DOCUMENT RESUME

ED 284 905 TM 870 458

AUTHOR	Byrne, Barbara M.
TITLE	Multitrait-Multimethod Analyses of Three Self-Concept Scales: Testing Equivalencies of Construct Validity across ability.
PUB DATE	Apr 87
NOTE	45p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability Grouping; *Construct Validity; Correlation; Factor Analysis; High Schools; Mathematical Models; *Multitrait Multimethod Techniques; Rating Scales; *Self Concept; *Self Concept Measures; Self Esteem; Semantic Differential; Statistical Studies; Track System (Education)
IDENTIFIERS	*Confirmatory Factor Analysis; Convergent Validation; *Covariance Structure Models; Discriminant Validity; Guttman Scales; Likert Scales

ABSTRACT

The construct validity of four self-concept (SC) traits (general SC, academic SC, English SC, mathematics SC), as measured by three different scales (Likert, semantic differential, Guttman) for low- (n=252) and high-track (n=588) Canadian high school students, was assessed using both the Campbell-Fiske criteria, and a comparison of hierarchically nested covariance structure models. Confirmatory factor analysis was used to model hypotheses related to convergent and discriminant validity and to test directly equivalencies of traits and methods. Findings indicate that assumptions of invariant construct validity cannot be taken for granted; differences in both the measurement and structure of SC were found. Academic SC, as measured by the Likert and Guttman scales, was problematic for the high track. These scales appeared to elicit different types of responses from high and low ability students. Tests of invariance formally confirmed this result. Discriminant validity of the trait factors was also less clear for the high track, but this may have been a measurement problem. Method bias was clearly more evident for the high than for the low track. Method bias effects for each scale type, as well as all but one trait correlation, were found to be noninvariant. A 5-page list of references and eight tables are included. (LPG)

Multitrait-multimethod Analyses of Three Self-concept Scales: Testing Equivalencies of Construct Validity Across Ability

Barbara M. Byrne

University of California, Los Angeles

Paper presented at the American Educational Research Association Annual Meeting, Washington, D.C. 1987.

Abstract

The construct validity of four self-concept (SC) traits (general SC, academic SC, English SC, mathematics SC), as measured by three different measurement scales (Likert, semantic differential, Guttman) for low ($\underline{n}$ = 252) and high ($\underline{n}$ = 588) track high school students was assessed using both the Campbell-Fiske criteria, and a comparison of hierarchically nested covariance structure models. Confirmatory factor analysis was used to model hypotheses related to convergent and discriminant validity, and to test directly, equivalencies of traits and methods. Findings indicate that assumptions of invariant construct validity cannot be taken for granted; differences in both the measurement and structure of SC were found. The study has important implications for substantive research that focuses on the comparison of mean differences in multidimensional SCs across populations, and in particular, in general, academic, English, and mathematics SCs across ability levels of high school students.

Multitrait-multimethod Analyses of Three Self-concept Scales: Testing Equivalencies of Construct Validity Across Ability

A wealth of self-concept (SC) resea has focused on mean differences in multidimensional SCs across ability (see Byrne, 1984; Wylie, 1979). An important assumption in testing for these differences is (a) evidence of the construct validity of SC measures and constructs within each group and, (b) the equivalence of SC measures and constructs across groups (Cole & Maxwell, 1985). In substantive research, however, this assumption is implicit in the comparison of groups, and is rarely tested directly. The present study, in broad terms, assesses the construct validity of a multidimensional SC structure as measured by three different measurement scales, and tests the equivalencies of construct validity across two ability levels of high school students.

In construct validation, a researcher seeks empirical evidence in support of hypothesized construct relations (a) among facets of the same construct (within-network relations), and (b) among different constructs (between-network relations). These theoretical linkages represent the nomological network of an hypothesized construct (Cronbach & Meehl, 1955). Although construct validation encompasses an interplay of theory construction, test development, and data collection (Shavelson, Hubner, & Sta... the two processes are complementary, rather than c... hat is to say, given an adequate theory, one ca... e instrument; given an adequate instrument, th... can be tested. Construct validation, then, is an ongoing process involving hypotheses that need to be challenged repeatedly with counterhypotheses (Anastasi, 1986; Cronbach, 1971; Cronbach & Meehl, 1955).

Campbell and Fiske (1959) posited that claims of construct validity must be accompanied by evidence of both convergent and discriminant validity. As such, a measure should correlate highly with other measures to which it is theoretically linked (convergent validity), and correlate negligibly with those that are theoretically unrelated (discriminant validity). To determine evidence of construct validity, they proposed that measures of multiple traits be assessed by multiple methods and that all trait-method correlations be arranged in a multitrait-multimethod (MTMM) matrix. The assessment of construct validity then focuses on comparisons among three blocks of correlations: (a) scores on the same traits measured by different methods (monotrait-heteromethod values i.e., convergent validity), (b) scores on different traits measured by the same method (heterotrait-monomethod values i.e., discriminant validity) and, (c) scores on different traits measured by different methods (heterotrait-heteromethod values i.e., discriminant validity). Specific criteria guide the inspection of these values and are described later.

While the seminal work of Campbell and Fiske (1959) represents a major contribution to the field of psychometrics, researchers have noted several shortcomings in their procedure (see e.g., Hubert & Baker, 1978; Kavanagh, MacKinney, & Wolins, 1971; Marsh & Hocevar, 1983; Schmitt, 1978; Widaman, 1985). In particular, many researchers have criticized the subjectivity of the criteria upon which construct validity is based, and have proposed alternative quantitative methodologies (for a review, see Schmitt & Stults, 1986).

One methodologically more sophisticated approach to assessing construct validity within the MMTM framework is the analysis of covariance structures using the confirmatory factor analytic (CFA) procedure originally proposed by Joreskog (1971), and now commercially avaliable to researchers through the LISREL VI computer program (Joreskog & Sorbom, 1985). The relative merits of CFA in analyzing MTMM matrices is now well documented (see e.g., Marsh & Hocevar, 1983; Schmitt & Stults, 1986; Widaman, 1985). As compared with the Campbell-Fiske procedure, a summary of the major advantages of CFA relevant to the present paper are as follows: (a) the MTMM matrix is explained in terms of the underlying latent constructs, rather than the observed variables, thus obviating influences of measurement error; (b) the evaluation of convergent and discriminant validities can be made at both the matrix and individual parameter levels; (c) based on a series of hierarchically nested models, hypotheses related to convergent and discriminant validities can be tested statistically, and (d) separate estimates of variance due to traits, methods, and error/uniquenesses are provided.

