

## DOCUMENT RESUME

ED 305 401

TM 013 029

AUTHOR Pilotte, William J.; Gable, Robert K.  
 TITLE Using Confirmatory Factor Analysis To Study the Impact of Mixed Item Stems on a Computer Anxiety Scale.  
 PUB DATE Feb 89  
 NOTE 28p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Savannah, GA, February 1989).  
 PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.  
 DESCRIPTORS \*Affective Measures; Computer Science Education; Error of Measurement; \*Factor Analysis; Factor Structure; Goodness of Fit; High Schools; High School Students; \*Item Analysis; \*Rating Scales; Test Bias; Test Items

IDENTIFIERS \*Computer Anxiety Scale; \*Confirmatory Factor Analysis; Factor Invariance; LISREL Computer Program; Parallel Test Forms

## ABSTRACT

Confirmatory factor analysis (LISREL VI) is the method best suited to the comparison of measurement models when those models are based on a priori assumptions. Traditionally, positive and negative item stems were mixed on affective scales to reduce response set bias since the item pairs were considered to be parallel. Recent studies indicate that positive and negative item stems may form separate factors, implying that they represent different constructs. In this study, the differences between positive and negative item stems were assessed using two forms of a computer anxiety scale to ascertain if the negation of an item produces a parallel item and to compare the factor structures and measurement errors to determine if factor invariance can be claimed. Three forms (Forms A, B, and C) of a computer anxiety scale were administered to a random sample of students (20 homerooms) at a small city high school. Reverse scoring was used for all items on Form B and for appropriate items on Form C. The results are consistent with those of other researchers, providing more evidence that the use of reverse scored items on an affective scale can alter students' responses to an item. One should view results with caution when the instrument includes mixed item stems, since the negation of an item tends to lead to an increase in the error variance related to the item. In general, positive and negative forms of this scale do not meet the criteria for factor invariance or for parallel tests. (TJH)

\*\*\*\*\*  
 \* Reproductions supplied by EDRS are the best that can be made \*  
 \* from the original document. \*  
 \*\*\*\*\*

ED305401

IMPACT OF MIXED STEMS

U.S. DEPARTMENT OF EDUCATION  
Office of Educational Research and Improvement  
EDUCATIONAL RESOURCES INFORMATION  
CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.  
 Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS  
MATERIAL HAS BEEN GRANTED BY

WILLIAM J. PILOTTE

TO THE EDUCATIONAL RESOURCES  
INFORMATION CENTER (ERIC)."

Using Confirmatory Factor Analysis to Study  
the Impact of Mixed Item Stems  
on a Computer Anxiety Scale  
William J. Pilotte  
and  
Robert K. Gable  
University of Connecticut

Paper presented at the annual meeting of the  
Eastern Educational Research Association,  
Savannah, Georgia, February, 1989.

M013029

## Abstract

Confirmatory factor analysis (LISREL VI) is the method best suited to the comparison of measurement models when those models are based on a priori assumptions (Hayduk, 1987; Jorskog & Sorbom, 1986). Traditionally, positive and negative item stems were mixed on affective scales to reduce response set bias since the item pairs were considered to be parallel (Gable, 1986; Fleishman & Benson, 1987; Nunnally, 1978). Recent studies indicate that positive and negative item stems may form separate factors, implying that they are representative of different constructs (Benson & Hocevar, 1985; Pilote & Gable, 1988; Schmitt & Stults, 1985; Wright & Masters, 1982). In this study, the differences between positive and negative item stems is studied using two forms of a computer anxiety scale to ascertain if the negation of an item produces a parallel item and to compare the factor structures and measurement errors to determine if factor invariance can be claimed. The results of this study are consistent with those of other researchers. One should view results with caution when the instrument includes mixed item stems, since the negation of an item tends to lead to an increase in the error variance associated with that item. In general, positive and negative forms of this scale do not meet the criteria for factor invariance or for parallel tests.

Using Confirmatory Factor Analysis to Study the Impact  
of Mixed Item Stems on a Computer Anxiety Scale

The purpose of this study was to investigate the changes in factor structure and error variance associated with items transformed to a negative stem. This section focuses on the measurement model, group comparisons, LISREL goodness of fit indices, and the use of positive and negative item stems.

