and the development of survey instruments designed to measure these attitudes (Antonak & Livneh, 1988). These instruments include the Inclusive School Program Survey (McLeskey, Waldron, & So, 2001), Opinions Relative to Integration of Students With Disabilities (Antonak & Larrivee, 1995), the Attitudes Toward Inclusive Education Scale (Wilczenski, 1995), and the Teacher Integration Attitude Questionnaire (Sideridis & Chandler, 1995).

Although these instruments measure similar constructs (e.g., the benefits of integration), the items differ in the levels of specificity of the disabilities referred to in the attitude statements. For example, Opinions Relative to Integration of Students With Disabilities contains the generic phrase "students with disabilities" as the referent for the attitude statements (e.g., "The integration of students with disabilities can be beneficial for students without disabilities"). In contrast, the Inclusive School Program Survey uses the more specific phrase "students with mild disabilities" as the referent for the items (e.g., "Students with mild disabilities improve their social skills when placed in a general education classroom"), whereas the Attitudes Toward Inclusive Education Scale uses specific social, physical, academic, and behavioral characteristics to describe the referent for the attitude statement (e.g., "Students who cannot move without help from others should be in regular classes").

Methodological research on the effects of question wording conducted outside special education has shown that even small wording changes can lead to very different results and conclusions (Schuman & Presser, 1996; Schwarz, 1999). For example, an experimental analysis of changing one word ("Do you think the United States should *allow* public speeches against democracy?" vs. "Do you think the United States should *forbid* public speeches against democracy?") found a large effect on the univariate response distribution; "approximately 20% more people were willing to 'not allow' such speeches than were willing to 'forbid' them" (Rugg, 1941, p. 92).

More recently, methodological research has moved from examining the effects of question wording on univariate distributions (e.g., means, marginal percentages) to examining whether changes in question wording affect the internal relationships among attitude items (e.g., factor structure) or the relations between attitude items and external variables (e.g., educational level) (Schuman & Presser, 1996). The fact that wording changes in attitude items produce a level shift does not necessarily mean that these changes will alter the internal relations between attitude items (e.g., factor structure) or the relations of items with external variables. For example, the correlation between two attitude items answered in terms of students with *severe* disabilities will be identical to the correlation between the same two attitude items answered in terms of students with *mild* disabilities if the wording change produces a systematic shift in the responses to the items. In contrast, if a wording change produces a change in response that is not systematic, the strength of the relation

between the items may change, thus altering the psychometric properties of the items used to represent the underlying construct. This type of change threatens the validity of the scores and poses serious challenges to researchers attempting to build explanatory models that incorporate these measures.

Given that many of the philosophical arguments on inclusion have revolved around the differentiation between mild and severe disability, there is a need, from a methodological standpoint, to examine what effects changes from *mild disability* to *severe disability* to *disability* have on responses to attitude items on a survey instrument. To address this need, four items (see Table 1) measuring a unidimensional construct (the Negative Effect of Inclusion) from the Beliefs/Perceptions About Inclusive Education Scale (B-PIES; Marfo, Harris, & Dedrick, 2002) were experimentally manipulated by altering the reference in the attitude statement to "students with mild disabilities," "students with severe disabilities," or "students with disabilities." Participating general and special education teachers were randomly assigned to the conditions of the manipulated independent variable (i.e., type of referent), and the effects of these manipulations were evaluated in terms of the measurement properties of the items and teachers' mean levels of response to the items on the Negative Effect of Inclusion scale. The following three research questions were examined:

1. What effect does altering the referent have on the psychometric properties of the scale (factor pattern coefficients, error variances for items, and factor variances)?
2. What effect does altering the referent have on the relationships between the scale and the following external variables: teacher gender, type of teacher (general educator, special educator, or other), total years of teaching experience, years of experience at current school, and involvement in training for inclusion (yes or no)?
3. What effect does altering the referent have on teachers' mean response levels to the scale?

Method

Instrument and Experimental Conditions

The four items measuring the Negative Effect of Inclusion were part of the B-PIES, a 41-item instrument that used a 5-point, Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*) to measure general and special education teachers' perceptions of various issues related to inclusion. The B-PIES was developed in stages beginning with reviews of the literature on the philosophical, policy, and instructional issues related to inclusion. Initial themes and items generated for the B-PIES were reviewed by the National Technical Advisory Board, field professionals from a local school district, and members of the research team. Subsequent psychometric analyses of the B-PIES have provided support for the reliability and validity of the scores from the instrument (Marfo et al., 2002).