The validity of SC has been examined within a MTMM design using both Campbell-Fiske and CFA procedures. Evidence of convergent and discriminant validity for both trait and method factors, and support for the multidimensional structure of SC for students in grades 5 through college have been reported (Marsh & O'Neill, 1984; Marsh, Parker, & Smith, 1983; Marsh, Smith, Barnes, & Butler, 1983). In particular, general SC, academic SC, English SC and mathematics SC, although correlated, could be measured as separate constructs. Other construct validity studies of SC measures have generally reported moderate evidence of convergent validity with other SC measures and/or external criteria. However, evidence of discriminant validity is inconsistent (see Byrne, 1984 for a review).

The construct validity of different measurement scales has also been examined within a MTMM framework using both Campbell-Fiske and CFA procedures. Findings have been consistent in reporting evidence of convergent validity for Likert, semantic differential, and Guttman scales (Flamer, 1983; Jaccard, Weber, & Lundmark, 1975; Kothandapani, 1971; Ostrom, 1969). Evidence of discriminant validity, however, has been inconsistent. Modest method bias for the Likert and Guttman scales has been reported (Kothandapani, 1971). However, in a reanalysis of the Ostrom and Kothandapani MTMM data using CFA, Bagozzi (1978) and Schmitt (1978) reported opposing conclusions regarding the convergent and discriminant validity findings (but see Widaman, 1985). Finally, Flamer's CFA analysis confirmed his former findings, and also reported evidence of a method-trait interaction; Likert and semantic differential scales differed in the way they measured a particular trait.

Although each of these studies used either Campbell-Fiske or CFA procedures to examine construct validity within a MTMM framework, none examined data either for a particular ability group (e.g., low track), or across ability groups (e.g., low track vs high track). Cole and Maxwell (1985) however, have noted that evidence of construct validity within one population in no way guarantees its equivalence across populations. As a case in point, Byrne and Shavelson (in press) found differences in the way English and mathematics SCs related to general and academic SCs for adolescent males and females; they also found significant gender differences in the reliability of certain measuring instruments. Indeed, findings from substantive studies of academic tracking in high school suggest the possibility of parallel construct validity differences based on SC responses from low and high ability students. For example, low track students have been shown to have weaker reading comprehension skills than high track students (Addy, Henderson, & Knox, 1980). As such, their interpretation of test items on particular measurement scales may differ from those of their high track peers. Such findings would bear importantly on the construct validity of the measures, and the traits underlying them.

The present study has three purposes. First, to assess the construct validity of four SC traits (general SC, academic SC, English SC, mathematics SC) as measured by three different measurement scales (Likert, semantic differential, Guttman), for low and high track students. Second, to compare construct validity findings based on two different approaches to analyzing MTMM matrices -- Campbell-Fiske criteria and confirmatory factor analysis. Finally, to test directly, the equivalencies of SC measurements and structure across academic high school tracks.

Method

Sample and Procedure

The original sample comprised 988 (324 low track, 664 high track) grades 11 and 12 students from two suburban high schools in Ottawa, Canada. Following listwise deletion of missing data, the final sample size was 840 (252 low track, 588 high track). The data approximated a normal distribution with skewness ranging from -1.19 to .19 ($\bar{X}$ = -.27) for low-track, and from -1.26 to .10 ($\bar{X}$ = -.50) for high-track students; kurtosis ranged from -.53 to 1.60 ($\bar{X}$ = .23) for the low track, and from -.92 to 1.83 ($\bar{X}$ = .27) for the high track. Since English is part of the core curriculum for high schools in Ontario (i.e. compulsory), it was known that all students were enrolled in at least one English course, and therefore, only mathematics classes were tested for the study.

In the province of Ontario, tracking in high school is applicable only to the core curricula. For each academic subject (e.g. mathematics, science, history, geography, English, French), two courses are structured; one designed to meet the needs of high ability students (advanced level courses) and the other, low ability students (general level courses). General level courses are considered "appropriate preparation for employment or further education in colleges and other non-university educational institutions" (Ontario Ministry of Education, 1979-81, p.7). Although a definition of high and low academic tracks has not been formalized by the Ontario Ministry of Education, most Ontario secondary schools in general (King & Hughes, 1985), and the participating schools in the present study in particular, classify low-track students as those taking two or more of their mathematics and science courses in any given year, at the general level; all other students are considered high-track.

A battery of SC instruments (described below) were administered to intact classroom groups during one 50-minute period. The testing was completed approximately two weeks following the April report cards. The students therefore had the opportunity of being fully cognizant of their academic performance prior to completing the tests for the study. This factor was considered important in the measurement of academic and subject specific SCs.

Instrumentation

The SC test battery consisted of 12 instruments; three measures for each of general SC, academic SC, English SC, and mathematics SC. All instruments were self-report rating scale formats and were designed for use with a high school population. They were selected because they purported to measure (with some justification) the SC facets in the theory to be tested.

Likert scale. The Self Description Questionnaire III (SDQ; Marsh & O'Neill, 1984) 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. Academic SC, English SC, and mathematics SC were measured by the Academic SC, Verbal SC, and Mathematics SC subscales, respectively; each contained 10 items. Internal consistency reliability coefficients ranging from .86 to .93 (Md α= .90) for each of these subscales, and strong support for their construct validity based on interpretations consistent with the Shavelson et al. (1976) model of SC have been reported (Byrne & Shavelson, 1986; Marsh & O'Neill, 1984).

Semantic differential scale. The Affective Perception Inventory (API; Soares & Soares, 1979) is a semantic differential scale with a forced-choice format containing four categories maintained along a continuum between two dichotomous terms (e.g. "happy", "unhappy"). The Self Concept, Student Self, English Perceptions, and Mathematics Perceptions subscales were used to measure general SC, academic SC, English SC, and Mathematics SC, respectively. The number of items comprising each of the API subscales is as follows: Self Concept 25; Student Self 25; English Perceptions 22;

Mathematics Perceptions 17. Internal consistency coefficients ranging from .79 to .95 (Md = .85) have been reported for these subscales (Byrne & Shavelson, 1986; Soares & Soares, 1980). Convergent validity coefficients ranging from .49 to .55 (Md r = .50 with peer ratings, and from .37 to .74 (Md r = 48.5) with teacher ratings for the same subscales, as well as evidence of discriminant validity, have also been reported (Soares & Soares, 1980).

