ACT Research Report Series 96-4

An Empirical Investigation of the Accuracy of a Step–Up Method for Estimating Test Score Conditional Variances

Imelda C. Go

David J. Woodruff

ACT

June 1996

For additional copies write;
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

An Empirical Investigation of the Accuracy of a Step-up Method for Estimating Test Score Conditional Variances

Imelda C. Go

David J. Woodruff

Abstract

In previous works, Woodruff derived expressions for three different conditional test score variances: the conditional standard error of prediction (CSEP), the conditional standard error of measurement in prediction (CSEMP), and the conditional standard error of estimation (CSEE). He also presented step-up formulas that require only weak assumptions and that allow the estimation of full-length test score conditional variances from two parallel half-length tests. This study empirically investigates the accuracy of the step-up formulas using real test data and concludes that the step-up formulas work fairly well for the CSEP and the CSEMP but less well for the CSEE. The CSEMP is also compared with two other procedures for estimating the conditional standard error of measurement (CSEM).

An Emprical Evaluation of the Accuracy of a Step-up Method for Estimating Test Score Conditional Variances

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) list as a secondary standard (one that is desirable but often not feasible) the recommendation that, conditional on critical score values, the standard error of measurement (SEM) be computed and reported. Under the classical test theory model, $X = T + E$, the conditional standard error of measurement (CSEM) is defined as the conditional observed score (or error score) variance for a fixed value of true score, that is, $\sigma^2(X|T=t) = \sigma^2(E|T=t)$. In practice, true scores are usually not known so methods have been developed that estimate $\sigma^2(E|X=x)$ in place of $\sigma^2(E|T=t)$. However, when the conditioning is on X rather than T, it can be shown that $\sigma^2(E|X) = \sigma^2(T|X) = -\sigma(T, E|X)$ for all values of X. Hence, $\sigma^2(E|X)$ is artificially constrained in a way that $\sigma^2(E|T)$ is not. Also, Woodruff (1990) shows that if the reliability of X is less than one, then $\mu[\sigma^2(E|X)] < \mu[\sigma^2(E|T)]$, where μ denotes expectation. Hence, on average, $\sigma^2(E|X)$ is underestimating $\sigma^2(E|T)$. Such considerations led Woodruff (1990, 1991) to develop an alternative method for estimating conditional test score variances. The purpose of this paper is to empirically evaluate the accuracy of this alternative procedure and to compare the alternative procedure with other procedures.

The Procedures

Consider two classically parallel full-length tests, $X1 = T_X + E_{X1}$ with m_{X1} items and $X2 = T_X + E_{X2}$ with m_{X2} items, both of which are administered to N examinees. It is shown in appendix A that it is reasonable to assume that $\sigma(T_X, E_{X2}|X1) = 0$ so that the following decomposition holds:

$$\sigma^2(X2|X1) = \sigma^2(T_X|X1) + \sigma^2(E_{X2}|X1). \qquad (1.)$$

Woodruff (1990, 1991) calls $\sigma^2(X2 \mid X1)$ the squared conditional standard error of prediction (CSEP), $\sigma^2(T_X \mid X1)$ the squared conditional standard error of estimation (CSEE), and $\sigma^2(E_{X2} \mid X1)$ the squared conditional standard error of measurement in prediction (CSEMP). All three of these conditional variances offer information about the accuracy of test scores at specific locations on the score scale, but it is the CSEMP that is most closely related to the CSEM. In Appendix A, it is shown that the average value of the CSEMP equals the average value of the CSEM, and this strongly supports the recommendation that the CSEMP be used as a substitute for the CSEM. Another advantage of using the CSEMP is that the CSEMP requires only the relatively weak assumptions of classical test theory.

