DOCUMENT RESUME

ED 294 904 TM 011 564

AUTHOR Byrne, Barbara M.
TITLE Testing for Equivalent Factorial Validity across Academic Track: A Confirmatory Factor Analysis of the Self Description Questionnaire III.
PUB DATE Apr 88
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Factor Analysis; Grade 11; Grade 12; High Schools; High School Students; Self Concept; *Self Evaluation (Individuals); Statistical Analysis; *Test Validity; *Track System (Education)
IDENTIFIERS *LISREL Computer Program; *Self Description Questionnaire III


ABSTRACT

Confirmatory Factor Analysis (CFA; LISREL) was used to: (1) validate the Self Description Questionnaire III (SDQ III) subscales measuring general, school, English, and mathematics self-concepts for a sample of 898 (285 low track and 613 high track) students in grades 11 and 12, and (2) test the equivalency of the factor structure across academic tracks. The results confirm a four-factor structure and reveal all factor loadings and covariances to be invariant across groups; 11 of 21 uniquenesses were non-invariant. Although, in a strict statistical sense, the differentially reliable items bore implications of bias in favor of the high track, these discrepancies, on the basis of absolute values, were judged to be of little practical significance. Overall, in light of the stringency of LISREL CFA procedures in general, and those used in this study in particular, the SDQ III demonstrated exceptionally superior psychometric properties. (Author)

# Testing for Equivalent Factorial Validity Across Academic Track: A Confirmatory Factor Analysis of the Self Description Questionnaire III

Barbara M. Byrne

University of Ottawa



Paper presented at the American Educational Research Association Annual Meeting, New Orleans, 1988

## Abstract

Confirmatory factor analysis (CFA; LISREL) was used to (a) validate the Self Description Questionnaire III (SDQ III) subscale measuring general, school, English, and mathematics self-concepts for a sample of 898 (285 low track, 613 high track) grades 11 and 12 students, and (b) test the equivalency of the factor structure across academic track. The results confirmed a 4-factor structure and revealed all factor loadings and covariances to be invariant across groups; 11 of 21 uniquenesses were noninvariant. Although, in a strict statistical sense, the differentially reliable items bore implications of bias in favor of the high track, these discrepancies, on the basis of absolute values, were judged to be of little practical significance. Overall, in light of the stringency of LISREL CFA procedures in general, and those used in this study in particular, the SDQ III demonstrated exceptionally superior psychometric properties.

# Testing for Equivalent Factorial Validity Across Academic Track: A Confirmatory Factor Analysis of the Self Description Questionnaire III

The Self Description Questionnaire III (SDQ III; Marsh & O'Neill, 1984) is designed to measure multidimensional academic and nonacademic SCs for late adolescents, and is theoretically linked to the hierarchical model of self-concept (SC; Shavelson, Hubner, & Stanton, 1976). Development of the SDQ III was based on construct validation research bearing on two earlier versions of the instrument --- the SDQ for preadolescents, and the SDQ II for early adolescents (see Marsh, Barnes, & Hocevar, 1985 for a research summary).

The SDQ III contains 136 items measuring 13 SC facets -- one general SC, three academic SCs (English, mathematics, general school), and nine nonacademic SCs (physical ability, physical appearance, social (same sex), social (opposite sex), parent relations, emotional stability, problem solving/creative thinking, religion/spirituality, and honesty/reliability). Only the general and academic SC subscales are relevant to the present investigation.

Exploratory and confirmatory factor analyses of the SDQ III have yielded well-defined general, and academic SC facets that were relatively distinct from each other (Byrne & Shavelson, 1986; Marsh & O'Neill, 1984; Marsh et al., 1985; Marsh, Richards, & Barnes, in press). Internal consistency reliability coefficients ranging from 0.79 to 0.95 (mean $\alpha$ = 0.90; Byrne & Shavelson, 1986; Marsh & O'Neill, 1984; Marsh et al., 1985), and test-retest reliability coefficients ranging from 0.66 to 0.94 (mean $r$ = 0.86; Marsh et al., in press) have been reported. Tests of factorial invariance across gender have demonstrated equivalence for all but two item-pairs of the General Self subscale (Byrne, in press a). Finally, results from multitrait-multimethod analyses have shown strong evidence of convergent and discriminant validities (for a review, see Byrne, in press a).