Table 1
Descriptive Statistics for Items Measuring Negative Effect of Inclusion by Form

Item	Form A: *Mild*	Form B: *Severe*	Form C: *Generic*
1. Including students with ____ in the general education classroom places an undue burden on general education teachers.			
M (*SD*)	2.49 (1.04)	3.80 (1.22)	3.18 (1.36)
Skewness/kurtosis	0.67/–0.06	–0.69/–0.70	–0.07/–1.26
Corrected item-to-total *r*	.62	.58	.69
2. Schools which include students with ____ in the general education classroom risk lowering their performance on statewide and national tests.			
M (*SD*)	2.84 (1.09)	3.30 (1.22)	3.05 (1.14)
Skewness/kurtosis	0.18/–0.85	–0.17/–1.05	0.11/–0.82
Corrected item-to-total *r*	.58	.68	.63
3. Including students with ____ in the general education classroom is a problem because the academic curriculum is too demanding for these students.			
M (*SD*)	2.56 (0.96)	3.38 (1.22)	3.02 (1.07)
Skewness/kurtosis	0.79/0.05	–0.25/–1.15	0.08/–0.69
Corrected item-to-total *r*	.71	.66	.67
4. Including students with ____ in the general education classroom can affect their self-concept negatively.			
M (*SD*)	2.46 (1.00)	3.13 (1.15)	2.77 (1.11)
Skewness	0.83/0.34	0.01/–0.88	0.04/–0.80
Corrected item-to-total *r*	.63	.62	.63
Negative Effect of Inclusion			
M (*SD*)	2.59 (0.82)	3.40 (0.96)	3.01 (0.95)
Skewness/kurtosis	0.71/0.61	–0.09/–0.83	–0.11/–0.35

Note: Negative Effect of Inclusion was created by averaging the four items. Sample sizes for Forms A, B, and C were 102, 104, and 82, respectively. Depending on the form, the underlined part of the item contained the referent "mild disability," "severe disability," or "disability."

The four items selected from the B-PIES for experimental manipulation were chosen because they were viewed as ones for which differential dispositions toward mild and severe disability had the potential to influence respondents' ratings. Three experimental variations (Forms A, B, and C) were examined. Each variation of the instrument had the same set of demographic and professional profile items, while differing from the others in terms of the specific labels used on the four-item attitudinal scale. Form A used the referent "students with mild disabilities," Form B used the referent "students with severe disabilities," and Form C used the undifferentiated label "students with disabilities." Cronbach's α values

and 95% confidence intervals for the reliability coefficients for the four-item scale were .81 (.75 to .87) for Form A, .81 (.75 to .87) for Form B, and .83 (.76 to .88) for Form C.

Participants

The study was conducted in one Florida school district with a population of more than 900 general and special education teachers in 21 elementary, middle, and high schools. Within each school, teachers were stratified into general and special education teachers and then randomly assigned to one of three experimental conditions (Form A, B, or C). This method of random assignment was used to ensure that general and special education teachers would be represented within each of the three experimental conditions in roughly the same ratio with which they were represented in their schools, and each school within the district would have teachers across all three variations of the survey instrument.

Surveys were distributed toward the end of the academic year, a typically busy period for instructional staff members. The return rates for Forms A (n = 102) and B (n = 104) were approximately 44% each; Form C's response rate was 36% (n = 82). Demographically, 53% of the teachers were from elementary schools, 27% from high schools, and 20% from middle schools. The majority of the participants were women (82%). Sixty-nine percent were general education teachers, 20% were special education teachers, and 11% were other types of teachers (e.g., a combination of general and special education). Total years of teaching experience ranged from 0.5 to 36 years (M = 13.49 years, SD = 9.30 years). Years teaching at one's current school ranged from 0.5 to 32 years (M = 6.93 years, SD = 6.69 years). Comparisons of these characteristics across the three experimental conditions indicated no statistically significant differences ($p > .05$), thus supporting the effectiveness of the randomization procedures for controlling various extraneous variables (a demographic comparison table is available on request).

Data Analysis

Research Question 1 focused on the measurement invariance of the four items of the Negative Effect of Inclusion scale across the three forms of the instrument. Multigroup confirmatory factor analysis was used to test the equality of the factor pattern coefficients (loadings) and error variances associated with each observed variable. All confirmatory factor analyses were based on the covariance matrix of the observed variables and used maximum likelihood estimation conducted using Mplus version 3.0 (Muthén & Muthén, 1998-2004). The Negative Effect of Inclusion factor was scaled by fixing the first factor pattern coefficient to 1.00.

