ASSESSING STUDENT PERCEPTIONS OF AFFECTIVE OUTCOMES OF SPECIAL EDUCATION PROGRAMS: INSTRUMENT DEVELOPMENT, VALIDATION, AND COMPARISONS TO REGULAR EDUCATION STUDENTS

ROBERT K. GABLE AND CHAUNCY N. RUCKER
University of Connecticut

EVERETT V. SMITH, JR.
University of Oklahoma

The development of the 45-item Likert-type response scale Student Survey for assessing the following five affective goals of *Connecticut's Common Core of Learning* is described: Academic Competence, Social Competence, Social Integration, Involvement in Educational Decision Making, and Consumer Satisfaction. Content validity literature-based and judgmental procedures are described; supportive evidence of construct validity is provided using a sample of 436 special education students (Grades 4, 6, and 8) from eight school districts who were eligible to take the statewide Connecticut Mastery Test. These data were analyzed using confirmatory factor analysis and item response theory techniques. Alpha reliabilities for the special education sample ranged from .70 to .91 for the five dimensions. Differences between special education students and a sample of 1,229 regular education students from the same districts are described.

Consistent with the federal office of Special Education Program's State Agency–Federal Evaluation Studies program and state mandates, the Connecticut State Department of Education (1987) has been developing strategies and instrumentation for evaluating special education students' progress

The authors extend their appreciation to Paula Colen of EASTCONN and to Tom Gillung and Ray Pecheone of the Connecticut State Department of Education for their support for the project from which this article is based. Requests for an expanded version of the text and data tables, or any other correspondence, should be sent to Robert Gable, University of Connecticut, 249 Glenbrook Road, U-4, Storrs, CT 06269.

Educational and Psychological Measurement, Vol. 57 No. 4, August 1997 685-697

toward academic and nonacademic components of *Connecticut's Common Core of Learning*. This study describes the development and validation of a nonacademic student attitude survey.

Method

Instrument Development: Content Validity

The first phase consisted of reviewing and editing a literature-based item bank in the area of self-concept (Marsh, Byrne, & Shavelson, 1988; Shavelson, Hubner, & Stanton, 1976). The second phase consisted of reviews by an eight-member special education advisory committee (item/dimension fit, readability, and age appropriateness). Then 10 special education teachers and 20 students (Grades 4 and 8) reviewed drafts of the survey with various response formats.

Survey Description

Demographic information included gender and grade level. A total of 50 items were included for rating on a 5-point Likert-type scale (1 = *disagree very much*, 5 = *agree very much*) containing numeric responses anchored at the top of each page by five variations of smiling/frowning faces. The five categories assessed by 10 items each were as follows:

- Academic Competence (things I can do): Students' perceptions of their ability to perform academic tasks;
- Social Competence (getting along with others): Students' perceptions of (a) their ability to perform skills necessary to maintain interactions and relationships with adults and peers in a school setting and (b) their ability to maintain behaviors necessary to function independently in a school setting;
- Social Integration (how people treat me): Students' perceptions of the extent to which they participate in and are accepted as members of the school community;
- Involvement in Educational Decision Making (my opinions): Students' perceptions of the nature of their participation in decisions that affect their educational programs or their access to educational resources; and
- Consumer Satisfaction (how I feel about school): Students' perceptions of the quality of educational services received.

Sample. The sample consisted of 436 special education students in Grades 4, 6, and 8 who were eligible to take the statewide Connecticut Mastery Test in eight school districts (urban, suburban, and rural). Individual and small group administration procedures were employed as determined by each district. A comparison group of 1,229 regular education students also completed the survey.

Data Analyses

Data were analyzed using descriptive statistics, confirmatory factor analysis (LISREL), reliability analyses, and Hotelling's multivariate analysis of variance procedures. Item response theory analyses also were run using Rasch latent trait procedures (Wright & Masters, 1982). Prior to running the confirmatory factor analysis, one item was deleted from each of the five categories to create a more favorable *n*:*p* ratio. These items were associated with item/category correlations < .40.