Guttman scales. The Self-esteem Scale (SES; Rosenberg, 1965) is a 10-item Guttman scale based on a 4-point format ranging from "strongly agree" to "strongly disagree; it was used to measure general SC. A test- retest reliability of .62 (Byrne, 1983), and an internal consistency reliability coefficient of .87 (Byrne & Shavelson, 1986) have been reported, as well as convergent validities ranging from .56 to .67 (see Byrne, 1983). The 8-item Self Concept of Ability Scale (SCAS; Brookover, 1962) also a Guttman scale, has a response format based on a 5-point format. Respondents are asked to rank their ability in comparison with others, on a scale from "1-I am the poorest" to "5-I am the best". Form A was used to measure academic SC. Forms B and C were used to measure English SC and mathematics SC, respectively. Items on Forms B and C are identical to those on Form A, except that they elicit responses relative to specific academic content (e.g. "how do you rate your ability in English (mathematics) compared to your close friends'?"). Test-retest and internal consistency reliability coefficients ranging from .69 to .72, and from .77 to .94, respectively, have been reported (see Byrne, 1983; Byrne & Shavelson, 1986).

Analysis of the Data

Responses to negatively worded items were reversed so that for all instruments, the highest response code was indicative of a positive rating of SC. Additionally, the first item on the API Self Concept subscale ("I am masculine----I am feminine") was recoded, so that it was contingent on gender.

The data were analyzed in three stages. First, zero-order correlations among all measures were arranged in a MTMM matrix, and then examined separately for evidence of construct validity based on the Campbell-Fiske criteria, for each track. Second, using CFA procedures, a 7-factor model of the data comprising four trait factors (general SC, academic SC, English SC, mathematics SC) and three method factors (Likert, semantic differential, Guttman scales) was proposed and tested separately for each track. A schematic representation of this model is presented in Figure 1. Finally, equivalencies of SC measurements and structure were tested across track.

Insert Figure 1 about here

Campbell-Fiske Criteria. Campbell and Fiske (1959) proposed four criteria for evaluating convergent and discriminant validity. These criteria are:

1. The convergent validities should be significantly different from zero and sufficiently large to warrant further investigation of validity.

2. The convergent validities should be higher than correlations between different traits assessed by different methods (heterotrait-heteromethod blocks).

3. The convergent validities should be higher than correlations between different traits assessed by the same method (heterotrait-monomethod blocks).

4. The pattern of correlations between different traits should be the same in both the heteromethod and monomethod blocks.

For each track, comparisons of various blocks of correlations involved determining the proportion of times that these criteria were satisfied.

Confirmatory Factor Analysis. For each track, a 7-factor model comprising four traits and three methods was hypothesized and tested for convergent and discriminant validity by means of

(a) comparisons with alternatively specified models, and (b) examination of individual parameter estimates. All CFA analyses were conducted using LISREL VI (Joreskog & Sorbom, 1985).

Traditionally, in covariance structure analysis, the extent to which a proposed model fits the observed data has been based on the χ^2 likelihood ratio test. However, problems related to the dependency of χ^2 on sample size have been noted (see e.g, Bentler & Bonett, 1980). Thus, in addition to the statistical fit of a model, a measure of its practical fit must also be considered (Widaman, 1985). To this aim, Bentler and Bonett proposed a normed index of fit (delta) that ranges from 0.0 to 1.0. Joreskog (Joreskog, 1971; Joreskog & Sorbom, 1985), among others, have posited that assessment of model fit should be based on multiple criteria. This was accomplished in the present study by using (a) the χ^2 likelihood ratio, (b) the χ^2 /degrees of freedom (df) ratio, (c) the delta index[1], (d) T-values and modification indices provided by the LISREL VI program, and (e) knowledge of substantive and theoretical research in this area.

To establish various validity criteria, the proposed 7-factor model was tested against a series of more restrictive models in which specific parameters were either eliminated or constrained to equal zero. Since the difference in χ^2 ($\Delta\chi^2$) is itself χ^2-distributed, with degrees of freedom equal to the difference in degrees of freedom for the two models, the fit differential between comparison models can be tested statistically. A significant $\underline{\Delta\chi^2}$ argues for the superiority of the less restrictive model. Additionally, the difference in practical fit can be noted. (see Widaman, 1985, for a more detailed discussion of these model comparisons).

The parameter estimates for trait and method factor loadings, trait intercorrelations, method intercorrelations, and estimated error uniquenesses were examined with respect to magnitude and statistical significance; the latter being determined by the z-ratio (parameter estimate/standard error) which is printed as a T-value by LISREL VI. T-values >2.00 are considered statistically significant at the .05 level (Joreskog & Sorbom, 1985).

Tests of Invariance. Testing for the equivalency of traits and methods involved the comparison of a series of models in which certain parameters were constrained to be equal across track, with less restrictive models in which these parameters were free to take on any value. The difference in $\underline{\chi^2}$, as described above, was used to determine the statistical significance of the hypotheses tested.

Results

Construct Validity Based on Campbell-Fiske Criteria

The matrices of zero-order correlations, computed separately for each track, are presented in Table 1, together with the means, standard deviations, and internal consistency alpha reliabilities. Results are entered below the main diagonal for the low track, and above the main diagonal for the high track.

Insert Table 1 about here

Criterion 1. Convergent validities were all statistically significant (p <.05) for both the low track (Md r = .60) and the high track (Md r = .69). Convergent validity for English SC as measured by the Likert and Guttman scales, however, was only moderate, even with findings of higher validity for the high track (low track, r = .43; high track, r = .56).

Criterion 2. Convergent validities were consistently higher than correlations between different traits assessed by different methods (heterotrait-heteromethod triangles) for both the low track (36 of 36 comparisons) and the high track (35 of 36 comparisons).

Criterion 3. Convergent validities were for the most part, consistently higher than correlations between different traits measured by the same method (heterotrait-monomethod triangles) for both the low track (14 of 18 comparisons) and the high track (15 of 18 comparisons). In particular, the semantic differential and Guttman scales both exhibited some method bias; this effect, however, was stronger for the Guttman scales.