Before conducting invariance tests, the fit of the one-factor model underlying the Negative Effect of Inclusion was evaluated for each form. The fit of the models was evaluated using the χ^2 test, Bentler's (1990, 1992) normed comparative fit index

(CFI), and the standardized root mean square residual (SRMR). Hu and Bentler's (1999) cutoff values of ≥.95 for the CFI and ≤.08 for the SRMR were used as general indicators of acceptable fit of the models; however, substantive issues such as the interpretability of the parameter estimates were also considered. This approach is consistent with Hu and Bentler's (1998) recommendations, derived from their Monte Carlo studies evaluating fit indices. On the basis of these studies, they concluded that "although our discussion has been focused on issues regarding overall fit indices, consideration of other aspects such as the adequacy and interpretability of parameter estimates, model complexity, and many other issues remains critical in deciding on the validity of a model" (p. 450). Marsh, Hau, and Wen (2004) offered similar recommendations, arguing that the complexity involved in interpreting fit indices precludes universal, rigid criteria.

Although the root mean square error of approximation (RMSEA; Steiger, 1990) is a widely used measure of fit, it was not used in the present study to evaluate model fit because of the relatively small samples in each group (n = 102, 104, and 82). Rigdon (1996) showed that although the quality of the RMSEA estimate is good in large samples, the estimate is not as good with small samples. Rigdon noted that "when sample size is low, RMSEA may suggest rejecting a model that otherwise would be accepted" (p. 375) and therefore recommended its use with larger samples. Hancock and Freeman (2001) offered similar cautions about using the RMSEA as a measure of fit for smaller confirmatory factor models with small sample sizes.

Following the evaluation of fit of the one-factor model for each form, measurement invariance across forms was examined. The equality of factor pattern coefficients and error variances associated with each observed variable was tested using a series of hierarchically ordered models of increasing restrictiveness. The least restrictive invariance model was Model 1, in which no equality constraints were imposed on the factor pattern coefficients, error variances, or factor variance across forms. Model 2, a more restrictive model in which equality constraints on the factor pattern coefficients for the items were imposed, was compared with Model 1 to evaluate the invariance of the factor pattern coefficients. Model 3, which imposed equality constraints on the measurement error variances, was compared with Model 2 to evaluate invariance of the error variances. As an additional test of invariance, the equality of the factor variances was examined by imposing equality constraints on the factor variance for Negative Effect of Inclusion (Model 4) and comparing the fit of this model with Model 3.

The strategy used to evaluate the various levels of measurement invariance was to compare the nested likelihood ratio χ^2 difference ($\Delta\chi^2$) relative to the difference in the degrees of freedom (Δdf) for the models being compared. These tests were supplemented by comparing the changes in the CFI and SRMR along with their actual values to determine if the equality constraints produced unacceptable fit on the basis of the guidelines (≥.95 for the CFI and ≤.08 for the SRMR) of Hu and Bentler (1999).

Research Question 2 examined if altering the referent in the attitude statements changed the relationship between the Negative Effect of Inclusion and a series of

external variables (e.g., teacher gender). This question was motivated by the fact that many research studies focus on relationships between constructs such as the Negative Effect of Inclusion and characteristics of study participants. Of interest in this study was whether the questionnaire form would alter the relationships between these variables. This question is equivalent to evaluating a series of statistical interactions between the form of the questionnaire and each external variable. To address this question, multigroup confirmatory factor analysis was used. Two models were run for each external variable. In the first model, the Negative Effect of Inclusion was regressed on the external variable within each form of the questionnaire, and the regression coefficients for this relationship were free to vary across forms. In the second model, the regression coefficients were constrained to be equal. As with previous analyses, the strategy used to evaluate the invariance of the regression coefficients was to compare the nested likelihood ratio χ^2 difference ($\Delta\chi^2$) relative to the difference in the degrees of freedom (Δdf) for the models being compared. Additionally, invariance was evaluated by comparing changes in the CFI and the SRMR for the least restrictive model (i.e., regression coefficients were free to vary) and the more restrictive model (i.e., regression coefficients were constrained to be equal). Evidence of a lack of invariance for the regression coefficients for an external variable would indicate a statistical interaction (i.e., form of questionnaire by external variable) and suggest that the relationship between the Negative Effect of Inclusion and the external variable was moderated by the form of the questionnaire.