Results and Discussion

Confirmatory Factor Analysis

The confirmatory factor analysis procedure was run on the total special education sample consisting of 349 complete sets of data. Table 1 contains the maximum likelihood factor structure coefficients. For the Academic Competence dimension, the coefficients were only moderate, due largely to measurement errors above .70 (theta delta) associated with items 1, 2, 4, and 7. The same was true for items 12, 16, and 20 in the Social Competence dimension. For the Social Competence dimension, another feature was present: The modification indexes indicated that all of the items could empirically fit with all four of the remaining dimensions. A good example of this situation is item 17: "I can talk with my teachers." The literature and content experts felt that this item best fit the Social Competence dimension, but the item also could empirically fit the Involvement in Education Decision Making (self-advocacy) dimension or the Consumer Satisfaction dimension. Although this empirical support for possible movement of items to other dimensions was present, final decisions regarding item fit were made based on the literature underlying the survey. The factor structure coefficients for the Social Integration, Involvement in Educational Decision Making, and Consumer Satisfaction dimensions were, with few exceptions, moderately high.

Empirical support for the relationship among the dimensions was found in the factor correlation matrix in which indexes ranged from .36 to .93. The latent constructs assessed by the Social Competence and Social Integration item sets were highly correlated (r = .93), as were those assessed by the Involvement in Educational Decision Making and Consumer Satisfaction item sets (r = .84). Although these relationships make conceptual sense, the assessment of the individual dimensions was deemed appropriate given the *Common Core of Learning*'s special education goals and the dimension-level reliability data.

Several tests of model fit and various opinions regarding the standards for each test are found in the literature (Byrne, 1989; Jöreskog & Sörbom, 1989;

Mulaik et al., 1989). Our chi-square degrees of freedom ratio of 2.50 is judged by most to indicate a very acceptable fit. Byrne (1989) prefers the ratio to be less than 2:1, whereas Wheaton, Muthen, Alwin, and Summers (1977) prefer it to be less than 5:1. Given the amount of error variance in some of our items and the information from the modification indexes that some items could be assigned to other dimensions on an empirical basis, our chi-square degrees of freedom ratio of 2.5:1.0 was deemed acceptable. Our goodness of fit index (.74) and root mean square index (.71) could ideally be a little higher. The root mean square residual is ideally less than .05. Our index of .097 is another indication that some items empirically fit other dimensions. Finally, the coefficient of determination of .999 is a generalized estimate of the reliability of the entire set of 45 items. In reality, we are more interested in the reliabilities of the five subsets of items defining the dimensions. Although support for construct validity of the dimensions is an ongoing process, the confirmatory factor analysis results offer support for the five-dimensional structure of the survey instrument.

Item Response Theory Analysis

In a prior study, it was shown that factor analytic results may not adequately address how well a set of item statements span a targeted construct. (See Gable, Ludlow, & Wolf, 1990, and Gable & Wolf, 1993, for a discussion of latent trait analyses in relation to the construct validity information obtained from classical factor analytic techniques.) The additional construct validity information provided by item response theory was illustrated. In this study, the data set for the special education sample was analyzed under the assumption that it fit the Rasch rating scale model (Andrich, 1978; Wright & Masters, 1982). The computer program *SCALE*, developed by Masters, Wright, and Ludlow (1981), was employed in the analysis.

ADEQUACY OF CONSTRUCT DEFINITION

The Rasch latent trait analysis was carried out to further examine the construct validity issue regarding meaningful score interpretations. This item response theory technique is important because it addresses the adequacy with which the attitude continuum underlying each construct is assessed by the respective items. More complete score interpretations are possible when the items defining the construct are "differentiated" or spread across the respective attitude continuum.

Variable maps showing the positions of students and items for each dimension or construct were developed. Tables 2 and 3 illustrate the data obtained for the Consumer Satisfaction and Involvement in Educational

(Text continues on page 695)

Table 1
Confirmatory Factor Analysis: Maximum Likelihood Dimensions and Loadings (N = *349)*

Dimension/Item Stem	Academic Competence	Social Competence	Social Integration	Involvement in Educational Decision Making	Consumer Satisfaction
Academic Competence					
1. I do well in science	.49				
2. I am satisfied with my work in reading	.48				
3. I am satisfied with how I do on tests	.59				
4. I like my math work	.50				
5. I am satisfied with my work in social studies	.55				
7. I usually know the answers to my teachers' questions	.46				
8. I am satisfied with what I write	.52				
9. I feel good about my schoolwork	.75				
10. Following teacher directions is easy for me	.51				
Social Competence					
11. People like me		.68			
12. I know how to behave in school		.35			
13. Kids like me to be on their teams		.72			
14. Making friends is easy		.70			
15. I get along well with kids		.66			
16. Teachers say I behave well		.41			
17. I can talk with my teachers		.45			
19. Being a good friend is easy for me		.57			
20. I get along well with most teachers		.47			
Social Integration					
21. Kids treat me fairly			.71		
22. Kids like me just the way I am			.77		
23. Kids respect me			.73		
25. Other kids pick me to be on their teams			.70		