Taken together, these findings provide strong support for the SDQ III as a potentially reliable and valid measure of adolescent SC. However, the assumption of equivalent factorial validity across levels of intellectual ability has not been directly tested. Indeed, previous research has shown the emergence of differential factor structures based on responses by children of different ability levels (see e.g., Byrne & Schneider, 1988; Silon & Harter, 1985). Thus, the validity of findings from research and program evaluations bearing on ability-group comparisons is dependent upon the factorial invariance of the measuring instrument. The purposes of the present study were twofold: (a) to test for the factorial validity of the SDQ III subscales measuring general, school, English, and mathematics SCs for low- and high-track high school students, and (b) to test for the factorial invariance of these subscales across academic track.

## Method

### Sample and Procedure

The present data were derived from a larger study that was designed to validate the structure of adolescent SC (Byrne & Shavelson, 1986), Following listwise deletion of missing data, the present sample comprised 285 low-track, and 613 high-track students from two high schools in Ottawa, Canada. The data were slightly negatively skewed with values ranging from -1.60 to 0.61 ($\bar{X}$ = -0.43) for the low track, and from -2.35 to 0.07 ($\bar{X}$ = -0.79) for the high track; kurtosis ranged from -.82 to 2.76 ($\bar{X}$ = 0.19) for the low track, and from -1.06 to 5.23 ($\bar{X}$ = 0.50) for the high track. Given mean skewness ranges between -1.00 and +1.00, however, little distortion to the parameter estimates was expected (see Muthén & Kaplan, 1985). (For a more extensive description of academic tracks, sampling procedures, and instrument administration, see Byrne, in press b).

### Instrumentation

The SDQ III is structured on an 8-point likert-type scale with responses ranging from "1-Definitely False" to "8-Definitely True". The General-Self subscale contains twelve items and was used to measure general SC. The Academic SC, Verbal SC, and Mathematics SC subscales each contain ten items and were used to measure general school, English, and mathematics SCs, respectively.

Analysis of the Data

All responses to negatively worded items were reflected so that the highest response code indicated a positive rating of SC. Using confirmatory factor analytic procedures (LISREL VI; Joreskog & Sorbom, 1985), the data were analyzed in two stages. First, the factorial validity of the SDQ III was tested separately for low- and high-track students. Second, the factorial invariance of the SDQ III was tested across academic track.

Consistent with Marsh and associates' research on the SDQ III, all analyses were conducted on item responses formed in pairs (for a description and rationale for this procedure, see Marsh & O'Neill, 1984; Marsh et al., in press). Assessments of model fit were based on the following criteria: (a) the chi square ( $X^2$ ) likelihood ratio test, (b) the $\chi^2$/degrees of freedom ratio, (c) Bentler and Bonett's normed index of fit (BBI)[1], and (d) T-values, normalized residuals and modification indices, all provided by the LISREL program.

Although EFA is widely used in construct validation research, it is limited in its ability to: (a) yield unique factorial solutions, (b) define a testable model, (c) assess the extent to which an hypothesized model fits the data, and suggest alternative parameterization for model improvement and, (d) adequately test factorial invariance across groups (Fornell, 1983; Long, 1983; Marsh & Hocevar, 1985). CFA, on the other hand, yields this information and is therefore a more powerful test of factorial validity. The CFA model in the present study hypothesized a priori that: (a) responses to the SDQ III could be explained by four factors. (b) each item-pair would have a non-zero loading on the SC factor it was designed to measure, and zero loadings on all other factors, (c) the four factors would be correlated and, (d) the error/uniqueness terms for the item-pair variables would be uncorrelated. Finally, since the same items were administered to both low- and high-track students, the measurements were hypothesized to be invariant. This hypothesis was tested directly by constraining the factor loadings and uniquenesses to be equivalent across track