Research Question 3 examined if altering the referent in the attitude statements had an impact on teachers' mean response levels to the Negative Effect of Inclusion. Multigroup confirmatory factor analysis with mean and covariance structure (MACS) was used to address this question. In these analyses, the means of the measured variables were included in the model (in the analyses addressing Research Questions 1 and 2, the observed variables were centered on their means, and thus the means of the observed variables were zero). The equation for a single group representing the vector of scores for the four attitudinal items ($\mathbf{X}$) and the one latent factor (ξ) of Negative Effect of Inclusion is

$$\mathbf{X} = \boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x\xi + \boldsymbol{\delta},$$

where $\boldsymbol{\tau}_x$ is a 4×1 vector of intercept parameters, $\boldsymbol{\Lambda}_x$ is a 4×1 vector of the factor pattern coefficients, and $\boldsymbol{\delta}$ is a 4×1 vector of uniqueness associated with the measured variables. In the MACS model, the expected value of $\mathbf{X}$ (the mean of the observed variables) equals $\boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x\kappa$, where κ is the mean of the latent variable of Negative Effect of Inclusion. When the latent variable mean is set to zero, the item intercept represents the estimated mean of the observed variable. Additional details of the models used to compare the item intercepts and the latent mean of Negative Effect of Inclusion across forms are provided in the results for Research Question 3.

Results

Research Question 1: Measurement Invariance

The χ^2 values for the one-factor model for Forms A and B were statistically significant ($p < .05$), whereas the χ^2 value for Form C was not statistically significant ($p = .361$). The results for the CFIs (.967, .948, and 1.000) and SRMRs (.039, .047, and .022) for Forms A, B, and C, respectively, however, indicated that the fit of the one-factor model for each form was acceptable according to the general guidelines presented by Hu and Bentler (1999) (Table 2).

Following the examination of the one-factor model for each questionnaire form, a baseline model (Model 1) was established by fitting a multigroup model with no equality constraints. The χ^2 value and degrees of freedom for this model were simply the sums of the individual χ^2 values and associated degrees of freedom for the models run separately by form. Although the χ^2 value for Model 1 was statistically significant, $\chi^2(6, N = 288) = 17.864$, $p = .007$, the alternative measures of fit indicated acceptable fit (CFI = .970, SRMR = .038). All factor pattern coefficients (i.e., loadings) were significantly different from zero ($p < .01$).

To test the invariance of the factor pattern coefficients, Model 2, in which the factor pattern coefficients were constrained to be equal, was compared with Model 1. The results indicated that constraining the pattern coefficients to be equal did not result in a significant decline in model fit ($\Delta\chi^2 = 6.895$, $\Delta df = 6$, $p > .05$; see Table 3). The CFI of .967 (ΔCFI = .003) and the SRMR of .067 (ΔSRMR = .029) indicated acceptable fit for Model 2. Taken together, these results suggest that there was insufficient evidence to reject the null hypothesis of equal factor pattern coefficients, and therefore, the factor pattern coefficients may be viewed as reasonably invariant across the forms.

To test the invariance of the error variances, Model 3, in which the error variances were constrained to be equal, was compared with Model 2. Results of the change in χ^2 relative to the change in degrees of freedom indicated that constraining the error variances to be equal did not result in a significant decline in model fit ($\Delta\chi^2 = 13.839$, $\Delta df = 8$, $p > .05$; see Table 3). The CFI of .952 (ΔCFI = .015) and the SRMR of .084 (ΔSRMR = .017) indicated acceptable fit for Model 3. Taken together, the results suggest that there was insufficient evidence to reject the null hypothesis of equal error variances, and therefore, the error variances may be viewed as reasonably invariant across the forms.

Finally, as an additional test of invariance, the variances of the Negative Effect of Inclusion were constrained to be equal across forms, and the fit of the model (Model 4) was compared with that of Model 3. The results indicated that constraining the factor variances to be equal did not result in a significant decline in model fit ($\Delta\chi^2 = 3.090$, $\Delta df = 2$, $p > .05$; see Table 3). The CFI of .950 (ΔCFI = .002) indicated acceptable fit for Model 4; however, the SRMR of .139 (ΔSRMR = .055) suggested less

Table 2
Parameter Estimates for Baseline Model for One-Factor Negative Effect of Inclusion Multigroup Confirmatory Factor Analysis With No Equality Constraints

