

# An Application of Classical CSEM Methods in a Corrections Academy Selection Test

Kasey R. Stevens<sup>1</sup>

Corrections Standards Authority

Lawrence S. Meyers

California State University, Sacramento

*In G. M. Hurtz (Chair), Integrating Conditional Standard Errors of Measurement into Personnel Selection Practices. Symposium conducted at the 23<sup>rd</sup> annual conference of the Society for Industrial and Organizational Psychology, San Francisco, California.*

William G. Mollenkopf (1949) proposed a method using a quadratic function to estimate the conditional standard error of measurement (C-SEM) that replicates the nonlinear shape of the C-SEM as a function of total test score. Feldt, Steffen, and Gupta (1985) explored the possibility that a fourth order polynomial would better fit the expected C-SEM values and concluded that a third degree polynomial would be sufficient to describe the function. We refer to the modification of Mollenkopf's method using a third degree polynomial based on the research of Feldt et al. as the Mollenkopf-Feldt (M-F) method to estimate the C-SEM.

The M-F method represents a difference method in which the total test is divided into two half test scores and estimates of the C-SEM as the differences between the two halves are generated through polynomial regression. The model is as follows:

$$Y^2 = [(Half_1 - Half_2) - (GM_1 - GM_2)]^2 = b_1X + b_2X^2 + b_3X^3$$

where  $Y^2$  is the adjusted squared difference score,  $Half_1$  and  $Half_2$  are the scores for half-test 1 and half-test2, respectively, for each test taker,

$GM_1$  and  $GM_2$  are the overall means for half-test 1 and half-test 2, respectively, and  $X$  is the total test score; the intercept is set to zero. The polynomial regression model that is built has the following general form: at each value of  $X$  (the total test score),  $b_1$ ,  $b_2$ , and  $b_3$  are the raw score regression weights for the linear ( $X$ ), quadratic ( $X^2$ ), and cubic ( $X^3$ ) representations, respectively, of the total test score.

This study provides an application of the M-F method focusing on a written selection exam for three entry-level adult and juvenile correctional positions with the minimum educational requirements of a high school education. Basic reading comprehension, grammar, and math skills are assessed by three sets of 30 items. Applicants must achieve a minimum score of 22 on each section to pass the exam as a whole.

## Test Takers

The data were based on the January to October 2007 administrations of the written selection exam for the three entry-level positions. During this time frame 36,825 applicants took the written selection exam. The ethnicity of the applicants was 38.0% ( $n = 13,991$ ) Hispanic American, 29.3% ( $n = 10,777$ ) European American, 15.6% ( $n = 5,727$ ) African American, and 7.9% ( $n = 2,910$ ) Asian American, with 8.5% ( $n = 3,133$ ) indicating their ethnicity as other or

---

<sup>1</sup> This paper is a portion of a thesis (in preparation) by Kasey Stevens to fulfill the requirements for the Master's degree in Psychology with a concentration in I/O Psychology at California State University, Sacramento.

not at all. The applicants were 57.9% male ( $n = 20,966$ ) and 38.2% female ( $n = 14,067$ ), with 4.7% ( $n = 1,735$ ) not indicating their sex.

### Test Splitting

Each test section was split into two halves utilizing the following iterative ABBA sequence: First the test items were sorted based on their difficulty (mean performance) from the most difficult to least difficult. For items ranked 1 and 2, the more difficult item was assigned to half-test 1. For items 3 and 4, the more difficult item was assigned to half-test 2. For items 5 and 6, the more difficult item was assigned to half-test 2. For items 7 and 8, the more difficult item was assigned to half-test 1. This sequence was continued until all items were assigned to test halves. Finally, the score for each test half was calculated.

The efficacy of each split was evaluated to determine if the two test halves were *tau equivalent* (Feldt et al., 1985), that is, if the two halves had comparable means and variances. The procedure used to determine the equivalence of the halves was influenced by the large sample size for each split. Each split had a sample size consisting of at least 3,000 cases. This large sample size provides too much statistical power for statistical significance testing to be used “*carte blanche*.” Instead, the judgment of tau equivalence was based on a combination of effect size and a  $t$  test.