For each value of $X1 = 0, 1, 2, \ldots, m_{X1}$, let the item scores for $X2$ be analyzed as a two-way persons (P) by measures (M) ANOVA with one observation per cell. In these conditional ANOVA's, let $MS_P(X2 \mid X1)$ denote the main effect mean square for persons and let $MS_{PM}(X2 \mid X1)$ denote the persons by measures interaction mean square. Following Woodruff (1990, 1991) estimates for the three conditional variances are given by:

$$[\mathrm{CSEP}(X2 \mid X1)]^2 = s^2(X2 \mid X1) = m_{X2} MS_P(X2 \mid X1), \qquad (2.)$$

$$[\mathrm{CSEMP}(X2 \mid X1)]^2 = s^2(E_{X2} \mid X1) = m_{X2} MS_{PM}(X2 \mid X1), \text{ and} \qquad (3.)$$

$$\begin{aligned}[\mathrm{CSEE}(X2 \mid X1)]^2 &= s^2(T_X \mid X1) = s^2(X2 \mid X1) - s^2(\mathrm{E}_{X2} \mid X1) \\ &= m_{X2}[MS_P(X2 \mid X1) - MS_{PM}(X2 \mid X1)]. \qquad (4.)\end{aligned}$$

In practice, scores on two full-length tests are rarely available. However, if a single full-length test can be divided into two parallel half-length tests, then estimates for the full-length test conditional variances can be obtained from the half-length test conditional variances by using the step-up formulas derived by Woodruff (1990, 1991). Suppose that the full-length test, X, can be divided into two classically parallel half-length tests, $Y1 = T_Y + E_{Y1}$ with m_{Y1} items and $Y2 = T_Y + E_{Y2}$ with m_{Y2} items. Let the linear transformation

$$X^* = X^*(Y1) = aY1 + b \tag{5.}$$

rescale the half-length test, $Y1$, to have the same mean and variance as the full-length test, X. The stepped-up estimates are:

$$\{[\mathrm{CSEP}^*[X(Y2)|X^*(Y1)]\}^2 = \left[\frac{2(1+3r_{Y1Y2})}{\left(1+r_{Y1Y2}\right)^2}\right] m_{Y2} MS_P[(Y2 \mid X^*(Y1)], \tag{6.}$$

$$\{[\mathrm{CSEMP}^*[\mathrm{X}(Y2) \mid X^*(Y1)]]\}^2 = 2m_{Y2} MS_{PM}[Y2 \mid X^*(Y1)\ ], \text{ and} \tag{7.}$$

$$\{\mathrm{CSEE}^*[X(Y2) \mid X^*(Y1)]]\}^2 = m_{Y2}\left\{\left[\frac{2(1+3r_{Y1Y2})}{\left(1+r_{Y1Y2}\right)^2}\right] MS_P[Y2 \mid X^*(Y1)] - 2MS_{PM}[Y2 \mid X^*(Y1)]\right\}. \tag{8.}$$

In the preceeding three equations, the conditioning is on $X^*(Y1)$, the two mean squares are computed from a two-way ANOVA on the item scores for $Y2$, and the notation $X(Y2)$ denotes that these half-length test mean squares have been stepped-up to full-length test mean squares. Finally, $r_{Y1\,Y2}$ denotes the sample correlation between $Y1$ and $Y2$.

In what follows, reference will be made to stepped-up half-length test conditional standard deviations and to full-length test conditional standard deviations. The stepped-up half-length test conditional standard deviations, as given on the left side in equations (6.), (7.), and (8.), will always have asterisks as part of their name whereas the full-length test conditional standard deviations, as given on the left side in equations (2.), (3.), and (4.), will not.

There are at least two methods that estimate $\sigma^2(E_X | X)$ in place of $\sigma^2(E_X | T_X)$. The first of these methods is the difference method due to Thorndike (1951). This method divides a single full-length test, X, into two parallel half-length tests, $Y1$ and $Y2$, and then calculates

$$\text{T-CSEM}(E_X | X) = s(Y1 - Y2 | X) \qquad (9.)$$

as a substitute estimate for $\sigma(E_X | T_X)$. Woodruff (1990) critically discusses the basis for this method. Another such method is presented by Feldt, Steffen, & Gupta (1985). This method is based on an ANOVA of the item responses of X. It substitutes as an estimate for $\sigma(E_X | T_X)$ the following estimate

$$\text{F-CSEM}(E_X | X) = [m_X(MS_{PM} | X)]^{1/2} \qquad (10.)$$

where $(MS_{PM} | X)]$ is a conditional interaction mean square from a measures by persons ANOVA of the item responses of X given a fixed value of X.