	Form A: *Mild*		Form B: *Severe*		Form C: *Generic*	
Item[a]	Factor Pattern Coefficient	Error Variance	Factor Pattern Coefficient	Error Variance	Factor Pattern Coefficient	Error Variance
1. Burden	1.000	0.602	1.000	0.852	1.000	0.673
2. Performance	1.015	0.686	1.179	0.622	0.742	0.652
3. Too demanding	1.169	0.269	1.184	0.604	0.751	0.468
4. Self-concept	1.094	0.429	1.049	0.631	0.720	0.626
Factor variance	0.472		0.616		1.159	
χ^2	6.580		9.264		2.020	
p	.0364		.0095		.3606	
CFI	.967		.948		1.000	
SRMR	.039		.047		.022	

Note: Sample sizes for Forms A, B, and C were 102, 104, and 82, respectively. For each form, the χ^2 value had two degrees of freedom. All factor pattern coefficients were significantly different from zero ($p < .01$). CFI = comparative fit index; SRMR = standardized root mean square residual.
a. See Table 1 for a complete listing of the items.

Table 3
Model Fit Indices Resulting From Factorial Invariance Tests for Forms A (*mild*), B (*severe*), and C (*generic*)

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	SRMR
1. No equality constraints	17.864	6	—	—	.970	.038
2. Equal factor pattern coefficients	24.759	12	6.895	6	.967	.067
3. Equal measurement error variance	38.598	20	13.839	8	.952	.084
4. Equal factor variance	41.688	22	3.090	2	.950	.139

Note: CFI = comparative fit index; SRMR = standardized root mean square residual. None of the $\Delta\chi^2$ values was statistically significant at the .05 level.

than acceptable fit. Taken together, these results suggest some differences in the variances of the factor across forms. Examination of the variances of the Negative Effect of Inclusion across forms for Model 3 (equal factor pattern coefficients and error variances) showed that the smallest variance was for Form A (variance = 0.536, *SE* = 0.118), followed by Form C (variance = 0.748, *SE* = 0.169) and Form B (variance = 0.808, *SE* = 0.168). Follow-up comparisons indicated that the largest difference was between Form A and Form B, with $\Delta\chi^2 = 2.861$, $\Delta df = 1$, $p > .05$, and CFI = .948 (ΔCFI = .004) and SRMR = .133 (ΔSRMR = .049).

Table 4
Invariance Tests of Relationships Between Each External Teacher Predictor Variable and Negative Effect of Inclusion Across Forms A (*mild*), B (*severe*), and C (*generic*)

Predictor	Model	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	SRMR
Female	Free	53.566	29	—	—	.939	.079
	Equal	63.971	31	10.405**	2	.919	.112
Type of teacher	Free	52.245	38	—	—	.961	.067
	Equal	54.056	42	1.811	4	.967	.075
Total years of teaching experience	Free	49.787	29	—	—	.947	.080
	Equal	54.951	31	5.164	2	.939	.080
Years of experience at current school	Free	44.919	29	—	—	.959	.075
	Equal	49.489	31	4.570	2	.952	.096
Training in inclusion	Free	46.155	29	—	—	.956	.075
	Equal	46.542	31	0.387	2	.960	.078

Note: Sample sizes varied for each predictor. For gender, Forms A to C had 102, 103, and 82 cases, respectively. For type of educator, Forms A to C had 96, 100, and 81 cases, respectively. For total years of experience Forms A to C had 100, 100, and 80 cases, respectively. For years at current school, Forms A to C had 100, 100, and 80 cases, respectively. For training in inclusion, Forms A to C had 100, 104, and 80 cases, respectively. Free = regression coefficients were free to vary across forms; equal = regression coefficients were constrained to be equal across forms. In all models, factor pattern coefficients and error variances were constrained to be equal. CFI = comparative fit index; SRMR = standardized root mean square residual.
**$p < .01$.

Research Question 2: Invariance of the Regression Coefficients for Negative Effect of Inclusion on External Teacher Variables

Table 4 summarizes the model fit indices for the invariance tests of the regression coefficients for the Negative Effect of Inclusion on each of the teacher demographic variables (in all models, factor pattern coefficients and error variances were constrained to be equal on the basis of previous analyses supporting the tenability of this specification). In the first analysis, Negative Effect of Inclusion was regressed on teacher gender (a dummy variable with 1 = female and 0 = male) within each form, and the regression coefficients were allowed to vary across forms. The fit of this baseline model was marginal, $\chi^2(29, N = 287) = 53.566$, $p = .004$, CFI = .939, SRMR = .079. When the regression coefficients were constrained to be equal there was a significant decline in the fit of the model ($\Delta\chi^2 = 10.405$, $\Delta df = 2$, $p < .01$, CFI = .919, ΔCFI = .020, SRMR = .112, ΔSRMR = .033). The freely estimated regression coefficients of Negative Effect of Inclusion on teacher gender for Forms A, B, and C were –.503 (*SE* = .216, $p < .05$), .475 (*SE* = .241, $p < .05$), and –.442 (S*E* = .274, $p > .05$), respectively. These results indicated that for Form A, male teachers, compared with female teachers, were more likely to report negative effects associated with inclusion, whereas for Form B, female teachers were more