Criterion 4. For both tracks, the pattern of correlations among the different traits was fairly similar across methods; three correlations derived from the semantic differential and Guttman measures were differentially disproportionate across track.

Construct Validity Based on Confirmatory Factor Analyses

Goodness-of-fit indices for the series of MTMM models tested are presented in Tables 2 and 3 for the low and high tracks, respectively. Model 1 is the most restrictive model, representing the null hypothesis that each observed measure is an independent factor; it serves as the null model against which competing models are compared in order to determine the delta index. Models 2-4 represent decreasingly restrictive models, such that Model 4 is the least restrictive, having both correlated traits and correlated methods; it serves as the baseline model since it represents hypothesized relations among the traits and methods and, typically, demonstrates the best fit to the data.[2]

Insert Tables 2 and 3 about here

Although for both tracks Model 4 represented the best fit to the data, the fit, based on statistical criteria, was not good. This lack of fit indicated some degree of misspecification in the model (see Kaplan, 1987); it was expected that the subsequent analyses would identify possible areas of misspecification. Due to problems of estimation, as well as other considerations (see Widaman, 1985), additional fitting of the hypothesized model was not conducted. Model 4, then, indicated that both the trait and method factors were correlated. These correlations for the low track, however, were extremely weak, as indicated by the small difference, albeit significant ($p < .05$), in statistical ($\Delta\chi^2_3 = 9.48$) and practical (χ^2/df = 0.0; delta = .02) fit criteria between Model 4 and Model 3 in which the methods were uncorrelated. These results suggest that for the low track, the three measurement scales were operating independently.

Evidence of convergent validity was tested by comparing Model 4 with Model 5 in which no trait factors were specified. As shown in Table 4, the $\Delta\chi^2$ was highly significant for both tracks, thus providing strong evidence of convergent validity for the trait factors. Since complete discriminant validity of traits argues for zero intercorrelations, evidence of same can be tested by comparing the baseline model (Model 4) with one in which perfect correlations among traits are hypothesized (Model

6). The results in Table 4 indicate that for both tracks, discriminant validity of the traits was evident as indicated by the highly significant differences in ___. Finally, the discriminant validity of method factors (i.e. no method bias) was tested by comparing Model 4 with Model 2 in which no method factors were specified. Again, for both tracks, this comparison yielded statistically significant ____'s, suggesting fairly strong evidence of method bias effects.

Insert Table 4 about here

To determine the extent to which each measurement scale was contributing to the method bias, Model 4 was further compared with three additional models, each of which eliminated one of the three methods. With one exception, each of the comparisons indicated significant method effects; those associated with the semantic differential, for the low track, were not significant. The results in Table 4 demonstrate that while the Likert measures made the heaviest contribution to method bias for the low track, the Guttman measures were more important for the high track. Scales contributing the least to method bias were the semantic differential for the low track, and the Likert for the high track.

More precise assessments of trait- and method-related variance can be ascertained by examining the individual parameter estimates as specified for Model 4. These results are presented in Tables 5 and 6 for the low and high tracks, respectively. The magnitude of the trait loadings for both tracks are shown to be generally consistent with the earlier convergent validity findings (see Table 4); all loadings for the low track, and all but one for the high track were significant. With the exception of academic SC, as measured by the Likert and Guttman scales for the high track, each trait factor was well defined by the hypothesized model.

Insert Tables 5 and 6 about here

Method factor loadings, overall, tended to be larger for the high, than for the low track. Method-related variance for the high track was substantial for all but three measurements; all parameter estimates were statistically significant. In contrast, only seven of the 12 method parameters were significant for the low track. Interestingly, the measurement of general SC was associated with a modest degree of method effects for each of the scales.

Discriminant validity of traits and methods are determined by examining the factor correlation matrices. Results generally supported earlier findings from the overall measures of goodness-of-fit (see Table 4). However, evidence of trait discriminant validity for the high track was less clear than for the low track. Marsh and Hocevar (1983) noted that only when correlations are extreme (i.e., approach unity) should researchers be concerned about a lack of discriminant validity. As such, claims of discriminant validity of the traits appears justified for both tracks. However, Marsh and Hocevar also argued for trait correlations consistent with the underlying theory. This is not the case for the high track; trait correlations are not totally consistent with SC theory involving these particular traits. In particular, correlations between academic SC and mathematics SC, and between English SC and mathematics SC, typically, yield values of approximately .50 and .01, respectively (see e.g., Byrne & Shavelson, 1986; Marsh & Shavelson, 1985). As such, discriminant validity of the traits for the high track cannot be clearly interpreted on the basis of these findings.

Lack of discriminant validity among method factors was clearly more evident for the high, than for the low track. These findings suggest that whereas, for the most part, each measurement scale operated independently for the low track, this was not so for the high track; a higher degree of method bias was evident.

Tests of Invariance

In testing for invariance, the parameters were estimated simultaneously for each track. The first step was to test the assumption of overall invariance across ability (i.e., is there, or is there not, a difference in the low and high track varance-covariance matrices?). Since this assumption was rejected (χ^2_{78} = 199.64, $p<.001$), hypotheses related to the invariance of traits and methods across ability were formally tested by comparing a series of increasingly restrictive models. Results from tests for the invariance of SC measurements and structure are presented in Tables 7 and 8 respectively.

Insert Table 7 about here

The simultaneous 4-factor solution for each group yielded a reasonable fit to the data (χ^2/df = 3.79). These results suggest that for both tracks, the data were fairly well described by the general, academic, English, and mathematics SC factors. Thus, a series of models were tested by comparing one in which certain parameters were constrained to be equal across track, against one in which these parameters were free to take on any value. For example, the hypothesis of an invariant pattern of trait loadings was tested by constraining these parameters to be equal across track, and then comparing this model (Model 2) with Model 1, in which only the number of factors was held invariant. Since the difference in χ^2 was significant ($\Delta\chi^2_{12}$ = 239.09), this hypothesis was considered untenable. Similarly, the hypothesis of an invariant pattern of general SC loadings was tested, but found tenable.

Given findings of a nonsignificant $\Delta\chi^2$, specified factor loading parameters were held cumulatively invariant, thus providing an extremely powerful test of factorial invariance. Space limitations preclude further elaboration of the invariance testing procedures. However, detailed elsewhere, are descriptions of the procedure in general (e.g., Joreskog, 1971), and an application similar to the present one, in particular (Byrne & Shavelson, in press).