## Results and Discussion

### Confirmatory Factor Analyses

Although, for both tracks, the hypothesized 4-factor model represented a statistically unacceptable fit to the data (low track, $\chi^2_{183}$ = 425.18; high track, $\chi^2_{183}$ = 805.45), the normed fit index for the high track represented a psychometrically reasonable fit to the data (BBI = .91) indicating that over 90% of the data covariation was accounted for; the index of fit for the low track was less adequate (BBI = .86). The factor loading and uniqueness estimates for these models are presented in Tables 1 and 2.

---

Insert Tables 1 and 2 about here

---

One factor considered to be instrumental in undermining the model fit was the presence of corrlelated uniquenesses.[2] Indeed, previous research has demonstrated that LISREL models involving psychological constructs in general (see e.g., Joreskog, 1982; Newcomb, Huba, & Bentler, 1986; Huba, Wingard, & Bentler, 1981), and the SC construct in particular (see e.g., Byrne & Shavelson, 1986, 1987), often requires the researcher to specify correlated uniquenesses in order to obtain a well-fitting model; such parameter specifications, of course, being theoretically and empirically jusified (see Fornell, 1983; Gerbing & Anderson, 1984). Correlated uniquenesses frequently result from nonrandom error introduced by a particular measurement method; one example is that of method effects due to the item format associated with subscales of the same measuring instrument.

To investigate the misfit in the model, then, a sensitivity analysis was conducted (see Byrne, Shavelson, & Muthén, 1987; Tanaka & Huba, 1984). As such, model fitting for each track was continued beyond the initially fitted models. Several additional modifications that included both correlated uniquenesses and secondary loadings (item-pair loadings on non-target factors), resulted in a statistically better fitting model for both the low track ($\chi^2_{162}$ = 192.48, p = .06; BBI = .94) and the high track ($\chi^2_{149}$ = 171.52, p = .10; BBI = .98). Given the probability of method effects as noted earlier, along with the known moderate correlations among the four SC factors under study, these parameters were not unexpected.

Several considerations, however, bore on the decision to reject these final models in favor of the more parsimonious intial models. First, the uniqueness covariance estimates, while statistically significant, were relatively minor. ranging from -.14 to .15 ($\bar{X}$ = .05) for the low track, and from -.06 to .27 ($\bar{X}$ = .04) for the high track. Second, the estimated secondary factor loadings, while statistically significant, were also relatively minor, ranging from -.30 to .44 ($\bar{X}$ = .04) for the low track, and from -.24 to .29 ($\bar{X}$ = .03) for the high track. Third, the estimated factor loading and factor variance-covariance estimates in the final model correlated .93 and .99 respectively, for the low track, and .94 and .97 respectively, for the high track, with those in the initially hypothesized model (see Byrne et al., 1987; Newcomb et al., 1986; Tanaka & Huba, 1984); these results substantiated the stability of the initial models. Fourth, although each of the model respecifications resulted in a statistically significant improvement in model fit, these increments, based on the normed index of fit, were considered of little practical significance (see also, Huba et al., 1981). Fifth, the sensitivity of the likelihood ratio test to trivial departures of the observed from an hypothesized model, with large samples, is now widely known (see Bentler & Bonett, 1980; Huba et al., 1981; Marsh & Hocevar, 1985). Finally, given the exploratory nature of these supplementary analyses and thus, the risk of capitalization on chance factors (see Long, 1983), the estimates derived from these final models were considered dubious. For these reasons, then, the initial model for each track was used as the baseline model in tests of invariance.

Factorial Invariance Across Academic Track

The next step in the analyses involved estimating parameters simultaneously for low and high tracks, in order to test for equivalencies of item-pair measurements, and factor covariances. These results are summarized in Table 3.