To assess the magnitude of the differences between the means of the two test halves, Cohen’s  $d$  statistic was computed as the mean difference between the two halves divided by the average standard deviation. According to Cohen an effect size of .2 or less is considered small.

To assess the magnitude of the differences between the variability of the two test halves, a  $t$  test was used. A formula for comparing the standard deviations of correlated (paired)

distributions is provided by Guilford and Fruchter (1978, p. 167):

$$t = \frac{(SD_1 - SD_2)\sqrt{N-2}}{2(SD_1)(SD_2)\sqrt{1-r^2}}$$

In the above formula,  $SD_1$  is the standard deviation of test half 1,  $SD_2$  is the standard deviation of test half 2,  $N$  is the number of test takers with both half test scores, and  $r$  is the Pearson correlation between the two half test scores, with  $t$  evaluated at  $N-2$  degrees of freedom.

Tau equivalency was said to be achieved when the mean difference effect size ratio was less than .2 and the probability of  $t$  evaluating the variance was greater than .10. Occasionally, tau equivalence was not achieved with the first split. When this occurred minor changes were made to the split halves in an iterative manner until it was achieved. For example, the math section contained one item with a mean difficulty value around .68 with the remaining 29 items having mean difficulties ranging from .83 to .98. To overcome the imbalance of one difficult item on only one test half, an item in the .8s on the same test was swapped for an easier item in the .9s on the other half. Thus two moderately difficult items on a test half was required to compensate for one difficult item on the test half. This swapping of items occurred only when tau equivalence was not achieved with the original split.

The process for splitting the test into two halves and determining tau equivalence was conducted separately for the three test sections for all applicants and separately for the ethnic groups of African American, Asian, European American and Hispanic American within each test section. For males and females, the mean item difficulties within each test section were similar to the mean item difficulties within each section for the full sample split. For this reason the split developed for all applicants within each

test section was used to create the split halves for males and females.

The results of the analyses conducted to determine tau equivalence for each split are provided in Table 1. As may be seen from Table 1, the mean difference effect size ratios for the splits ranged from .01 to .14 and the  $t$  values evaluating the variances ranged from .01 to 1.59 with probabilities greater than .10. Correlations between the split halves ranged from .45 to .67. The Spearman-Brown prophesied reliability with the length doubled range from .62 to .81.

### Supplementary Test Analyses

Supplementary analyses of the test indicated that there was no adverse impact using the 80% rule of thumb. Further, logistic regression analyses of the individual items supported the proposition that there was no DIF. It is therefore possible that examining estimated C-SEM could provide test developers with another tool to investigate group differences.

## Results and Discussion

### *M-F Method for Estimating C-SEM*

The regression coefficients for the full sample, ethnic groups, and sex groups for the reading, writing, and math sections are provided in Table 2. Polynomial regression resulted in significant linear and quadratic regression coefficients. The only exception is that the reading quadratic coefficient for African Americans was not significant. The cubic component was not significant in any of the three test sections.

### *Graphs of Regression Results*

The polynomial regression estimates of the C-SEM for each test section by ethnic group and sex group resulted in bow shaped curves with a strong quadratic component (Figures 1 through 6). The shape of these functions is as expected based on curves shown by Nelson (2007) which

were derived from data provided by Feldt (1984, Table 1) and Qualls-Payne (1992, Table 3).

It is important to note that there are horizontal tails drawn in the higher score range for the reading test (full sample, Asian, male, and female), writing test (African American), and math test (full sample, African American, male, and female). When calculating the predicted C-SEM for these functions, the value under the radical became negative. This situation was arbitrarily remedied by assigning the smallest C-SEM a constant value. For example, the horizontal tail for the full sample reading function begins at the test score of 28. For the reading test score of 28 the predicted C-SEM is .785. Calculating the predicted C-SEM for the reading test score of 29 and 30 is impossible since due to negative values under the radical. Since the C-SEM is unlikely to be zero, the C-SEM for the test score of 29 and 30 was set to .785; the smallest C-SEM before the value under the radical became negative.