Insert Table 8 about here

Overall, the results indicate that whereas all measures of general SC and English SC were invariant across track, this was not so for academic and mathematics SCs. Academic SC, as measured by the SDQ III and SCAS, differed for the two groups. Likwise, the API measurement of mathematics SC was not consistent across track. Each of the method factors and, all but one trait correlation, were found to differ significantly across track; the correlation between general and academic SC was equivalent.

Summary and Discussion

The construct validity of four SC traits (general SC, academic SC, English SC, mathematics SC) as measured by three different measurement scales (Likert, semantic differential, Guttman) for low and high track students was assessed using both the Campbell-Fiske criteria and CFA procedures. The results from both analyses, in general, supported fairly strong evidence of convergent validity and evidence of method bias for both groups. CFA procedures, including tests of the invariance of traits and methods across tracks, provided a more detailed insight into the group-specific aspects of these findings.

Overall, construct validity findings yielded four major differences between low- and high-track students. First, academic SC, as measured by the Likert and Guttman scales, was problematic for the high track. Relatedly, the strongest method loadings were associated with these same measures. It appears that items on the Likert and Guttman scales measuring academic SC elicited different types of responses from high and low ability students. Quite possibly, different perceptions of academic SC by the two groups of students bear importantly on the problems of model misspecification noted earlier.

Second, discriminant validity of the trait factors was less clear for the high, than for the low track. However, this finding may, in fact, be a measurement, not a structural problem. The fact that the Likert and Guttman scales were in some way measuring academic SC differently from the semantic differential scale for the high track, indicates a trait-method interaction effect and likely contributes to the poor discrimination among the trait factors.

Third, method bias was clearly more evident for the high, than for the low track. The large method intercorrelations indicate that responses by high ability students to items measuring a particular trait would be similar, regardless of which of the three scaling formats were used. In other words, given a particular score on general SC as measured by the Likert scale say, high track students would be equally likely to obtain a similar score on either the semantic differential or Guttman scales. When the impact of each method factor was examined separately, these effects differed across track. Whereas the Likert scales contributed the most method bias to scores by the low track, the Guttman scales contributed the most for the high track. Contributing the least to method bias were the semantic differential and Likert scales for the low and high tracks, respectively. However, these results, particularly with respect to the Likert scale, are not consistent with earlier findings based on the Campbell-Fiske criteria.

Finally, tests of invariance formally tested, and confirmed, earlier findings that the Likert and Guttman scales differed in the measurement of academic SC across abilities; this was also found to be so for mathematics SC, as measured by the semantic differential scale. Furthermore, method bias effects for each scale type, as well as all but one trait correlation, were found to be noninvariant.

Taken together, the findings from this study demonstrate that assumptions of equivalent construct validity across groups cannot be taken for granted. Differences were found with respect to both the measurement and structure of SC. These results yield important implications for substantive research focusing on mean differences in multidimensional SCs across populations, and in particular, in measurements of general, academic, English, and mathematics SCs across ability levels of high school students.

References

Addy, R.J., Henderson, C., & Knox, W.G. (1980). The general store: Meeting the needs of the general level student (Resource Booklet No.ISBNO-920930-04-2). Toronto: Professional Development Committe, Ontario Secondary School Teachers' Federation.

Anastasi, A. (1986). Evolving concepts of test validation. Annual Review of Psychology, 37, 1-15.

Bagozzi, R.P. (1978). The construct validity of the affective, behavioral, and cognitive components of attitude by analysis of covariance structures. Multivariate Behavioral Research, 13, 9-31.

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

Brookover, W.B. (1962). Self-concept of Ability Scale. East Lansing, Mich.: Educational Publication Services.

Byrne, B.M. (1983). Investigating measures of self-concept. Measurement and Evaluation in Guidance, 16, 115-126.

Byrne, B.M. (1984). The general/academic self-concept nomological network: A review of construct validation research. Research of Educational Research, 54, 427-456.

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. (in press). Adolescent self-concept: Testing the assumption of equivalent structure across gender. American Educational Research Journal.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cole, D.A. & Maxwell, S.E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. Multivariate Behavioral Research, 20, 389-417.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational measurement (pp.443-507). Washington D.C.: American Council on Education.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Flamer, S. (1983). Assessment of the multitrait-multimethod matrix validity of Likert scales via confirmatory factor analysis. Multivariate Behavioral Research, 18, 275-308.

Hubert, L.J. & Baker, F.B. (1978). Analyzing the multitrait-multimethod matrix. Multivariate Behavioral Research, 13, 163-179.

Jaccard, J., Weber, J., & Lundmark, J. (1975). A multitrait-multimethod analysis of four attitude assessment procedures. Journal of Experimental Social Psychology, 11, 149-154.

Joreskog, K.G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426.

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

Kaplan, D. (1987). The impact of specification error on the estimation, testing, and improvement of structural equation models. Multivariate Behavioral Research. Manuscript accepted for publication.

Kavanagh, M.J., MacKinney, A.C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

King, A.J.C. & Hughes, J. (1985). Secondary school to work: A difficult transition (Report No. ISBN-0-920930-21-2). Toronto: The Research Committee of the Ontario Secondary School Teachers' Federation.

Kothandapani, V. (1971). Validation of feeling, belief, and intention to act as three components of attitude and their contribution to prediction of contraceptive behavior. Journal of Personality and Social Psychology, 19, 321-333.

Marsh, H.W. & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 20, 231-248.

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

Marsh, H.W., Parker, J.W., & Smith, I.D. (1983). Preadolescent self-concept: Its relation to self-concept as inferred by teachers and to academic ability. The British Journal of Educational Psychology, 53, 60-78.

Marsh, H.W. & Shavelson, R.J. (1985). Self-concept: Its multifaceted, hierarchical structure. Educational Psychologist, 20, 107-123.

Marsh, H.W., Smith, I.D., Barnes, J., & Butler, S. (1983). Self-concept: Reliability, stability, dimensionality, validity and the measurement of change. Journal of Educational Psychology, 75, 772-790.

Ostrom, T.M. (1969). The relationship between the affective behavioral, and cognitive components of attitude. Journal of Experimental Social Psychology, 5, 12-30.

Schmitt, N. (1978). Path analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 2, 157-173.

Schmitt, N. & Stults, D.M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10, 1-22.