---

Insert Table 3 about here

---

Tests of invariance involved specifying a model in which certain factor loading parameters were constrained to be equal across track and then comparing that model with a less restrictive model in which these parameters were free to take on any value. Since the difference in $\chi^2$ ($\Delta\chi^2$) is distributed as $\chi^2$, with degrees of freedom equal to the corresponding difference in degrees of freedom, it provides a basis for determining the tenability of the hypothesized equality constraints; a significant $\Delta\chi^2$ indicating noninvariance. For example, Model 2 in which all factor loadings were specified as equal across track was compared against Model 1 in which only the number of factors was held invariant; the pattern of factor loadings was unconstrained. The difference in $\chi^2_{17}$ was 23.91, which was not significant. This finding indicated that the pattern of factor loadings was equivalent across academic track; items were thus measuring designated SC facets in the same way for each track.

Some researchers have argued that claims of factorial invariance should provide additional evidence of equivalent uniquenesses across groups (see e.g., Benson, 1987). Non-equivalent uniquenesses, then, would suggest that an instrument is more reliable for one group than it is for another. Green (1975) has further contended that in such instances, the instrument is clearly measuring something different for each group and, therefore, is differentially valid.

To test for the equivalence of uniquenesses across track, a model in which the number of factors, pattern of factor loadings and all uniquenesses were constrained to be equal across track (Model 3) was compared with one in which the uniquenesses were unconstrained (Model 2); the $\chi^2$ differential was highly significant ($\Delta\chi^2_{21}$ = 139.13). Tests of invariance proceeded next to identify item-pair measurements that were differentially reliable across track. As such, a series of models was specified in which all factor loadings, and one uniqueness under test were held invariant across track. Given findings of invariant uniquenesses, however, these parameters too were cumulatively constrained to be equal across track. For example, in testing for the invariance of MSC5 (item-pair 21), all factor loadings and 9 uniquenesses, in addition to the one being tested, were held invariant (see nonasterisked uniquenesses in Tables 1 & 2). Finally, a test for the invariance of factor covariances resulted in a $\Delta\chi^2$ that was not significant ($\Delta\chi^2_{32}$ = 41.41).[3]

## Conclusions

CFA procedures were used to test the factorial validity of the general and academic subscales of the SDQ III. The results demonstrated a well-defined factor structure yielding one general SC facet, one school SC facet, and two subject matter facets -- English SC and mathematics SC. Tests of invariance revealed item-pair measurements that had equivalent factor loadings, albeit partially nonequivalent uniquenesses; all factor covariances were equivalent.

The major finding was that while the SDQ III measured the same SC facet in the same scale units for both tracks (i.e., invariant factor loadings), it did so with a differential degree of reliability (i.e., noninvariant uniquenesses).

Specifically, 10 items were more reliable for the high, than for the low track; the reverse was true for one item only. Since reliability casts the upper bound for validity, this finding translates into one of differential validity which, in turn, implies item bias in favor of the high track for 10 of the 11 items (see Benson, 1987; Green, 1975).

However, in assessing the factorial invariance of a measuring instrument, practical, as well as statistical significance, must be taken into account. The practical importance of differences in item reliabilities can be more realistically judged by examining the absolute magnitude of these discrepancies (see Werts, Rock, Linn, & Joreskog, 1976). As such, The discrepancies were considered of little practical significance. Indeed, it is highly likely that sample-specific artifacts in the data, such as differential correlated errors across track, contributed importantly to the non-equivalence of item-pair reliabilities (see e.g., Marsh, in press; Tanaka & Huba, 1984). Thus, while the claim that factorial validity is justified only with evidence of both invariant factor loadings and uniquenesses is technically correct (see Benson, 1987), this criterion is considered to be excessively stringent (Muthén, personal communication, January, 1987). For all practical purposes, then, mean comparisons across academic track, should not be detrimentally affected by these differential item-pair reliabilities in the SDQ III.