Measurement Model. The measurement model used to explain the covariation in a set of observed variables is important since reliability depends on how closely the model can reproduce the covariance matrix (Bollen, 1982; Fleishman & Benson, 1987; Hayduk, 1987; Long, 1983). Confirmatory factor analysis allows for the testing of different measurement models based on a set of a priori assumptions concerning the number of restrictions placed on the scale items. The most restrictive model dictates that all items are equally accurate indicators and that the error associated with the individual items is not correlated. In the least restrictive model the factor loadings for the different scale items are free to vary, error variance is not constrained to be equal and the correlations between the disturbances for the observed variables are no longer forced to zero (Fleishman & Benson, 1987; Kenny, 1979). Previous research has proven that correlated measurement error will bias reliability estimates and that the reliability of an instrument can vary across subgroups (Fleishman & Benson, 1987). The congeneric

model, being the least restrictive, is probably the most realistic model.

Group Comparisons. In many instances a researcher would like to compare different groups with respect to a certain trait. These types of comparisons "share the implicit assumption that the measure of interest assesses a common latent construct across populations" (Newton, Kameoka, Hoelter, & Tanaka-Matsumi, 1984, p. 100). Since construct equivalence is a necessary condition for cross-group comparisons, factorial invariance must be established prior to score interpretation. Factorial invariance remains specific to the instrument and the population under study; consequently, it must be examined each time two or more groups are compared (Newton et al., 1984).

Linear structural relationships (LISREL) provides the flexibility to test different measurement models and to compare those measurement models for improved fit, factor invariance across groups, and equal error variance assumptions (Bollen, 1982; Fleishman & Benson, 1987; Hayduk, 1987; Joreskog & Sorbom, 1986; Newton et al., 1984). The LISREL measurement model specifies how hypothetical constructs are measured in terms of observed variables and can be used to describe the reliabilities and validities of those observed variables (Joreskog & Sorbom, 1986).

Goodness of Fit. The measurement model can be examined for goodness of fit using several different criteria. The  $\chi^2$  measure indicates if the model and the set of coefficient estimates are

consistent with the covariance matrix for the observed variables (Bollen, 1982; Hayduk, 1987; Hoelter, 1983). This measure has the disadvantage of being sample size dependent (Hayduk, 1987; Hocevar et al., 1987; Jorskog & Sorbom, 1986; Kroonenberg & Lewis, 1982; Long, 1983; Marsh, 1985, 1987). Since  $\chi^2$  is sample size dependent, other measures have been designed to address the concept of fit. Bentler and Bonnet (1982) have suggested a comparison between the model under consideration and the null model which assumes no common factors (Hayduk, 1987; Hocevar et al., 1987; Hoelter, 1983; Kroonenberg & Lewis, 1982; Long, 1983; Marsh, 1985, 1987; Newton et al., 1984). The Bentler-Bonett Index  $((\chi^2_{\nu} - \chi^2_0) / \chi^2_{\nu})$  scales the chi-square between 0 and 1.0, with 1.0 indicative of perfect fit (Bollen, 1982; Hayduk, 1987; Marsh, 1985). The Bentler-Bonett Index for acceptable measurement models should be greater than .90 (Bollen, 1982; Kenny, personal communication, 1988). A second comparison with the null model is the Tucker-Lewis Index  $((\frac{\chi^2_{\nu}}{df_{\nu}} - \frac{\chi^2_0}{df_0}) / (\frac{\chi^2_{\nu}}{df_{\nu}} - 1))$ . As with the Bentler-Bonett Index, larger values are indicative of better model fit. Hayduk (1987) advocates the use of competing models in assessing the goodness of fit rather than using the traditional null model. Carmines and McIver (1981) suggest using the ratio of  $\chi^2$  to the degrees of freedom, with values between 2 and 3 being reasonable and indicative of model fit (Hayduk, 1987; Hoelter, 1983; Marsh, 1985). It seems that the best method to assess the fit is to combine these criteria with the normalized residuals obtained from the LISREL program. The normalized residuals are "standard errors for the

estimated loadings, factor correlations and uniqueness" (Kroonenberg & Lewis, 1982, p. 69). If all of the normalized residuals are less than 2.0, then the model appears to adequately reproduce the given covariance matrix (Hayduk, 1987; Jorskog & Sorbom, 1986).