Table 1 *Continued*

Dimension/Item Stem	Academic Competence	Social Competence	Social Integration	Involvement in Educational Decision Making	Consumer Satisfaction
26. Kids in my class like to do schoolwork with me			.63		
27. I feel like I am part of the group			.74		
28. I get a lot of phone calls from friends			.49		
29. Kids like to do things with me			.77		
30. I am popular in school			.66		
Involvement in Educational Decision Making					
32. In my class, teachers ask what students think				.60	
33. I ask questions in class				.67	
34. Teachers listen to my opinion				.79	
35. I ask for help when I need it				.58	
36. When I decide to do something, I do it well				.61	
37. I like to make decisions				.49	
38. I contribute to discussions in my class				.69	
39. Teachers care what I think				.79	
40. I give my opinions in class				.75	
Consumer Satisfaction					
41. Teachers make learning interesting					.74
42. Teachers treat me fairly					.79
43. This is a good school					.72
44. Most teachers are interested in how I do in school					.70
45. Teachers help students					.75
47. I like the teachers here					.79
48. My schoolwork is interesting					.71
49. I like coming to school					.60
50. Teachers care about students					.76

Note. The minimum t value for the loadings was 6.50.

Table 2
Construct Map for the Consumer Satisfaction Dimension

Score	Person Position	Error	People ($N = 387$)	Item ($L = 9$)	Item Value	Fit
35 (20)	3.20	.98	XXXXXXXXXXXXXXXXXXXX			
34 (18)	2.51	.70	XXXXXXXXXXXXXXXXXX			
33 (20)	2.10	.57	XXXXXXXXXXXXXXXXXXXX			
32 (28)	1.81	.51	XXXXXXXXXXXXXXXXXXXXXXXXXXXX			
31 (23)	1.57	.46	XXXXXXXXXXXXXXXXXXXXXX			
30 (23)	1.37	.43	XXXXXXXXXXXXXXXXXXXXXX			
29 (19)	1.19	.40	XXXXXXXXXXXXXXXXXX			
28 (22)	1.04	.38	XXXXXXXXXXXXXXXXXXXXX			
27 (13)	0.90	.37	XXXXXXXXXXXXX			
26 (27)	0.77	.35	XXXXXXXXXXXXXXXXXXXXXXXXXX	49 (I like coming to school)	.78 (.1)	4
25 (16)	0.64	.34	XXXXXXXXXXXXXXXX			
24 (17)	0.53	.33	XXXXXXXXXXXXXXXX	48	.54 (.1)	–2
23 (13)	0.42	.33	XXXXXXXXXXXXX			
22 (12)	0.31	.32	XXXXXXXXXXXX	41	.26 (.1)	–1
21 (9)	0.21	.32	XXXXXXXXX	43	.23 (.1)	2
20 (9)	0.11	.31	XXXXXXXXX			
19 (12)	0.02	.31	XXXXXXXXXXXX	47	–.03 (.1)	–4
18 (11)	–0.08	.31	XXXXXXXXXXX	42	–.14 (.1)	–2
17 (8)	–0.17	.31	XXXXXXXX			

Table 2 *Continued*

Score	Person Position	Error	People ($N = 387$)	Item ($L = 9$)	Item Value	Fit
16 (10)	–0.27	.31	XXXXXXXXXX	44	–.35 (.1)	1
15 (7)	–0.36	.31	XXXXXXX	50	–.44 (.1)	1
14 (10)	–0.45	.31	XXXXXXXXXX			
12 (14)	–0.65	.32	XXXXXXXXXXXXXX			
11 (4)	–0.75	.32	XXXX			
10 (4)	–0.86	.33	XXXX	45 (Teachers help students)	–.86 (.1)	–1
9 (3)	–0.97	.34	XXX			
8 (4)	–1.09	.35	XXXX			
7 (2)	–1.21	.36	XX			
6 (1)	–1.35	.38	X			
5 (3)	–1.51	.41	XXX			
4 (2)	–1.70	.45	XX			
3 (1)	–1.94	.52	X			
2 (2)	–2.26	.62	XX			
1 (0)	–2.82	.90				

Note. Numbers in parentheses after scores are frequencies. Numbers in parentheses after item values are standard errors. The columns on the left side of the map contain raw score frequencies for the set of nine items, the number of students with each score for the set of nine items, the scale value or position on the continuum for each person's score, and the associated error statistic. The middle panel of the map contains an "X" for each person. The column labeled "items" contains important information regarding *each item's location on the attitude continuum underlying the respective constuct*. The indexes are *scale value* locations listed under the column labeled "item value." Following these indexes are the standard error and item fit statistics.