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

Soares, A.T. & Soares, L.M. (1979). The Affective Perception Inventory - Advanced Level. Trumbell, Conn.: ALSO.

Soares, A.T. & Soares, L.M. (1980). The Affective Perception Inventory: Test manual/advanced level. Trumbell, Conn.: ALSO.

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.

Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement, 9, 1-26.

Wylie, R. (1979). The self-concept volume 2: Theory and research on selected topics. Lincoln: University of Nebraska Press.

Footnotes

1. A χ^2/df ratio ranging from 1.00 to 5.00 (Wheaton, Muthen, Alwin, & Summers, 1977), and a delta index >.90 (Bentler & Bonett, 1980) are considered a reasonable fit to the data.

2. For reasons related to identification and estimation problems, trait-method factors were fixed to zero for all analyses (see Schmitt & Stults, 1986; Widaman, 1985).

Table 1

Multitrait-multimethod Matrix of Zero-order Correlations Among Self-concept Measures for Low and High Tracks[a]

Measure	Likert (SDQIII) GSC	ASC	ESC	MSC	Semantic Differential (API) GSC	ASC	ESC	MSC	Guttman (SCAS) GSC	ASC	ESC	MSC
Likert												
GSC	--	.33	.28	.17	.66	.56	.16	.18	.81	.26	.11	.13
ASC	.32	--	.40	.43	.33	.63	.42	.46	.32	.66	.42	.40
ESC	.30	.28	--	-.01	.22	.35	.73	.05	.30	.35	.56	.02
MSC	.24	.35	-.06	--	.20	.33	-.03	.89	.23	.50	-.01	.84
Semantic Differential												
GSC	.61	.26	.20	.28	--	.62	.20	.27	.67	.27	.12	.20
ASC	.45	.57	.38	.35	.55	--	.42	.41	.57	.54	.34	.35
ESC	.15	.43	.62	.03	.18	.47	--	.07	.21	.35	.70	.01
MSC	.25	.39	.05	.78	.26	.42	.23	--	.27	.52	.04	.82
Guttman												
GSC(SES)	.75	.26	.27	.26	.59	.46	.11	.24	--	.31	.15	.19
ASC	.27	.58	.25	.23	.23	.52	.37	.35	.27	--	.54	.61
ESC	.24	.37	.43	.01	.26	.41	.50	.02	.25	.51	--	.09
MSC	.24	.35	-.02	.72	.21	.37	.08	.75	.22	.45	.07	--
Low Track												
M	76.00	49.58	54.92	41.69	76.88	70.29	57.82	44.88	31.18	24.80	25.33	23.02
SD	13.40	12.40	9.45	13.37	9.07	8.84	10.62	10.61	4.84	4.47	4.84	5.82
α	.91	.86	.73	.87	.83	.82	.87	.94	.85	.79	.84	.89
High Track												
M	75.71	57.77	57.47	49.00	76.76	73.72	61.75	47.24	31.45	30.26	28.90	26.25
SD	14.58	11.78	9.93	16.92	9.44	9.59	11.21	11.64	5.07	4.94	5.73	7.97
α	.94	.89	.81	.94	.86	.85	.89	.95	.88	.86	.90	.95

[a] Correlations for low track are below the main diagonal, and for high track above main diagonal.

Note: The underlined values are convergent validities. The values in solid triangles are discriminant validities (heterotrait-monomethod correlations); those in broken triangles are discriminant validities (heterotrait-heteromethod correlations).

All correlations > .11 are significant ($p<.05$) α = alpha reliability coefficient; GSC = general self-concept (SC); ASC = academic SC; ESC = English SC; MSC = mathematics SC; SDQ III = Self Description Questionnaire III; API = Affective Perception Inventory; SCAS = SC of Ability Scale.

Table 2

Goodness-of-fit Indices for Multitrait-multimethod Models

Low Track ($\underline{n}$=252)

	Model	X^2	df	X^2/df	delta
1.	12 uncorrelated factors (null model)	1681.05	66	25.47	--
2.	4 correlated trait factors no method factors	216.26	48	4.51	.871
3.	4 correlated trait factors 3 uncorrelated method factors	114.69	36	3.19	.914
4.	4 correlated trait factors 3 correlated method factors (baseline model)	105.21	33	3.19	.937
5.	no trait factors 3 correlated method factors	868.09	51	17.02	.484
6.	4 perfectly correlated trait factors, freely correlated method factors	403.61	39	10.35	.760
7.	4 correlated trait factors 2 correlated method factors (semantic differential, Guttman)	154.14	39	3.95	.908
8.	4 correlated trait factors 2 correlated method factors (Likert, Guttman)	110.73	39	2.83	.932
9.	4 correlated trait factors 2 correlated method factors (Likert, semantic differential)	133.00	39	3.41	.921

Table 3

Goodness-of-fit Indices for Multitrait-multimethod Models

High Track ($\underline{n}$=588)

	Model	χ^2	df	χ^2/df	delta
1.	12 uncorrelated factors (null model)	5480.71	66	83.04	--
2.	4 correlated trait factors no method factors	642.79	48	13.39	.883
3.	4 correlated trait factors 3 uncorrelated method factors	302.70	36	8.41	.944
4.	4 correlated trait factors 3 correlated method factors (baseline model)	185.98	33	5.64	.966
5.	no trait factors 3 correlated method factors	3114.75	51	61.07	.432
6.	4 perfectly correlated trait factors, freely correlated method factors	1484.21	39	38.06	.729
7.	4 correlated trait factors 2 correlated method factors (semantic differential, Guttman)	310.09	40[a]	7.75	.943
8.	4 correlated trait factors 2 correlated method factors (Likert, Guttman)	338.44	40[a]	8.46	.938
9.	4 correlated trait factors 2 correlated method factors (Likert, semantic differential)	463.12	40[a]	11.58	.915

[a] To offset the estimation of a Heywood case, the error variance of the self-concept of Ability Scale Form A was fixed to .01; this accounted for the extra degree of freedom.