Positive and Negative Item Stems. In developing an affective scale, the researchers have traditionally been advised to include an equal number of positive and negative item stems in order to reduce response set bias (Benson & Hocevar, 1985; Schmitt & Stults, 1985; Wright & Masters, 1982; Nunnally, 1982). In following the traditional advice, the researcher must assume that the items are parallel or at least tau-equivalent (Fleishman & Benson, 1987). For this assumption to hold true, the positive and negative item stems need to define the same construct for the population under study (Benson & Hocevar, 1985). Previous studies have confirmed that positive and negative items are not unidimensional and that a two factor measurement model best represents the observed covariance matrix, with the negatively worded items defining the second factor (Benson & Hocevar, 1985; Pilotte & Gable, 1988; Schmitt & Stults, 1985). In part, this must be true since studies have shown that wording changes can make significant differences in the factor structure and in the item validities (Benson & Hocevar, 1985; Bentler, Jackson & Messick, 1971; Schmitt & Stults, 1985).

High school students react differently to positive and negative item stems and their responses are affected by the emotionality of the words (Simpson et al., 1976). A controlled experiment involving

upper division undergraduate students indicated that using reverse scored items resulted in more student inaccuracies that were both practically and significantly different (Schriesheim & Hill, 1981). Schriesheim and Hill also concluded that the negatively worded items were less valid (i.e., result in less accurate responses which impairs the validity of the results). The use of mixed item stems to balance response set bias appears to be ineffective; however, Masters and Wright (1982) advocate the use of "For and Against statements" to "expose persons with unusual response tendencies" (p. 135).

In summary, in order to study the differences that exist between positive items and their negative transformations a measurement model must be established. The most realistic being the congeneric or least restrictive model. These differences can be detected using the LISREL VI multiple group procedure. The model comparisons must be made after assessing multiple goodness of fit indicies.

Purpose. This study employed confirmatory factor analysis techniques to assess the issue of model fit and to compare different plausible models based on the model's ability to reproduce the original covariance matrix. First, it will address the issues of fit and improvement of fit. Second, this study will illustrate that transforming an affective item stem from positive to negative wording will alter the item, resulting in the emergence of a "negative" factor for the high school population.



METHODSample

A random sample of students attending a small city high school was obtained using homeroom assignments. This school population is representative of other small city schools within the state. Although the sample size is small, it is within acceptable bounds for a confirmatory LISREL analysis.

Instrumentation

Three parallel forms of an instrument to measure computer anxiety were developed to study the impact of item phrasing on the validity of a Likert-type affective scale used in a high school setting. The three forms differed only in phrasing. The first scale was composed of nine items that indicate computer anxiety as defined below.

An unpleasant, emotional state marked by worry, apprehension and attention associated with thinking about, using, or being exposed to a computer.

The scale resulted from items generated by the first author and rated for content validity by seven experts in the field of computer education and high school students. The experts were sent a list of statements, a short review of the literature, and directions for rating the items. Some statements had to be eliminated based on the raters' comments. A specific example from the form "Only smart people can master a computer" elicited the additional "...and I am smart so" or "I am not smart so" which is more indicative of the students' general academic self confidence than of computer anxiety.

The construct validity of this form was assessed by a factor analysis with varimax and oblimin rotations, with the number of factors extracted determined by Kaiser's criterion. The original instrument contained 10 items on a 5-point Likert scale, with 1 assigned to strongly disagree, 5 to strongly agree and 3 to neutral. The factor structure from this exploratory analysis was indicative of a one factor structure when only 9 items were used.

The second form was devised by negating each item from the original form to provide a parallel instrument. Five of these statements were negated using the word not, while the remaining four statements were negated by changing the target word to one opposite in meaning. Traditionally, these items should reflect computer anxiety when reverse scored. A third form, C, consisted of 5 items from Form A and 4 items from Form B. An example of each type of item is presented below:

(computer anxiety) I feel threatened by computers.

(nonanxious) I do not feel threatened by computers.

(computer anxiety) I feel stupid around computers.

(nonanxious) I feel intelligent around computers.

#### Analyses

The three forms of the instrument were combined into packages and distributed by the building principal to 20 different homerooms, equally divided among grades 9 through 12. Each student responded anonymously to one of the three forms. Reverse scoring was employed for all items on Form B and for the

appropriate items on Form C, such that a response of 5 was indicative of computer anxiety.