likely to report negative effects; the relationship between teacher gender and Negative Effect of Inclusion was not statistically significant ($p > .05$) within Form C. These results suggest a gender-by-form statistical interaction such that the relationship between teacher gender and Negative Effect of Inclusion differed depending on the form of the questionnaire.

For each of the other four predictor variables, constraining the regression coefficients of Negative Effect of Inclusion on each predictor to be equal across forms did not result in a significant decrease in the fit of the models (see Table 4). These results indicated that the assumption of invariant regression coefficients across forms (i.e., no statistical interaction) was tenable. For type of teacher (the three categories of general education, special education, and other types of teachers were recoded into two dummy variables, with other types of teachers as the reference group), the regression coefficient for the dummy variable of special education teachers was –.778 ($SE = .200$, $p < .01$), indicating that special education teachers reported fewer negative effects of inclusion compared with other types of teachers. The regression coefficient for the dummy variable of general education teachers (compared with other types of teachers) was not statistically significant (regression coefficient = –.154, $SE = .165$, $p > .05$).

Total years of teaching experience was not significantly related to Negative Effect of Inclusion (regression coefficient = .005, $SE = .006$, $p > .05$), whereas years of experience at one's current school was significantly related to the Negative Effect of Inclusion (regression coefficient = .025, $SE = .008$, $p < .01$). Teachers who were at their current schools for longer periods of time reported more negative effects of inclusion. Finally, the predictor variable of training in inclusion was found to be significantly related to the Negative Effect of Inclusion, with teachers who received training in inclusion reporting fewer negative effects of inclusion (regression coefficient = –.464, $SE = .155$, $p < .01$).

Overall, these results indicate a lack of statistically significant interactions between the form of the questionnaire and four of the five predictor variables: type of teacher, total years of teaching experience, years of experience at current school, and training in inclusion. Gender was the only predictor variable that exhibited a lack of invariance in its regression coefficients across forms.

In the previous analyses, invariance testing of the regression coefficients across forms was carried out for one predictor at a time. To determine if these results would change with simultaneous examination of the predictors, Negative Effect of Inclusion was regressed on teacher gender, type of teacher, total years of teaching experience, years of experience at one's current school, and training in inclusion and the invariance of the regression coefficients across forms was examined. The results considering the predictors simultaneously were consistent with those considering the predictors individually. Teacher gender was the only variable that exhibited a statistically significant interaction effect with the form of the questionnaire, and this effect was primarily due to differences evidenced on Form B.

Research Question 3: Mean Differences

Research Question 3 examined if altering the referent in the attitude statements measuring the Negative Effect of Inclusion had an impact on teachers' mean response levels. Multigroup confirmatory factor analysis with MACS was used to address this question. Item intercepts in these models represented the extent to which teachers endorsed the items, with larger numbers representing stronger levels of endorsement.

For the MACS models, the Negative Effect of Inclusion was scaled by fixing the factor pattern coefficient to 1.00 for the first item (the reference indicator) within each form. For identification purposes, the factor mean for Negative Effect of Inclusion was set to zero for Form A. Factor pattern coefficients and measurement error variances were constrained to be equal across forms on the basis of analyses conducted for Research Question 1.

Two MACS models were run. The first allowed the intercepts to be free across forms, whereas the second constrained the intercepts to be equal. For the model that allowed the intercepts to be free across forms, the fit was acceptable, $\chi^2(20, N = 288) = 38.598$, $p = .007$, CFI = .952, SRMR = .071. The latent mean for the Negative Effect of Inclusion for Form B was significantly greater than the latent mean for Form A (estimate/$SE = 7.91$, $p < .01$) and the latent mean for Form C (estimate/$SE = 3.39$, $p < .01$). The latent mean for Negative Effect of Inclusion for Form C was also significantly greater than the latent mean for Form A (estimate/$SE = 3.96$, $p < .01$). The analysis of the latent means for Negative Effect of Inclusion indicated that teachers' responses to inclusion became more negative as the referent in the questionnaire items changed from "students with mild disabilities" to "students with disabilities" to "students with severe disabilities."