The Empirical Investigation

The data for this study was a random sample of 40,000 examinees with scores on the October1986 ACT Assessment Program (American College Testing [ACT], 1987). The ACT Assessment Program (AAP) then consisted of 219 dichotomously-scored items from four subtest areas: 75 from English, 40 from Mathematics, 52 from Social Studies, and 52 from Natural Sciences. Though data from 219 items were available, the goal was to divide the items into four parallel groups of items so the first three English items were eliminated. The remaining 216 AAP items were treated as an item pool from which parallel tests and half-tests could be constructed. In particular, four 54-item half-length tests were created and these were combined to yield two 108-item full-length tests. The four half-length tests were denoted $Y1$, $Y2$, $Y3$, and $Y4$. The two half-length tests $Y1$ and $Y2$ were combined to yield the full-length test $X1$, and the two half-length tests $Y3$ and $Y4$ were combined to yield the full-length test $X2$.

All four half-length tests were carefully constructed to be balanced in content and to have similar test score statistics. The two full-length tests also were constructed to be balanced in content and to have similar test score statistics.

The first step in constructing the four parallel half-length tests and the two parallel full-length tests was to compute the correlations between item position and item difficulty within each one of the four AAP subtests. Because the items within these four AAP subtests were ordered by item difficulty, negative correlations of –.36, –.88, –.56, and –.62 were found for the English, Mathematics, Natural Sciences, and Social Sciences AAP subtests, respectively. As a consequence, a systematic selection of the subtest items in their original test order was used. Table 1 shows the systematic item selection scheme for half-length tests Y1, Y2, Y3, and Y4. For example, to construct test Y1, the 1st out of every 4 English items, the 4th out of every 4 Mathematics items, the 3rd out of every 4 Social Studies items, and the 2nd out of every 4 Natural Sciences items were used. As a result, each one of the four parallel half-length tests had 18 English items, 10 Mathematics items, 13 Social Studies items, and 13 Natural Sciences items; and each one of the two parallel full-length tests had 36 English items, 20 Mathematics items, 26 Social Studies items, and 26 Natural Sciences items. The full-length tests and half-length tests were not homogeneous in content, but they were parallel in content. This illustrates an advantage of the current method, namely, an assumption of unidimensionality is not required.

Tables 2 presents some relevant test score statistics for the two 108-item parallel full-length tests, X1 and X2, and the four 54-item parallel half-length tests: Y1, Y2, Y3, and Y4. The statistics in Table 2 indicate that the two full-

Table1. Systematic Item Sampling Scheme for Constructing Parallel Half-Length Tests.

	AAP Subtests			
Half-Length Test	English	Mathematics	Social Studies	Natural Sciences
Y1	1	4	3	2
Y2	2	3	4	1
Y3	3	2	1	4
Y4	4	1	2	3

Table 2. Test Score Statistics for the Full-Length and Half-Length Tests.

	Mean	SD	Correlations, KR20's, and Dissattenuated Correlations*					
Test			*X1*	*X2*	*Y1*	*Y2*	*Y3*	*Y4*
X1	62.2	16.8	0.93	0.93	—	—	—	—
X2	59.9	16.7	1.00	0.93	—	—	—	—
Y1	31.4	8.7	—	—	0.87	0.87	0.87	0.87
Y2	30.8	8.7	—	—	1.00	0.87	0.87	0.87
Y3	29.5	8.7	—	—	1.00	1.00	0.87	0.87
Y4	30.4	8.6	—	—	1.00	1.00	1.00	0.87

*Correlations are above the diagonal, KR20's are on the diagonal, and dissattenuated correlations are below the diagonal.

length tests have nearly identical test score statistics except for a modest difference between the means, and that should have little effect on the procedures under study. The same is true for the four half-length tests. Relevant correlations, KR20's, and relevant dissattenuated correlations (using the KR20's) are also presented in Table 2. These support the claim that the half-length tests and the full-length tests are indeed parallel.