A measurement model was developed for Form A and Form B. A traditional null model with no common factors and a competing model were also developed for each form. The Bentler-Bonett and Tucker-Lewis Indices were calculated and used in conjunction with the chi-square to degree of freedom ratio, chi-square statistic, and normalized residuals in order to assess the goodness of fit. The two models were analyzed for factor invariance using the LISREL VI multiple groups procedure. The issue of model fit was addressed using the chi-square statistic and normalized residuals. The chi-square differencing technique was used to test for a significant increase in fit between nested models.

### Results

#### Reliability

The three forms were analyzed for the degree of internal consistency using Chronbach's Alpha and the results are presented in Table 1.

---

Insert Table 1 about here

---

#### Factor Structure

A confirmatory factor analysis (LISREL VI) used to determine if the positive and negative item stems were measuring the same construct indicated a two factor model was preferable (Pilotte &

Gable, 1988). The factor loadings and factor correlations are included in Table 2.

---

Insert Table 2 about here

---

Prior to testing for factor invariance, measurement models were independently developed for Form A and Form B. The items were worded such that agreement was indicative of computer anxiety for Form A and lack of anxiety for Form B. The consistency of item stems used on each form suggested that those items should define a single factor. Since an a priori factor structure was to be tested, confirmatory factor analysis (LISREL VI) was used to test and refine the measurement models. The covariance matrix for the students' responses were input in all cases and a one factor solution specified. Psi was set equal to 1 and the program was allowed to estimate the factor loadings and the disturbances associated with each item. The final measurement model was refined to allow for correlated measurement error between pairs of statements, which theoretically share some common error variance. The inclusion of the correlated measurement error is necessary when the omission of a common cause contributes to the measurement error (Hayduk, 1987). The factor loadings and correlations for Form A are found in Table 3.

---

Insert Table 3 about here

---

### Model Fit

In testing the measurement model the null hypothesis is "There is no significant difference between the input data and the model" (Benson, 1987; Jorskog & Sorbum, 1986; Marsh, 1985). One indication of adequate model fit is a nonsignificant chi-square. The final measurement model for Form A had a  $X^2$  of 19.61, (df = 23,  $p = .665$ ), which seems to indicate that the model reproduces the original covariance matrix. The normalized residuals for the measurement model were inspected for values greater than 2.0 in order to ascertain if any entry in the original covariance matrix could not be accounted for by the given model. The largest normalized residual was .63, which supports the decision to accept this measurement model. The chi-square to degrees of freedom ratio, Bentler-Bonett Index, and Tucker-Lewis Index also indicate that the given measurement model has adequate fit. The results of these tests can be found in table 4.

---

Insert Table 4 about here

---

The measurement model for Form A was also compared to a competing model. A previous study indicated that the two factor model, positive item stems on one factor and negative item stems

on the second factor, fit the data better than the single factor model (Pilotte & Gable, 1988). Consequently, the competing model was defined by assigning the same items to the same two factors. Failing to accept this two factor model also supports the previous conclusion that the item stem defined the factor (for a more complete discussion see Pilotte & Gable, 1988). The accepted method for comparing nested models is chi-square differencing (Benson, 1987; Bollen, 1982; Hayduk, 1987; Marsh, 1985). This method states that  $\chi^2_{1-2} = \chi^2_1 - \chi^2_2$  with degrees of freedom equal to  $df_1 - df_2$  with model 1 being the most restrictive model.

This competing model was not capable of reproducing the original covariance matrix as evidenced by the 20 normalized residuals that were greater than 2.0. In this case the chi-square differencing test yielded a chi-square of 235.83, ( $df = 4$ ), which is highly significant. Therefore, the one factor model being tested better explains the data.

The same procedures were applied to Form B to obtain and test the measurement model. The factor loadings and error variances are found in Table 5.

---

Insert Table 5 about here

---

The measurement model for Form B has a chi-square statistic of 26.60, ( $df = 21$ ), which is indicative of fit between the model and the covariance matrix. All of the normalized residuals were less

than 2.0, which indicates that the model can adequately reproduce all the cells within the original covariance matrix. The other tests for goodness of fit, see Table 6, support the hypothesis that this measurement model is consistent with the original data.

---

Insert Table 6 about here

---

The chi-square difference test for Form B, with respect to a two factor model, yielded a value of 139.83, (df = 6), which is highly significant. This indicates that the model being tested provides better fit. The LISREL program estimate of the correlation between the two factors was .91. This high a correlation also supports the conclusion to use a one factor model.