It is interesting that the M-F polynomial functions occasionally produced negative values. This may highlight a potential shortcoming with the various difference methods for calculating C-SEM. However, this issue was not a concern with this exam since the cut point was made at the test score of 22 and not in the higher score range. It is possible that the lack of variance in the higher score range contributed to this phenomenon since these particular tests would be considered to be relatively "easy"; thus, test takers at the highest ability levels as estimated by total test score show very little variability and therefore little uncertainty regarding their estimated true test scores. It would be interesting to determine if this phenomenon appears using IRT methods for calculating C-SEM or if this is obtained with more difficult exams that have greater variance in the higher score range.

### Comparison of C-SEM Functions by Test Section, Ethnicity and Sex

Analyses were conducted to determine if the estimated C-SEM varied by ethnicity (Figures 1, 3, and 5) and sex (Figures 2, 4, and 6) for each test section. For reading, the functions were similar to the full sample function for all ethnic groups through the midrange but differed between the midrange and the highest score values in that the functions bowed less sharply and peaked later in the upper score range for Hispanic, European, and African Americans. The estimated C-SEM at the cut point ( $X = 22$ ) ranged from 1.89 to 2.50 for the different ethnic groups. There were minor differences in the estimated reading C-SEM by sex, with estimated C-SEMs at the cut point of 1.72 for males and 1.85 for females. The function peaked earlier and had a steeper bow for males compared to females and the full sample function. Only the Asian, male and female functions had a horizontal tail similar to the full sample function.

The writing test produced results very similar to reading except that (a) the Hispanic and European American functions were very similar to the overall full sample function across the entire score range exhibiting a relatively smooth bow in contrast to the Hispanic and American reading function which bowed less and peaked much higher compared to the reading full sample function, (b) the Asian function peaked slightly later and bowed less past the midrange, and (c) the African American function peaked earlier, had a more intense bow in the higher score range, and a horizontal tail. Compared to the full sample function, in the higher score range the female function had a steeper bow while the male function bowed less.

For math, the full sample and African American functions were similar. Compared to the full sample function the Hispanic, Asian, and European functions bowed less in the higher score range. Only minor differences were

obtained between the math functions for males and females in the upper score ranges. Only the African American, male and female functions had the horizontal tails similar to the full sample function.

In summary, the main differences between the estimated CSEM for each test section by ethnic group and sex with the full sample functions are limited to the high end of the score range. This might have been obtained because the exam is a rather easy. The average item difficulty is around .85 and only about 11% of the applicants fail the exam. Because the test is easy, many applicants obtain perfect or near perfect scores thus reducing the information (variance) of scores in the upper range.

Overall, the M-F method does a good job of estimating the C-SEM without requiring special software and a high level of statistical savvy. While the method does have some drawbacks (occasional horizontal tails or negative values if one does not compensate), it is a useful tool for test development in order to incorporate the use of C-SEM in line with professional testing Standards 2.2 and 2.14 (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

### References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44*, 883-891.

- Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, 9, 351-361.
- Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed). New York: McGraw-Hill.
- Mellenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika*, 14, 189-229.
- Nelson, L. R. (2007). Some issues related to the use of cut scores. *Thai Journal of Educational Research and Measurement*, 5 (in press).
- Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement*, 29, 213-225.