Overall, given the stringency of the LISREL CFA procedures in general, and the degree of statistical rigor applied in this study in particular, the SDQ III demonstrated excellent psychometric properties. The instrument is also easily administered, easily scored, time-efficient, and easily adapted to specific assessment needs (i.e., each subscale is an independently valid measure of one particular SC facet) --- all important considerations for testing at the secondary school level. Given the scarcity of psychometrically-sound measures of adolescent SC, coupled with the concerns of school counselors, school psychologists, and school administrators for the self-perceptions of low ability students in academically-tracked schools, the SDQ III can become an invaluable assessment tool for these professionals in their measurement of general, school, English, and mathematics SCs for high school students.

References

Benson, J. (1987). Detecting item bias in affective scales. Educational and Psychological Measurement, 47, 55-67.

Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Byrne, B. M. (in press a). Measuring adolescent self-concept: Factorial validity and equivalency of the SDQ III across gender. Multivariate Behavioral Research.

Byrne, B.M. (in press b). Adolescent self-concept, ability grouping, and social comparison: Reexamining academic track differences in high school. Youth and Society.

Byrne, B.M. & Schneider, B.H. (1988). Perceived Competence Scale for Children: Testing for factorial validity and invariance across age and ability. Applied Measurement in Education, 1, 171-187.

Byrne, B.M. & Shavelson, R.J. (1986). On the structure of adolescent self-concept. Journal of Educational Psychology, 78, 474-481.

Byrne, B.M. & Shavelson, R.J. (1987). Adolescent self- concept: Testing the assumption of equivalent structure across gender. American Educational Research Journal, 24, 365-385.

Byrne, B.M., Shavelson, R.J., & Muthén, B. Reexamining multigroup comparisons and the invariance of factor structures: A review and app cation. Manuscript submitted for publication.

Fornell, C. (1983). Issues in the application of covariance structure analysis: A comment. Journal of Consumer Research, 9, 443-448.

Gerbing, D.W. & Anderson, J.C. (1984). On the meaning of within-factor correlated measurement errors. Journal of Consumer Research, 11, 572-580.

Green, D.R. (1975). What does it mean to say a test is biased? Education and Urban Society, 8, 33-52.

Huba, G.J., Wingard, J.A., & Bentler, P.M. (1981). A comparison of two latent variable causal models for adolescent drug use. Journal of Personality and Social Psychology, 40, 180-193.

Joreskog, K.G. (1982). Analysis of covariance structures. In C. Fornell (Ed.), A second generation of multivariate analysis Vol. 1: Methods (pp. 200-242). New York: Praeger.

Joreskog, K.G. & Sorbom, D. (1985). LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood. Mooresville, Ind.: Scientific Software.

Long, J.S. (1983). Confirmatory factor analysis. Beverly Hills, CA.: Sage.

Marsh, H.W., Barnes, J., & Hocevar, D. (1985). Self-other agreement on multidimensional self-concept ratings: Factor analysis and multitrait-multimethod analysis. Journal of Personality and Social Psychology, 49, 1360-1377.

Marsh, H.W. & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. Psychological Bulletin, 97, 562-582.

Marsh, H.W. & O'Neill, R. (1984). Self Description Questionnaire: The construct validity of multidimensional self-concept ratings by late adolescents. Journal of Educational Measurement, 21, 153-174.

Marsh, H.W., Richards, G.E., & Barnes, J. (in press). Multidimensional self-concepts: The effect of participation in an outward bound program. Journal of Personality and Social Psychology.

Muthén, B. & Kaplan D. (1985). A comparison of methodologies for the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, 38, 171-189.

Newcomb, M.D., Huba, G.J., & Bentler, P.M. (1986). Determinants of sexual and dating behaviors among adolescents. Journal of Personality and Social Psychology, 50, 428-438.

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

Tanaka, J.S. & Huba, G.J. (1984). Confirmatory hierarchical factor analyses of psychological distress measures. Journal of Personality and Social Psychology, 46, 621-635.