#### Factor Invariance

The measurement models for Form A and Form B were tested for factor invariance. In constructing Form B, each item from Form A was negated in an attempt to establish parallel items. Factor invariance is a necessary condition for parallel items (Fleishman & Benson, 1987). In testing for invariance both forms are simultaneously fit to the same model, constraining some parameters to be equal (Hayduk, 1987; Jorskog & Sorbom, 1986; Marsh, 1985). The comparison was made between the model that allowed factor loadings to vary and the model with factor loadings constrained to be equal. This type of comparison is in keeping with the

sequential rules established by Jorskog and Sorbom (Jorskog & Sorbom, 1986; Kenny, personal communication, 1988; Marsh, 1985, 1987).

The least restrictive model allowed the factor loadings to vary over groups. This model has a chi-square of 46.26, (df = 44), which is indicative of adequate fit. The normalized residuals were all less than 2.0 indicating that this model reproduces the original covariance matrix. A competing two factor model yielded a chi-square of 421.85, (df = 54), which is indicative of poor fit. The chi-square difference of 375.59, (df = 10), indicates the one factor model gives a better fit.

The model that forced respective items on Form A and Form B to have equal factor loadings was then tested. This model resulted in a chi-square of 87.21, (df = 53,  $p = .002$ ), indicating lack of fit. Analysis of the normalized residuals indicated that this model failed to reproduce 11 entries of the original covariance matrix for Form B. Comparing this model with the previous model resulted in a chi-square difference of 40.95, (df = 9), suggesting that the less restrictive model is best. The goodness of fit statistics summarized in Table 7 all indicate that the less restrictive model shows a better fit with the original data.

---

Insert Table 7 about here

---



### Conclusions

This study provides more evidence that the use of reverse scored items on an affective scale can alter the students' response to an item. The two scales were constructed to form parallel items. A necessary condition for parallel items is factor invariance which was tested by the multiple groups procedure of LISREL VI. The analysis clearly indicated that respective items on the two scales were best represented by different factor loadings. This result is consistent with earlier research (Benson & Hocevar, 1987; Pilotte & Gable, 1988).

The LISREL model provides a measure of generalized reliability for the model, the total coefficient of determination, which decreased from .99 for Form A to .87 for Form B. This seems to support Schriesheim and Hill's (1981) study that indicated that negatively phrased items tend to be less valid partially because of increased student inaccuracies (Benson & Hocevar, 1985; Schriesheim & Hill, 1981). Further analysis of these data is necessary to study the effect that the positive/negative transformation may have had on item reliabilities.

The measurement models for Form A and Form B include error variance estimates. The error variances for the negatively phrased items, Form B, appear to be higher than for the respective item on Form A. This result is also consistent with previous research (Benson, 1987; Benson & Hocevar, 1985; Schriesheim & Hill, 1981). Since the results do not support the hypothesis of

factor invariance, the inclusion of mixed item stems on an affective instrument should be viewed with caution. A more complete study of the high school population needs to be undertaken to ascertain the extent to which reverse scored items affect the item and instrument reliabilities and the factor structures of the instrument itself.

Table 1

Internal Consistency of the Computer Anxiety Scales

Form	Item Classification	Alpha
A	Computer Anxiety	.95 (N=94)
B	Reverse Score	.87 (N=90)
C	Mixed Stems	.73 (N=87)

Note. From Pilotte, W. & Gable, R. (1988, November). The impact of positive and negative item stems on the validity of a computer anxiety scale. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, N.Y.

Table 2

Factor Loadings and Factor Correlations for Form C

Factor 1	Factor 2
.94 (10.49)	.00
.89 (11.31)	.00
.90 (8.69)	.00
.00	.69 (4.80)
.00	.94 (5.36)
.00	.35 (2.51)
.39 (3.42)	.00
.00	.40 (2.70)
.29 (2.87)	.00

note: All loadings significant using t-values from LISREL VI program. T-values given in parentheses next to factor loadings.

Psi Factor 1, Factor 2

.24 (1.97)

Note. From Pilotte, W. & Gable, R. (1988, November). The impact of positive and negative item stems on the validity of a computer anxiety scale. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, N.Y.

Table 3  
Factor Loadings and Item Error Variance Form A

Loading	Error Variance
.94 (11.3)	.19 (6.0)
.87 (11.8)	.12 (6.2)
.97 (10.4)	.33 (6.6)
.94 (11.6)	.16 (5.3)
1.0 (11.5)	.22 (5.4)
1.0 (11.8)	.18 (5.7)
.95 (11.2)	.22 (6.5)
.75 (6.6)	.90 (6.8)
.57 (6.6)	.53 (6.8)

note: All loadings and error variances are significant using t-values from LISREL VI program. T-values given in parentheses next to each value.