Table 3
Construct Map for the Involvement in Educational Decision Making Dimension

Score	Person Position	Error	People ($N = 390$)	Item ($L = 9$)	Item Value	Fit
35 (12)	3.09	.99	XXXXXXXXXXXX			
34 (17)	2.40	.71	XXXXXXXXXXXXXXXX			
33 (17)	1.99	.57	XXXXXXXXXXXXXXXX			
32 (23)	1.69	.50	XXXXXXXXXXXXXXXXXXXXXX			
31 (25)	1.46	.46	XXXXXXXXXXXXXXXXXXXXXXXX			
30 (21)	1.27	.42	XXXXXXXXXXXXXXXXXXXX			
29 (19)	1.10	.40	XXXXXXXXXXXXXXXXXX			
28 (29)	0.95	.38	XXXXXXXXXXXXXXXXXXXXXXXXXXXX			
27 (18)	0.81	.36	XXXXXXXXXXXXXXXXX			
26 (20)	0.69	.34	XXXXXXXXXXXXXXXXXX			
25 (22)	0.57	.33	XXXXXXXXXXXXXXXXXXXX			
24 (17)	0.46	.32	XXXXXXXXXXXXXXXX			
23 (19)	0.36	.31	XXXXXXXXXXXXXXXXX			
22 (17)	0.27	.31	XXXXXXXXXXXXXXXX	40	.26 (.1)	–4
21 (22)	0.17	.30	XXXXXXXXXXXXXXXXXXXX	38	.21 (.1)	–2
20 (14)	0.08	.30	XXXXXXXXXXXXXX	32, 33, 34, 39	.10 (.1)	4, 1, 2, 1
19 (10)	0.00	.29	XXXXXXXXXX			
18 (9)	–0.09	.29	XXXXXXXXX			
17 (9)	–0.17	.29	XXXXXXXXX	37	–.16 (.1)	3

Table 3 *Continued*

Score	Person Position	Error	People ($N = 390$)		Item ($L = 9$)	Item Value	Fit
15 (20)	−0.34	.29	XXXXXXXXXXXXXXXXXXXX	36		−.35 (.1)	−1
14 (3)	−0.42	.29	XXX				
13 (4)	−0.51	.29	XXXX	35		−.48 (.1)	2
12 (3)	−0.60	.30	XXX				
11 (4)	−0.69	.30	XXXX				
10 (3)	−0.78	.31	XXX				
9 (3)	−0.88	.32	XXX				
8 (1)	−0.99	.33	X				
7 (4)	−1.11	.35	XXXX				
6 (2)	−1.24	.37	XX				
5 (0)	−1.39	.40					
4 (1)	−1.57	.44	X				
3 (1)	−1.80	.51	X				
2 (0)	−2.12	.61					
1 (1)	−2.68	.91	X				

Note. Numbers in parentheses after scores are frequencies. Numbers in parentheses after item values are standard errors. The columns on the left side of the map contain raw score frequencies for the set of nine items, the number of students with each score for the set of nine items, the scale value or position on the continuum for each person's score, and the associated error statistic. The middle panel of the map contains an "X" for each person. The column labeled "items" contains important information regarding *each item's location on the attitude continuum underlying the respective constuct.* The indexes are *scale value* locations listed under the column labeled "item value." Following these indexes are the standard error and item fit statistics.

Decision Making dimensions. Primary statistics included are score frequencies, person positions in logits, and item location values in logits.

ITEM SCALE VALUES

Items listed with scale values or "difficulty indexes" toward the top of the map are those *most difficult to agree with*, whereas those toward the bottom are *least difficult to agree with*. A good example of this is found for the Consumer Satisfaction dimension illustrated in Table 2: Item 49, "I like coming to school," was the most difficult for students to agree with, whereas item 45, "Teachers help students," was the easiest to agree with. In addition to examining these item difficulty issues, the scale values allow examination of how well the items span the attitude continuum, which we refer to as how well the construct being assessed has been "differentiated."

The differentiation of an attitude construct is important because only with such differentiation can we contribute meaningful construct validity information to our score interpretations. Hence, differentiation leads to more complete score interpretations, which are the heart of the construct validity issue. The differentiation of the attitude construct can be illustrated by examining the spread of the item scale values across the attitude continuum. The spread of the items within each dimension was quite good for the types of items and students included in this study. The reliability of the item separation indexes for the five dimensions ranged from .89 to .98 (Wright & Masters, 1982). The items defining the Consumer Satisfaction dimension in Table 2 are particularly well spread across the attitude continuum, with items 45 and 49 bounding the end points of the continuum. Thus, it is meaningful and easier for us to describe a high- and low-scoring person on this dimension because we have a more comprehensive understanding of the construct through the content of the respective items.