Table I

*Evaluation of split half tau equivalence for each test section by full sample, ethnicity and sex*

Group Analysis	Mean	t	Mean		Std. Deviation		r	n
	Ratio		half 1	half 2	half 1	half 2		
<i>Reading Section</i>								
Full Sample	.07	.15*	13.33	13.45	1.68	1.67	.626**	33,398
European American	.01	1.40*	13.87	13.86	1.33	1.36	.507**	10,777
Asian	.01	.00*	12.96	12.93	1.99	1.99	.675**	2,909
Hispanic	.00	.36*	13.15	13.15	1.78	1.79	.596**	13,988
African American	.14	1.59*	13.19	13.43	1.63	1.71	.585**	5,726
Female	.08	.34*	13.33	13.47	1.63	1.64	.450**	13,346
Male	.03	.87*	13.42	13.36	1.69	1.72	.623**	20,029
<i>Writing Section</i>								
Full Sample	.07	.58*	13.49	13.35	1.75	1.73	.610**	33,355
European American	.01	.92*	13.76	13.79	1.43	1.45	.544**	10,776
Asian	.03	.81*	13.03	12.94	2.16	2.06	.689**	2,908
Hispanic	.07	1.11*	13.25	13.38	1.80	1.76	.622**	13,985
African American	.02	.03*	13.24	13.20	1.85	1.85	.623**	5,724
Female	.10	.13*	13.60	13.42	1.69	1.69	.601**	13,330
Male	.06	.54*	13.42	13.31	1.78	1.76	.613**	20,001
<i>Math Section</i>								
Full Sample	.05	.57*	13.64	13.73	1.52	1.51	.567**	33,383
European American	.07	.72*	13.99	13.89	1.30	1.29	.489**	10,772
Asian	.05	.21*	13.80	13.88	1.40	1.41	.552**	2,908
Hispanic	.01	.23*	13.60	13.61	1.54	1.55	.578**	13,981
African American	.03	.36*	13.35	13.29	1.74	1.76	.618**	5,723
Female	.15	.01*	13.46	13.70	1.51	1.51	.539**	13,342
Male	.01	.17*	13.76	13.74	1.52	1.51	.588**	20,017

Note. n = sample size, r = Pearson correlation between split halves, \* p &gt; .10, \*\* p &lt; .01

Table 2

*C-SEM regression coefficients for each test section by full sample, ethnicity, and sex*

Group Analysis	$b_1$	$b_2$	$b_3$	$n$	$R$
<i>Reading Section</i>					
Full Sample	.778**	-.027**	.000	33,396	.603
European American	.864**	-.027**	.000	10,776	.622
Asian	.824**	-.030*	.000	2,909	.572
Hispanic	.737**	-.020**	.000	13,987	.601
African American	.717**	-.019**	.000	5,726	.617
Female	.816**	-.030**	.000	13,344	.602
Male	.948**	-.037**	.000	20,029	.604
<i>Writing Section</i>					
Full Sample	.890**	-.028**	.000	33,354	.603
European American	.864**	-.027**	.000	10,775	.602
Asian	.705**	-.019**	.000	2,908	.589
Hispanic	.861**	-.028**	.000	13,984	.597
African American	.967**	-.035**	.000	5,724	.597
Female	.931**	-.031**	.000	13,329	.611
Male	.865**	-.026**	.000	20,001	.598
<i>Math Section</i>					
Full Sample	.964**	-.036**	.000	33,379	.608
European American	.862**	-.026**	.000	10,771	.614
Asian	.705**	-.019	.000	2,908	.602
Hispanic	.773**	-.024**	.000	13,981	.602
African American	.914**	-.035**	.000	5,723	.601
Female	.941**	-.035**	.000	13,341	.611
Male	.937**	-.035**	.000	20,017	.609

Note.  $b_1$  = linear raw score regression weight,  $b_2$  = quadratic raw score regression weight,  $b_3$  = cubic raw score regression weight,  $n$  = sample size,  $r$  = Pearson correlation between split halves, \*\*  $p < .01$ , \*  $p < .05$ .