$\chi^2 = 19.61$ ;  $p = .665$ ; total coefficient of determination = .990

Table 4

Goodness of Fit Indices: Form A

---

Index	Value
$\chi^2 / df$	.85
Bentler-Bonett	.98
Tucker-Lewis	1.0

---

Table 5

Factor Loadings and Error Variances: Form B

Loading	Error Variance
.88 (6.3)	1.17 (6.1)
.98 (9.4)	.41 (4.9)
.99 (9.4)	.42 (4.8)
.94 (7.2)	.91 (5.9)
.91 (7.3)	.84 (5.9)
.51 (5.8)	.67 (6.3)
.26 (1.9)*	1.37 (6.7)
.72 (5.9)	.96 (6.3)
.85 (8.5)	.45 (5.8)

Note: All loadings significant using t-values from LISREL VI program unless marked by \*. T-values given in parentheses next to values.

Table 6

Goodness of Fit Indices: Form B

Index	Value
$\chi^2 / df$	1.26
Bentler-Bonett	.97
Tucker-Lewis	.99



Table 7

Goodness of Fit Indices: Multiple Groups

---

Invariant	
X <sup>2</sup> /df	1.6
Bentler-Bonett	.94
Tucker-Lewis	.98
Unconstrained	
X <sup>2</sup> /df	1.05
Bentler-Bonett	.97
Tucker-Lewis	1.0

---

## References

- Benson, J. (1987). Detecting item bias in affective scales. Educational and Psychological Measurement, 47, 55-67.
- Benson, J. & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. Journal of Educational Measurement, 22, 231-240.
- Bollen, K. (1982). A confirmatory factor analysis of air quality. Evaluation Review, 6, 521-535.
- Fleishman J. & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliabilities. Educational and Psychological Measurement, 47, 925-939.
- Gable, R. (1986). Instrument development in the affective domain. Boston: Kluwer Nijhoff.
- Hayduk, L. (1987). Structural equation modeling with lisrel: essentials and advances. Baltimore: John Hopkins University Press.
- Hocevar, D., Khattab, A., & Michael, W. (1987). Significance testing and efficiency in lisrel measurement models. Educational and Psychological Measurement, 47, 45-49.
- Hoelter, J. (1983). The analysis of covariance structures. Sociological Methods and Research, 11, 325-344.
- Jorskog K. & Sorbom, D. (1986). LISREL VI analysis of linear structural relationships by the method of maximum likelihood. Indiana: Scientific Software, Inc.

- Kenny, D. A. (1979). Correlation and causality. New York: John Wiley.
- Kroonenberg, P. & Lewis, C. (1982). Methodological issues in the search for a factor model: Exploration through confirmation. Journal of Educational Statistics, 7, 69-89.
- Long, J. S. (1983). Confirmatory factor analysis: a preface to LISREL. Beverly Hills, CA: Sage.
- Marsh, H. (1985). The structure of masculinity/femininity: an application of confirmatory factor analysis to higher-order factor structures and factorial invariance. Multivariate Behavioral Research, 20, 427-449.
- Marsh, H. (1987). The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. Journal of Educational Measurement, 24, 17-40.
- Newton, R., Kameoka, V., Hoelter, J., Tanaka-Matsumi, J. (1984). Maximum likelihood estimation of factor structures of anxiety measures: a multiple group comparison. Educational and Psychological Measurement, 44, 179-193.
- Nunnally, J. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.
- Pilotte, W. & Gable, R. (1988, November). The impact of positive and negative item stems on the validity of a computer anxiety scale. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, N.Y.

- Schmitt, N. & Stults, D. (1985). Factors defined by negatively keyed items: the result of carless respondents? Applied Psychological measurement, 4, 367-373.
- Schriesheim, C. & Hill, K. (1981). Controlling acquiescence response bias by item reversals: the effects on question validity. Educational and Psychological Measurement, 41, 1101-1114.
- Simpson, R., Rentz, R., & Shrum, J. (1976). Influence of instrument characteristics on student responses in attitude assessment. Journal of Research in Science Teaching, 13, 275-281.
- Wright, B. & Masters, M. (1982). Rating scale analysis. Chicago: Mesa Press.