Silon, E.L. & Harter, S. (1985). Assessment of perceived competence, motivational orientation, and anxiety in segregated and mainstreamed educable mentally retarded children. Journal of Educational Psychology, 77, 217-230.

Werts, C.E., Rock, D.A., Linn, R.L., & Joreskog, K.G. (1976). Comparison of correlations, variances, covariances, and regression weights with or without measurement error. Psychological Bulletin, 83, 1007-1013.

Footnotes

1. A $X^2$ /df ratio < 1.50 (Muthen, personal communication, January, 1987), and a BBI >.90 (Bentler & Bonett, 1980) represent reasonable model approximations to the observed data.
2. The term "uniqueness" is used in the factor analytic sense to mean a composite of specific and random measurement error which, in cross-sectional studies, cannot be separated (for an extended discussion, see Gerbing & Anderson, 1984).
3. Equality constraints were specified for all factor loadings, all covariances, and all invariant uniquenesses (see Tables 1 and 2).

Table 1

Item Factor Loading, Uniqueness, and Reliability Estimates for Low Track ($n$ = 285)[a]

| | Factors[b] | | | | | |
|---|---|---|---|---|---|---|
| Item-pair Measurements | General SC | School SC | English SC | Mathematics SC | Uniqueness[cd] | Reliability |
| GSC 1 | 1.000[e] | 0.0 | 0.0 | 0.0 | .289* | .711 |
| GSC 2 | .957 | 0.0 | 0.0 | 0.0 | .349 | .651 |
| GSC 3 | .974 | 0.0 | 0.0 | 0.0 | .326 | .675 |
| GSC 4 | .749 | 0.0 | 0.0 | 0.0 | .601** | .561 |
| GSC 5 | .927 | 0.0 | 0.0 | 0.0 | .390** | .611 |
| GSC 6 | .770 | 0.0 | 0.0 | 0.0 | .578 | .422 |
| ASC 1 | 0.0 | 1.000[e] | 0.0 | 0.0 | .501 | .499 |
| ASC 2 | 0.0 | 1.159 | 0.0 | 0.0 | .330** | .670 |
| ASC 3 | 0.0 | 1.220 | 0.0 | 0.0 | .257*** | .744 |
| ASC 4 | 0.0 | 1.108 | 0.0 | 0.0 | .387*** | .614 |
| ASC 5 | 0.0 | .984 | 0.0 | 0.0 | .517** | .483 |
| ESC 1 | 0.0 | 0.0 | 1.000[e] | 0.0 | .678 | .322 |
| ESC 2 | 0.0 | 0.0 | 1.058 | 0.0 | .640 | .360 |
| ESC 3 | 0.0 | 0.0 | 1.040 | 0.0 | .652 | .348 |
| ESC 4 | 0.0 | 0.0 | 1.079 | 0.0 | .625* | .375 |
| ESC 5 | 0.0 | 0.0 | .791 | 0.0 | .799** | .202 |
| MSC 1 | 0.0 | 0.0 | 0.0 | 1.000[e] | .572** | .428 |
| MSC 2 | 0.0 | 0.0 | 0.0 | 1.244 | .337 | .663 |
| MSC 3 | 0.0 | 0.0 | 0.0 | 1.301 | .275** | .725 |
| MSC 4 | 0.0 | 0.0 | 0.0 | 1.266 | .314 | .686 |
| MSC 5 | 0.0 | 0.0 | 0.0 | 1.092 | .499 | .501 |

* $p < .05$  ** $p < .01$  ***$p < .001$

[a] Unstandardized solution
[b] All factor loadings were statistically significant and invariant across track
[c] All uniquenesses were statistically significant
[d] Asterisked values indicate noninvariance
[e] Fixed Parameter

SC = self-concept; GSC = general SC; ASC = school SC; ESC = English SC; MSC = mathematics SC; GSC1 = items 1 and 2 measuring GSC; GSC2 = items 3 and 4 measuring GSC, ---------MSC5 = items 9 and 10 measuring MSC