The full-length test score scale of 108 items was divided into intervals that comprised three score points starting with a score of 1. These intervals had midpoints of 2, 5, 8, ..., 104, and 107. CSEP, CSEMP, and CSEE estimates using full-length tests X1 and X2 were computed for each of these intervals except for some intervals at the bottom and top of the score scale that did not have a sufficient number of examinees for stable estimation. However, the expected guessing score on a 108-item test is 27 and the ACT Assessment Program (ACT, 1987) is designed so that few examinees obtain nearly perfect scores. Hence, the score interval midpoints of 26 through 96, for which stable CSEP, CSEMP, and CSEE estimates were obtained, covers the length of the score scale that the AAP was designed to most effectively measure. Two sets of such estimates were obtained: one conditioning on X1 and the other conditioning on X2.

Next, two sets of stepped-up half-length test estimates of the CSEP, CSEMP, and CSEE were computed using the two pairs of half-length tests (pair 1: Y1 and Y2, pair 2: Y3 and Y4) and the same three-point wide test score intervals. These stepped-up half-length test estimates of the CSEP, CSEMP, and CSEE were then compared to the full-length test estimates of the CSEP, CSEMP, and CSEE computed directly from X1 and X2. In particular, the

CSEP*[X(Y2)|X*(Y1)] was compared to the CSEP(X2|X1) and the CSEP*[X(Y4)|X*(Y3)] was compared to the CSEP(X1|X2). Similar comparisons were made for the CSEMP and the CSEE.

Figure 1a in Appendix B is a graph of the stepped-up half-length test estimates: CSEP*[X(Y2)|X*(Y1)], CSEMP*[X(Y2)|X*(Y1)], and CSEE*[X(Y2)|X*(Y1)] along with the full-length test estimates: CSEP(X2|X1), CSEMP(X2|X1), and CSEE(X2|X1). Figure 1b in Appendix B is the same as Figure 1a except that quadratic polynomials were used to smooth the CSEP*, CSEE*, CSEP, and CSEE estimates. Figures 2a and 2b in Appendix B are analogous to Figures 1a and 1b except that they compare the CSEP*[X(Y4)|X*(Y3)], CSEMP*[X(Y4)|X*(Y3)], and CSEE*[X(Y4)|X*(Y3)] estimates with the CSEP(X1|X2), CSEMP(X1|X2), and CSEE(X1|X2) estimates.

The CSEMP estimates also were compared to the CSEM estimates computed by the difference method (Thorndike, 1951) and the ANOVA method (Feldt et. al., 1985) using the same intervals of three score points that were used to compute the CSEMP estimates. Figure 3a is a graph of the CSEMP*[X(Y2)|X*(Y1)], the F-CSEM(E|X1), the T-CSEM(E|X1) estimates. Figure 3b is the same as Figure 3a except that the T-CSEM(E|X1) estimates have been smoothed using a quadratic polynomial. Figures 4a and 4b are analogous to Figures 3a and 3b except that Figures 4a and 4b compare the CSEMP*[X(Y4)|X*(Y3)] estimates to the F-CSEM(E|X2) and T-CSEM(E|X2) estimates.

Discussion

The primary purpose of the present study was to evaluate the accuracy of the step-up procedure. How well the stepped-up half-length test estimates: CSEP*, CSEMP*, and CSEE*, approximate the full-length test estimates: CSEP, CSEMP, and CSEE, can be seen in Figures 1 and 2. These figures indicate that the step-up procedure works very well for the CSEMP, fairly well for the CSEP, and less well for the CSEE.