Items appearing at the same locations on the attitude continuum do not add "new" information for differentiating the meaning of high- and low-scoring people. For example, the Involvement in Educational Decision Making dimension depicted in Table 3 illustrates that the following four items have the same location on the attitude continuum: item 32, "In my class, teachers ask what students think"; item 33, "I ask questions in class"; item 34, "Teachers listen to my opinions"; and item 39, "Teachers care what I think." Although the use of these four items enhances the alpha reliability of the data for the dimension, the use of only one or two items actually would provide an adequate representation of the scale value point on the attitude continuum.

ITEM FIT

Further preliminary study of individual items on the survey is accomplished through examination of the item fit statistics. When items fit the

Rasch measurement model, they are consistent in their ordering of students with high and low attitudes. That is, the set of items being analyzed for the respective dimension work together to form a single construct on which meaningful score interpretations for high- and low-scoring students may be given. This is possible because the item orders students on the continuum in relation to the order specified by the remaining eight items defining the dimension. On the other hand, when the information obtained from the student responses to an item is not consistent with that from the other eight items defining the same construct, the fit statistic t value exceeds +2, indicating "item misfit" and signaling the need for review of the item. For this type of attitude survey, the positive misfit t values are the primary concern.

Items with positive t values tend not to maintain their positions on the *easy to agree with* to *hard to agree with* continuum for both low- and high-scoring students. For example, as shown in Table 2, the high positive scale value for item 49 ("I like coming to school") indicates that it was difficult for many students to agree with this statement. Even though this is the case, the fit statistic of $t = 4$ indicates that the attitude information obtained from the item is not consistent with that obtained from the other eight items defining the construct. In other words, for some students the item is defining a concept that is not the same as the remaining eight items on the dimension. In this case, the examination of actual student responses in comparison to those predicted by the Rasch mathematical model indicated that the item misfit was created by several low-scoring students who tended to *agree very much* with the item. That is, a subset of students who tended to disagree with the other eight items defining the dimension tended to *agree very much* with the statement "I like coming to school." Although this is a subtle distinction, it makes sense that some students could disagree with statements such as "Teachers treat me fairly" and "My schoolwork is interesting" but still agree with the statement "I like coming to school." This situation is a good example of the usefulness of examining the item response theory misfit statistic (t). Three additional items associated with misfit t values > 2 were the following: item 4, "I like my math work" ($t = 6.40$); item 28, "I get a lot of phone calls from friends" ($t = 6.61$); and item 32, "In my class, teachers ask what students think" ($t = 4.28$). These items need to be examined further.

Reliability

The special education data alpha reliabilities for the three grade levels ranged from .70 (Academic Competence) to .91 (Social Integration). For the total special education sample data, the reliabilities were as follows: Academic Competence, .76; Social Integration, .89; Involvement in Educational

Decision Making, .86; and Consumer Satisfaction, .90. All of the reliabilities were considered to be very adequate for affective self-assessment data obtained from paper-and-pencil surveys with a special education population.

Comparison of Special Education and Regular Education Student Responses

Special education and regular education student responses to the Student Survey were compared using Hotelling's multivariate procedure. When a statistically significant difference was associated with a medium effect size > .30, follow-up univariate F tests using medium effect sizes of > .50 for a criterion were used to identify the main contributors to the overall difference. For all grade levels, regular education students had higher self-perceptions of their ability to perform academic tasks (i.e., Academic Competence) than did the special education students. At the Grade 6 level, regular education students had higher scores on perceived Social Competence.

References

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38*, 665-680.

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer-Verlag.

Connecticut State Department of Education. (1987). *Connecticut's common core of learning*. Hartford, CT: Author.

Gable, R. K., Ludlow, L. H., & Wolf, M. B. (1990). The use of classical and Rasch latent trait models to enhance the validity of affective measures. *Educational and Psychological Measurement, 50*, 869-878.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings*. Boston: Kluwer Academic.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS, Inc.

Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology, 80*, 366-380.

Masters, G. N., Wright, B. D., & Ludlow, L. H. (1981). *SCALE: A Rasch program for rating scale data*. Chicago: Mesa Press.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*, 407-441.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 84-136). San Francisco: Jossey-Bass.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.