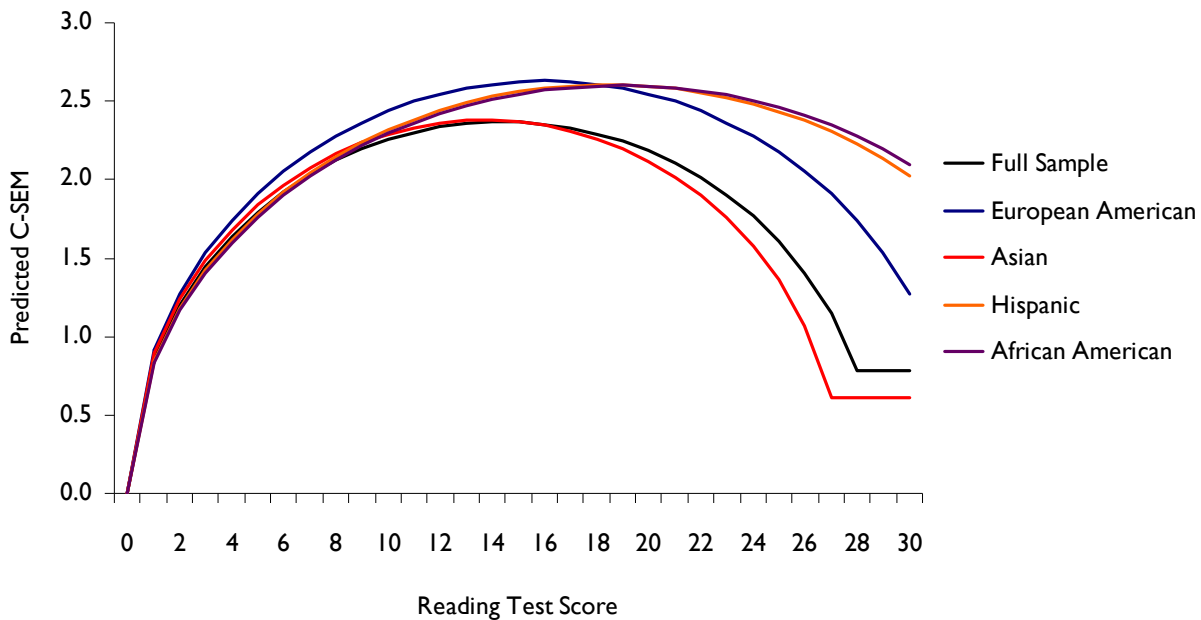


Figure 1. Predicted reading C-SEM by ethnicity and full sample

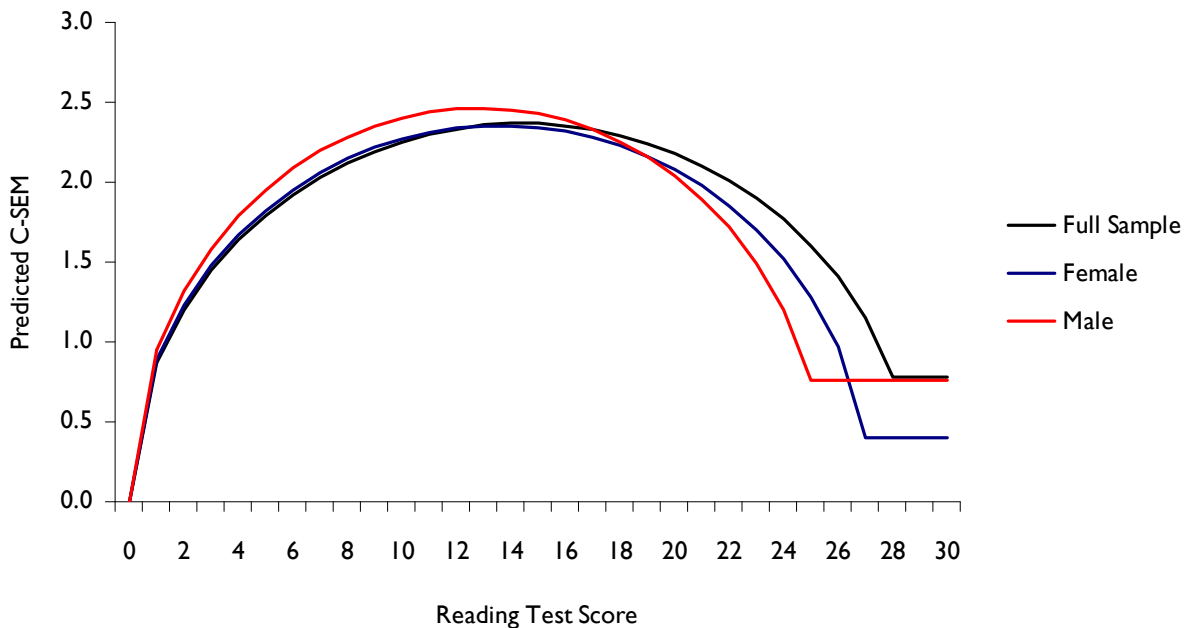