The secondary purpose of this paper was to compare the stepped-up half-length test estimate, CSEMP*, with the Feldt et. al (1985) and the Thorndike (1951) estimates of the CSEM, namely, F-CSEM and T-CSEM, respectively. Figures 3 and 4 show that the T-CSEM tends to be less than both the F-CSEM and the CSEMP*. Figures 3 and 4 also show that the F-CSEM tends to be less than the CSEMP* at both ends of the score scale but slightly greater than the CSEMP* in the middle of the score scale. These latter results agree with those found by Woodruff (1990). Because the average CSEMP equals the average CSEM, these results suggest that the T-CSEM is generally underestimating the CSEM, and that the F-CSEM may be slightly underestimating the CSEM at the ends of the score scale, but on average the F-CSEM appears closer to the CSEM than the T-CSEM.

Finally, all of the half-length and full-length test scores in the present study had unimodal approximately symmetrical distributions so the results reported here do not necessarily generalize to other types of test score distributions. However, Woodruff (1990) does report some limited results for skewed test score distributions, and those results are similar to the ones reported here.

References

American College Testing Program (1987). *ACT Assessment Program technical manual.* Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

DeGroot, M. H. (1989). *Probability and Statistics.* Reading, MA: Addison-Wesley

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**(4), 351-361.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational Measurement,* 560-620. Washington, DC: American Council on Education.

Woodruff, D. J. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement,* **27**(3), 191-208.

Woodruff, D. J. (1991). Stepping up test score conditional variances. *Journal of Educational Measurement,* **28**(3), 191-196.

Appendix A

Derivations

To show that $\sigma(T_X, E_{X2} \mid X1) = 0$ first recall that $X1 = T_X + E_{X1}$ and $X2 = T_X + E_{X2}$ are parallel measurements and that μ denotes expectation. The conditional covariance betweeen T_X and E_{X2} given $X1$ can be written as

$$\sigma(T_X, E_{X2} \mid X1) = \mu(T_X E_{X2} \mid X1) - \mu(T_X \mid X1)\mu(E_{X2} \mid X1). \qquad \text{(A1)}$$

Using the double expectation theorem (DeGroot, 1989, p 220) on the first term on the right hand side of (A1) gives

$$\begin{aligned} \sigma(T_X, E_{X2} \mid X1) &= \mu[\mu(T_X E_{X2} \mid X1) \mid T_X] - \mu(T_X \mid X1)\mu(E_{X2} \mid X1) \\ &= \mu[T_X \mu(E_{X2} \mid T_X, X1)] - \mu(T_X \mid X1)\mu(E_{X2} \mid X1). \end{aligned} \qquad \text{(A2)}$$

Making the assumption of linear experimental independence (Lord & Novick, 1968, p 45) between E_{X2} and $X1$ and between E_{X2} and $(X1, T_X)$ implies that

$$\mu(E_{X2} \mid X1) = 0 \text{ for all values of X1 and} \qquad \text{(A3)}$$

$$\mu(E_{X2} \mid X1, T_X) = 0 \text{ for all values of } (X1, T_X). \qquad \text{(A4)}$$

Substituting (A3) and (A4) into (A2) yields the desired result:

$$\sigma(T_X, E_{X2} \mid X1) = \mu(T_X 0) - \mu(T_X \mid X1)0 = 0.$$

To show that $\mu[\sigma(E_{X2} \mid X1)] = \mu[\sigma(E_{X2} \mid T_X)]$ note that by Theorem 2.6.2 of Lord & Novick (1968, p 35)

$$\sigma^2(E_{X2}) = \mu[\sigma^2(E_{X2} \mid X1)] + \sigma^2[\mu(E_{X2} \mid X1)] = \mu[\sigma^2(E_{X2} \mid T_X)] + \sigma^2[\mu(E_{X2} \mid T_X)].$$

It follows from the assumption that E_{X2} is linearly experimentally independent of both $X1$ and T_X that $\mu(E_{X2} \mid X1) = \mu(E_{X2} \mid T_X) = 0$. Hence, the above becomes

$$\sigma^2(E_{X2}) = \mu[\sigma^2(E_{X2} \mid X1)] = \mu[\sigma^2(E_{X2} \mid T_X)].$$