When the intercepts were constrained to be equal across forms the change in χ^2 relative to the change of degrees of freedom was statistically significant ($\Delta\chi^2 = 22.584$, $\Delta df = 6$, $p < .001$), and the values of the CFI and SRMR suggested less than acceptable fit (CFI = .910, ΔCFI = .042, SRMR = .100, ΔSRMR = .029). These results indicated that the intercepts were significantly different across forms.

To further explore these differences, the observed means on the four items were compared across forms. The four items representing the Negative Effect of Inclusion were endorsed most strongly (i.e., had the largest means) when the referent was "students with severe disabilities" (Form B). The next largest means were when the referent was "students with disabilities" (Form C), followed by "students with mild disabilities" (Form A). Effect sizes, calculated as $(M_1 - M_2)$/pooled SD, provide additional information about the magnitude of these differences (see Table 5). For Item 1, teachers reported more negative attitudes related to inclusion when the referent was "students with severe disabilities" compared with "students with disabilities" (a moderate effect of 0.47) or "students with mild disabilities" (a large effect of 1.15). The smallest effects were observed for Item 2; however, the pattern was the same, with teachers more likely to endorse this item when the referent was "students with

Table 5
Effect Sizes for Negative Effect of Inclusion for Pairwise Comparison of Forms A (*mild*), B (*severe*), and C (*generic*)

Comparison: 1 Versus 2	Item 1: Burden	Item 2: Lower Performance	Item 3: Too Demanding	Item 4: Affect Self-Concept Negatively	Negative Effect of Inclusion
Severe (Form B) versus *generic* (Form C)	0.47	0.21	0.34	0.32	0.41
Generic (Form C) versus *mild* (Form A)	0.59	0.20	0.44	0.30	0.48
Severe (Form B) versus *mild* (Form A)	1.15	0.40	0.76	0.62	0.91

Note: Sample sizes for Forms A, B, and C were 102, 104, and 82, respectively. Effect size was computed as $(M_1 - M_2)$/pooled *SD*. A positive effect indicates that the first group viewed the effect of inclusion as more negative.

severe disabilities" compared with "students with disabilities" (a small effect of 0.21) or "students with mild disabilities" (a moderate effect of 0.40). For the overall scale of Negative Effect of Inclusion (formed by averaging the four items), the largest effect was between the *severe* and *mild* conditions (effect size = 0.91), followed by the difference between the *generic* and *mild* conditions (effect size = 0.48), and the *severe* and *generic* conditions (effect size = 0.41).

Discussion

This experimental study was conceived in the course of implementing a comprehensive program of research examining various dimensions of inclusive education. While developing an instrument to measure teachers' beliefs and perceptions about inclusion, we were forced to assess the pros and cons of developing the instrument completely around the generic label "students with disabilities" or around the differentiated labels of "students with mild disabilities" and "students with severe disabilities."

An instrument requiring teachers to respond to all items on the basis of an undifferentiated label had the appeal of simplicity and efficiency in terms of item construction and instrument length. However, such an instrument ran the risk of ignoring the fact that some teachers' beliefs and perceptions regarding inclusive education may be influenced by differentiations teachers make between placement options for students with mild and severe disabilities. After reviewing the literature on inclusion, it became clear that although much of the philosophical and policy debate on inclusion has centered on the differentiation between mild and severe disability, there has been little research on the methodological consequences of altering the referent in survey items.

As an initial study focusing on one attitudinal construct related to inclusion (Negative Effect of Inclusion), this study found that these wording changes had little effect on the measurement properties of the questionnaire. On the basis of the results of a series of confirmatory factor analyses, this study found that the factor pattern coefficients (i.e., loadings) and the measurement error variances for the items were not significantly altered by the wording changes. These results, which were obtained using a moderately large sample of 288 teachers, suggest that the fundamental structure of the measures was not altered by changing a few critical words (i.e., *mild* and *severe*) that have been at the center of the debate on inclusion.