Figure 2. Predicted reading C-SEM by sex and full sample

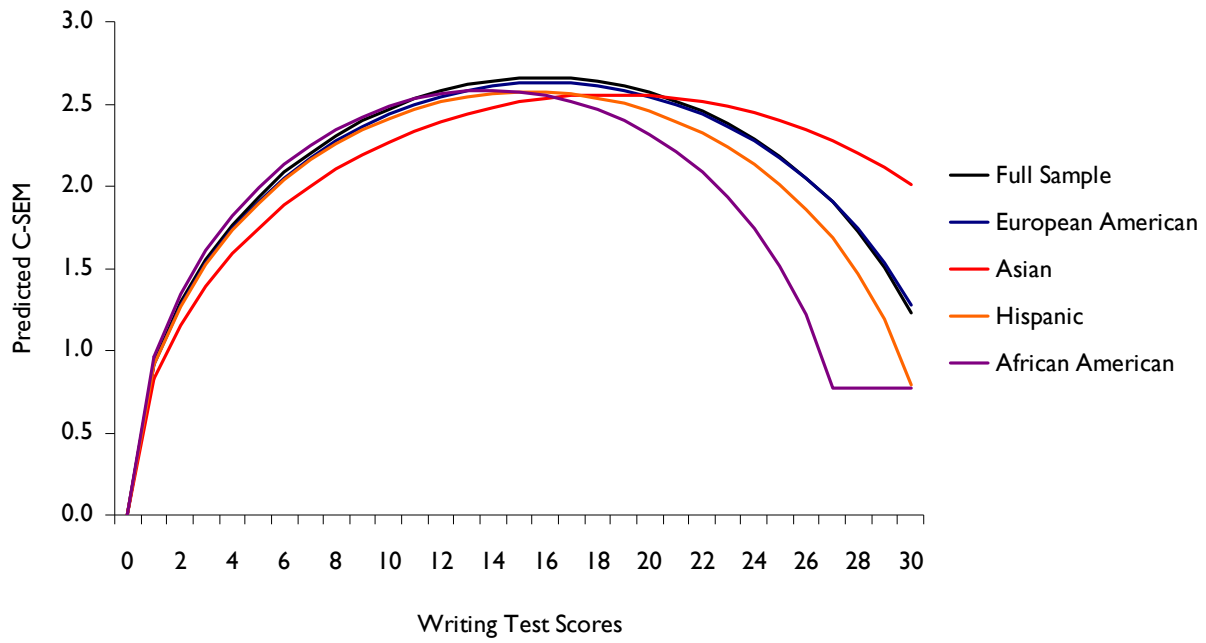


Figure 3. Predicted writing C-SEM by ethnicity and full sample