Table 2

Item Factor Loading, Uniqueness, and Reliability Estimates for

High Track (n = 613)[a]

| Item-pair Measurements | Factors[b] General SC | School SC | English SC | Mathematics SC | Uniqueness[cd] | Reliability |
|---|---|---|---|---|---|---|
| GSC 1 | 1.000[e] | 0.0 | 0.0 | 0.0 | .247* | .753 |
| GSC 2 | .920 | 0.0 | 0.0 | 0.0 | .362 | .637 |
| GSC 3 | 1.004 | 0.0 | 0.0 | 0.0 | .240 | .759 |
| GSC 4 | .873 | 0.0 | 0.0 | 0.0 | .426** | .574 |
| GSC 5 | 1.009 | 0.0 | 0.0 | 0.0 | .232** | .767 |
| GSC 6 | .851 | 0.0 | 0.0 | 0.0 | .455 | .545 |
| ASC 1 | 0.0 | 1.000[e] | 0.0 | 0.0 | .587 | .413 |
| ASC 2 | 0.0 | 1.328 | 0.0 | 0.0 | .271** | .728 |
| ASC 3 | 0.0 | 1.400 | 0.0 | 0.0 | .190*** | .809 |
| ASC 4 | 0.0 | 1.316 | 0.0 | 0.0 | .284*** | .715 |
| ASC 5 | 0.0 | 1.214 | 0.0 | 0.0 | .391** | .609 |
| ESC 1 | 0.0 | 0.0 | 1.000[e] | 0.0 | .616 | .384 |
| ESC 2 | 0.0 | 0.0 | 1.094 | 0.0 | .541 | .460 |
| ESC 3 | 0.0 | 0.0 | 1.051 | 0.0 | .576 | .424 |
| ESC 4 | 0.0 | 0.0 | 1.116 | 0.0 | .522* | .478 |
| ESC 5 | 0.0 | 0.0 | .619 | 0.0 | .853** | .147 |
| MSC 1 | 0.0 | 0.0 | 0.0 | 1.000[e] | .347** | .653 |
| MSC 2 | 0.0 | 0.0 | 0.0 | 1.087 | .228 | .772 |
| MSC 3 | 0.0 | 0.0 | 0.0 | 1.134 | .161** | .840 |
| MSC 4 | 0.0 | 0.0 | 0.0 | 1.103 | .207 | .794 |
| MSC 5 | 0.0 | 0.0 | 0.0 | 1.006 | .340 | .661 |

* $p < .05$ ** $p < .01$ ***$p < .001$

[a] Unstandardized solution
[b] All factor loadings were statistically significant and invariant across track
[c] All uniquenesses were statistically significant
[d] Aster' ked values indicate noninvariance
[e] Fixed Darameter

SC = self-concept; GSC = general SC; ASC = school SC; ESC = English SC; MSC = mathematics SC; GSC1 = items 1 and 2 measuring GSC; GSC2 = items 3 and 4 measuring GSC, ---------MSC5 = items 9 and 10 measuring MSC

Table 3

Simultaneous Tests of Invariance for Item-pair Measurements

| Competing Models | | $\chi^2$ | df | $\Delta\chi^2$ | $\Delta$df | $\chi^2$/df | BBI |
|---|---|---|---|---|---|---|---|
| 0 | Null Model | 11,895.61 | 420 | --- | --- | 28.32 | --- |
| 1 | Four SC factors invariant | 1,230.64 | 366 | --- | --- | 3.36 | .90 |
| 2 | Model 1 with all factor loadings invariant | 1,254.55 | 383 | 23.91 | 17 | 3.28 | .89 |
| 3 | Model 2 with all uniquenesses invariant[a] | 1,393.68 | 404 | 139.13 | 21*** | 3.45 | .88 |
| 4 | Model 2 with invariant uniquenesses and factor covariances invariant | 1,275.83 | 399 | 21.28 | 16 | 3.20 | .89 |

[a] Results from tests for the invariance of uniquenesses across track are reported in Tables 1 and 2.

SC = self-concept