Having found the assumption of measurement invariance to be tenable, this study examined if the wording changes affected the bivariate relations between a number of teacher demographic variables (e.g., type of teacher) and the Negative Effect of Inclusion. The results suggests that at least for this one construct and the five predictor variables examined in this study, the bivariate relationships are robust and similar across variations in question wording. In most cases, the same conclusions would be drawn no matter what referent was used in the attitude statement. This is an important finding for researchers who use correlational techniques (e.g., multiple linear regression) to build explanatory models that use attitudinal measures. Schuman and Presser (1977) characterized this type of result as an indication of "form-resistant correlations" (p. 153). The one exception to this finding occurred for teacher gender. Teacher gender had a negative relation to Negative Effect of Inclusion (i.e., female teachers reported fewer negative effects of inclusion) for Forms A and C, but for Form B, the regression coefficient was positive (.475, $p < .05$). This finding is difficult to explain, and additional research with more diverse samples is needed to test the generality of the finding related to teacher gender.

The largest effects from the alterations in question wording were observed in the means of the attitude statements. The results of mean comparisons are consistent with previous research suggesting that teachers' attitudes toward inclusion are affected by the severity of students' disabilities (Cook, 2001). Teachers' responses to inclusion became more negative as the referent in the survey items changed from "students with mild disabilities" to "students with disabilities" to "students with severe disabilities." The difference in means is also consistent with the many question-wording experiments done outside education that have found that even "seemingly innocuous word change can shift univariate item results noticeably" (Schuman & Presser, 1977, p. 152). In the present study, the item wording represented a fundamental change, and thus it was expected that the question wording would have a large effect on the univariate distribution of Negative Effect of Inclusion. The results of the present study suggest that school administrators and policy makers who are interested in determining whether teachers hold positive attitudes toward inclusion need to be aware that the answer to this question depends on how the items in a measurement instrument are framed and what referent is used. Furthermore, researchers and policy makers who are interested in determining if a consensus view toward inclusion is emerging need to

be aware that wording changes in a measurement instrument may also affect the variability in the construct being measured. In the present study, variability in the Negative Effect of Inclusion was limited in the *mild* condition, suggesting greater consensus, but increased significantly as one moved to the *generic* and *severe* conditions.

Although this study did not find major differences across forms in the factor structure underlying the Negative Effect of Inclusion, there is a need for researchers to continue to examine how variations in characteristics of measurement instruments may affect the quality of the data that are collected. Current advances in statistical methodology provide expanded ways to examine the effects of changing the wording in questions on data quality. These techniques include multigroup confirmatory factor analysis with MACS, as used in the present study, and differential item functioning analysis, which may be conducted using item response theory. By using a multitude of statistical approaches, researchers will be able to obtain more detailed information about how questionnaire design affects data quality.

It should be noted that in the present study, the items that were experimentally manipulated involved potential negative consequences of inclusion and were not contrasted with items measuring potential positive consequences of inclusion. Future research might investigate how question wording and the ordering of combinations of negatively and positively stated items might affect the quality of the responses to measurement instruments used to inform practice and policy.

References

Antonak, R. F., & Larrivee, B. (1995). Psychometric analysis and revision of the Opinions Relative to Mainstreaming scale. *Exceptional Children*, *62*, 139-149.

Antonak, R. A., & Livneh, H. (1988). *The measurement of attitudes toward people with disabilities: Methods, psychometrics and scales*. Springfield, IL: Charles C Thomas.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P. M. (1992). On the fit of models to covariances and methodology in the *Bulletin*. *Psychological Bulletin*, *112*, 400-404.

Cook, B. G. (2001). A comparison of teachers' attitudes toward their included students with mild and severe disabilities. *Journal of Special Education*, *34*, 203-213.

Hancock, G. R., & Freeman, M. J. (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement*, *61*, 741-758.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to unparameterized model misspecification. *Psychological Methods*, *3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Marfo, K., Harris, D. M., & Dedrick, R. F. (2002, April). *Empirical analysis of co-teaching and inclusive education*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320-341.

McLeskey, J., Waldron, N. L., & So, T. H. (2001). Perspectives of teachers toward inclusive school programs. *Teacher Education and Special Education*, *24*, 108-115.

Muthén, L. K., & Muthén, B. O. (1998-2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling*, *3*, 369-379.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly*, *5*, 91-92.

Schuman, H., & Presser, S. (1977). Question wording as an independent variable in survey analysis. *Sociological Methods & Research*, *6*, 151-170.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Thousand Oaks, CA: Sage.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Sideridis, G. D., & Chandler, J. P. (1995). Estimates of reliabilities for the teacher integration attitudes questionnaire. *Perceptual and Motor Skills*, *80*, 1214.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180.

Wilczenski, F. L. (1995). Development of a scale to measure attitudes toward inclusive education. *Educational and Psychological Measurement*, *55*, 291-299.