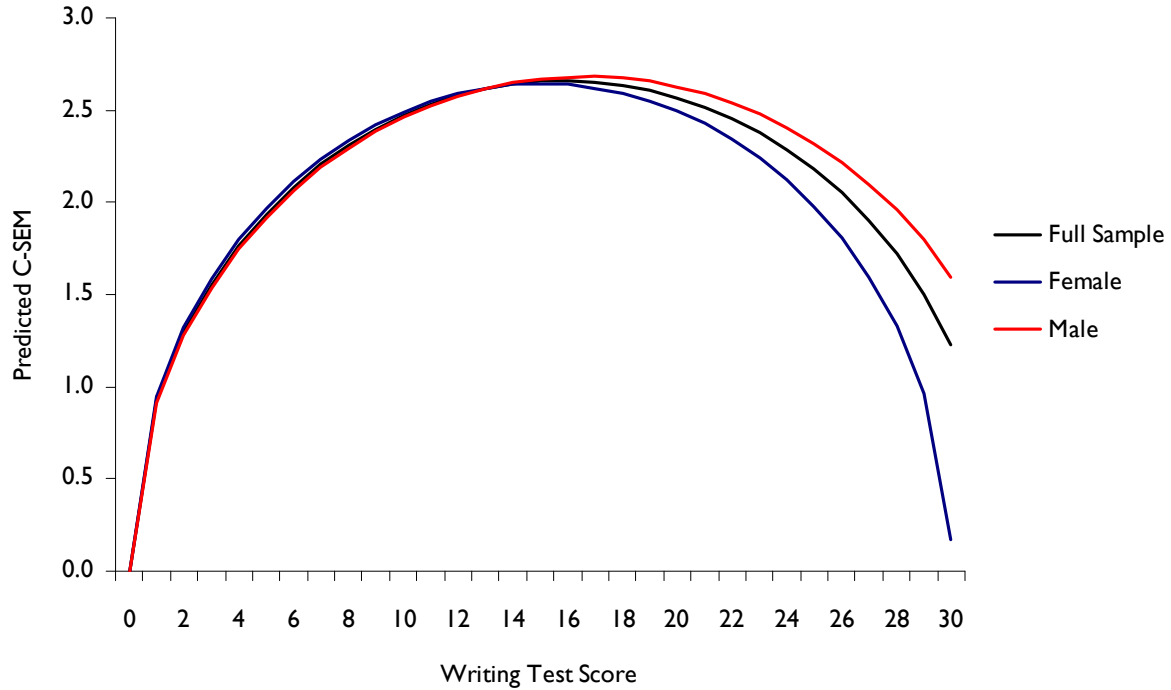


Figure 4. Predicted writing C-SEM by sex and full sample

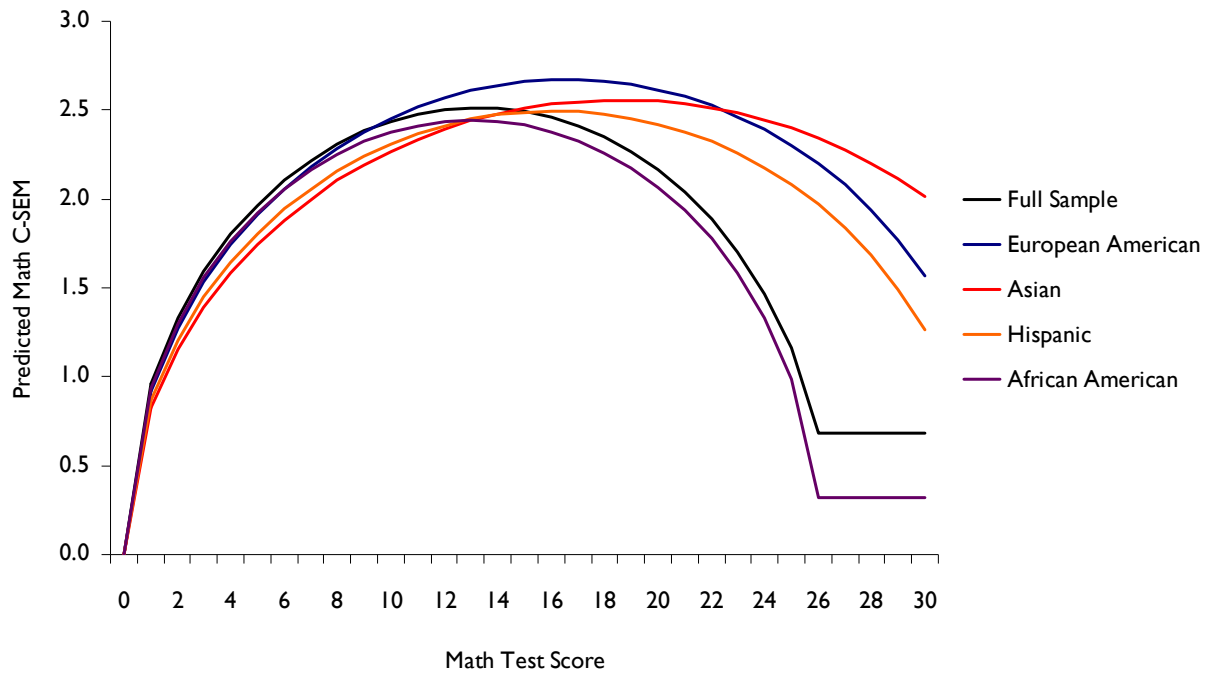