Table 4

Goodness-of-fit Indices for Comparison of Multitrait-multimethod Models[a]

	Low Track				High Track			
	Differences in				Differences in			
Model Comparison	χ^2	df	χ^2/df	delta	χ^2	df	χ^2/df	delta
Tests of Added Components								
Model 1 vs Model 2	1464.79	18	20.96	--	4837.92	18	69.65	--
Model 2 vs Model 3	101.57	12	.96	.04	340.09	12	4.98	.06
Model 3 vs Model 4	9.48*	3	0.00	.02	116.72	3	2.77	.02
Test of Convergent Validity								
Model 4 vs Model 5 (traits)	762.88	18	13.83	.45	2928.77	18	55.43	.53
Tests of Discriminant Validity								
Model 4 vs Model 6 (traits)	298.40	6	7.16	.18	1298.23	6	32.42	.24
Model 4 vs Model 2 (methods)	111.05	15	1.32	.07	456.81	15	7.75	.08
Tests of Method Bias								
Model 4 vs Model 7 (Likert)	48.93	6	.76	.03	124.11	7	2.11	.02
Model 4 vs Model 8 (semantic differential)	5.52[b]	6	.36	.00	152.46	7	2.82	.03
Model 4 vs Model 9 (Guttman)	27.79	6	.22	.02	277.14	7	5.94	.05

* $p<.05$

[a] unasterisked χ^2 difference values are statistically significant at $p<.001$

[b] not statistically significant

Table 5

Factor and Error/Uniqueness Loadings, and Factor Correlations for Baseline Model-Low Track[a]

Measure	Trait 1	Trait 2	Trait 3	Trait 4	Method I	Method II	Method III	Error/ Uniqueness
Likert								
general SC	.89*(.05)	.0	.0	.0	.07 (.07)	.0	.0	.20*(.05)
academic SC	.0	.73*(.06)	.0	.0	.31*(.11)	.0	.0	.37*(.07)
English SC	.0	.0	.78*(.07)	.0	.41*(.16)	.0	.0	.22 (.16)
mathematics SC	.0	.0	.0	.87*(.05)	.08 (.06)	.0	.0	.24*(.03)
Semantic Differential								
general SC	.67*(.06)	.0	.0	.0	.0	.46*(.16)	.0	.32*(.15)
academic SC	.0	.77*(.06)	.0	.0	.0	.43*(.15)	.0	.21 (.11)
English SC	.0	.0	.78*(.06)	.0	.0	.12 (.07)	.0	.37*(.06)
mathematics SC	.0	.0	.0	.88*(.05)	.0	.05 (.05)	.0	.21*(.03)
Guttman								
general SC	.84*(.06)	.0	.0	.0	.0	.0	.01 (.05)	.30*(.05)
academic SC	.0	.65*(.06)	.0	.0	.0	.0	.73*(.13)	.04 (.17)
English SC	.0	.0	.63*(.06)	.0	.0	.0	.27*(.07)	.53*(.06)
mathematics SC	.0	.0	.0	.84*(.05)	.0	.0	.24*(.06)	.25*(.04)

Factor Correlations

Trait 1	1.0						
Trait 2	.59 (.06)	1.0					
Trait 3	.33*(.07)	.72*(.05)	1.0				
Trait 4	.34*(.06)	.52*(.06)	.08 (.07)	1.0			
Method I	.0	.0	.0	.0	1.0		
Method II	.0	.0	.0	.0	.11 (.17)	1.0	
Method III	.0	.0	.0	.0	.39*(.13)	.03 (.12)	1.0

[a] All values of 1.0 and .0 are fixed values. All parameter estimates differing significantly from zero are asterisked. Parenthesized values are standard errors of associated parameters. SC = self-concept

Table 6

Factor and Error/Uniqueness Loadings, and Factor Correlations for Baseline Model-High Track

Measure	Trait 1	Trait 2	Trait 3	Trait 4	Method I	Method II	Method III	Error/ Uniqueness
Likert								
general SC	.88*(.04)	.0	.0	.0	.19*(.05)	.0	.0	.18*(.02)
academic SC	.0	.29*(.07)	.0	.0	.76*(.04)	.0	.0	.33*(.03)
English SC	.0	.0	.66*(.04)	.0	.46*(.04)	.0	.0	.39*(.03)
mathematics SC	.0	.0	.0	.78*(.03)	.51*(.04)	.0	.0	.07*(.01)
Semantic Differential								
general SC	.71*(.04)	.0	.0	.0	.0	.27*(.05)	.0	.40*(.03)
academic SC	.0	.83*(.10)	.0	.0	.0	.54*(.08)	.0	.01 (.12)
English SC	.0	.0	.82*(.04)	.0	.0	.53*(.05)	.0	.08*(.03)
mathematics SC	.0	.0	.0	.72*(.03)	.0	.59*(.04)	.0	.07*(.02)
Guttman								
general SC	.86*(.04)	.0	.0	.0	.0	.0	.24*(.05)	.20*(.02)
academic SC	.0	.14(.07)	.0	.0	.0	.0	.97*(.04)	.04 (.03)
English SC	.0	.0	.62*(.03)	.0	.0	.0	.59*(.04)	.33*(.03)
mathematics SC	.0	.0	.0	.68*(.03)	.0	.0	.55*(.04)	.17*(.01)

Factor Correlations

Trait 1	1.0						
Trait 2	.63*(.07)	1.0					
Trait 3	.11*(.06)	.20*(.07)	1.0				
Trait 4	.10 (.06)	.08 (.08)	-.46*(.05)	1.0			
Method I	.0	.0	.0	.0	1.0		
Method II	.0	.0	.0	.0	.89*(.02)	1.0	
Method III	.0	.0	.0	.0	.86*(.02)	.78*(.03)	1.0

[a] All values of 1.0 and .0 are fixed values. All parameter estimates differing significantly from zero are asterisked. Parenthesized values are standard errors of associated parameters. SC = self-concept

Table 7

Simultaneous Tests for the Invariance of Trait and Method Factor Loadings Across Track

Competing Models	χ^2	df	$\Delta\chi^2$	Δdf
Traits				
1. Four SC factors invariant[a]	311.41	82	--	--
2. Model 1 with all SC loadings invariant	550.23	94	239.09***	12
3. Model 1 with all general SC loadings invariant	312.23	85	1.09	3
4. Model 1 with all general and academic SC loadings invariant	456.67	88	145.53***	6
5. Model 1 with all general and English SC loadings invariant	318.13	88	6.99	6
6. Model 1 with all general, English, and mathematics SC invariant	334.68	91	23.54**	9
7. Model 5 with SDQASC invariant	401.89	89	83.76***	1
8. Model 5 with APIASC invariant	318.19	89	.06	1
9. Model 8 with SCAASC invariant	462.32	90	144.19***	2
10. Model 8 with SDQMSC invariant	321.67	90	3.54	2