Figure 5. Predicted math C-SEM by ethnicity and full sample

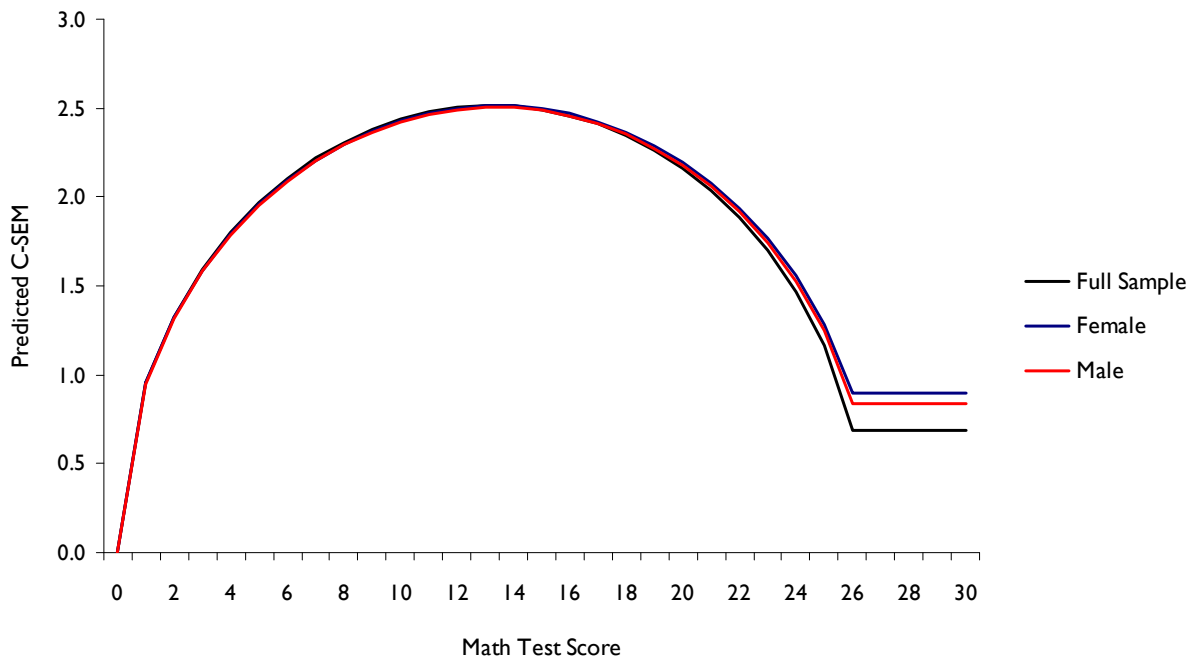


Figure 6. Predicted math C-SEM by sex and full sample