(table continues)

Model	χ^2	df	$\Delta\chi^2$	Δdf
11. Model 10 with APIMSC invariant	332.82	91	14.69**	3
12. Model 10 with SCAASC invariant	321.79	91	3.66	3
Methods				
1. Model 12 with Likert method factor invariant	588.94	93	267.15***	2
2. Model 12 with semantic differential factor invariant	468.17	93	146.38***	2
3. Model 12 with Guttman factor invariant	426.14	94	104.35***	3

** $p < .01$

*** $p < .001$

[a] Baseline models with nonsignificant parameters fixed to 0.0

SC = self-concept; SDQASC = Self Description Questionnaire III (SDQIII) Academic SC subscale; APIASC = Affective Perception Inventory (API) Student Self subscale; SCAASC = Self-concept of Ability Scale (SCAS) Form A; SDQMSC = SDQ III Mathematics SC subscale; APIMSC = API Mathematics Perceptions subscale; SCAMSC = SCAS Form C

Table 8

Tests for the Invariance of Trait Correlations

Competing Models	χ^2	df	$\Delta\chi^2$	Δdf
Traits				
1. Invariant measurement model[a]	321.79	91	--	--
2. Model 1 with all trait correlations invariant	489.60	95	167.81***	4
3. Model 1 with trait correlations made independently invariant				
a) general/academic SC	321.85	92	.06	1
b) general/English SC	339.84	92	18.05***	1
c) general/mathematics SC	344.77	92	22.98***	1
d) academic/English SC	397.28	92	75.49***	1
e) academic/mathematics SC	393.69	92	71.90***	1
f) English/mathematics SC	359.06	91	--	--

*** $p < .001$

[a] Model 12 in Table 7

SC = self-concept

Figure Caption

Figure 1. Multitrait-multimethod Model of Data

M = method

T = trait

LIK = Likert scale

SD = semantic differential scale

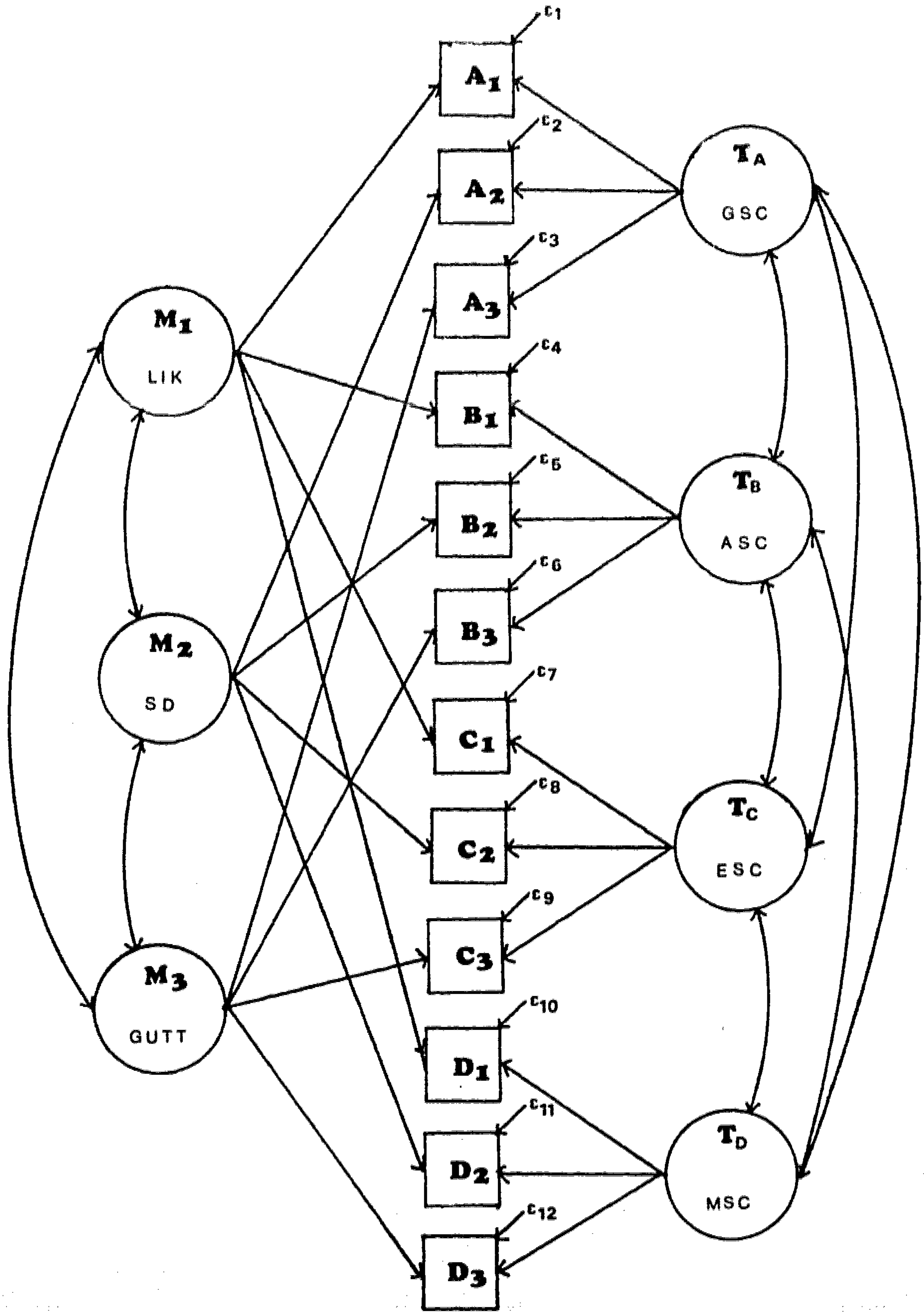
GUTT = Guttman scale

GSC = general self-concept

ASC = academic self-concept

ESC = English self-concept

MSC = mathematics self-concept



M1
LIK
M2
SD
M3
GUTT
A1
A2
A3
B1
B2
B3
C1
C2
C3
D1
D2
D3
e1
e2
e3
e4
e5
e6
e7
e8
e9
e10
e11
e12
TA
GSC
TB
ASC
TC
ESC
TD
MSC