Appendix B

Figures

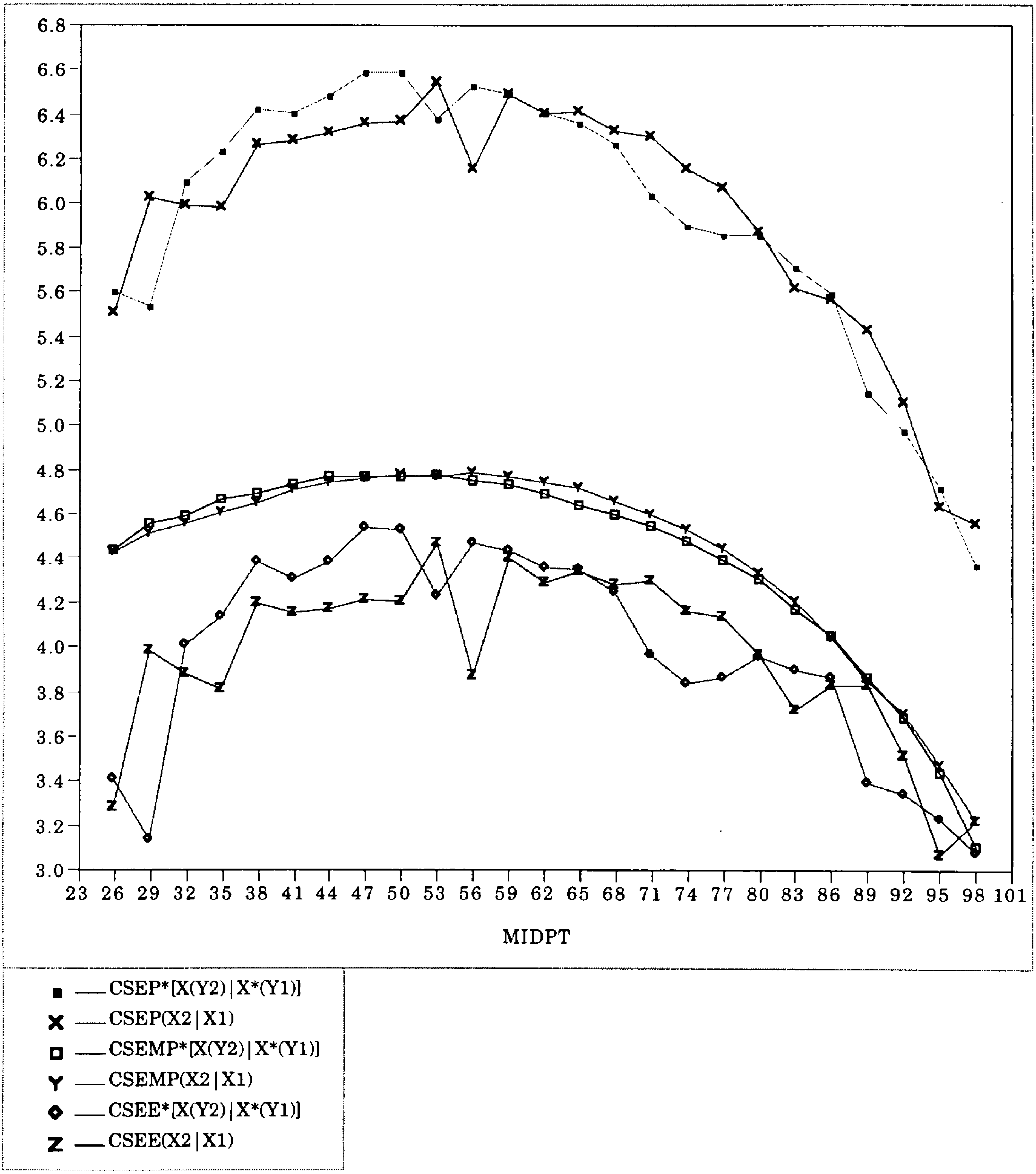

Figure 1a.

Plot of the full-length CSEP(X2|X1), CSEMP(X2|X1), and CSEE(X2|X1) against the stepped-up half-length CSEP*[X(Y2)|X*(Y1)], CSEMP*[X(Y2)|X*(Y1)], and CSEE*[(X(Y2)|X*(Y1)].

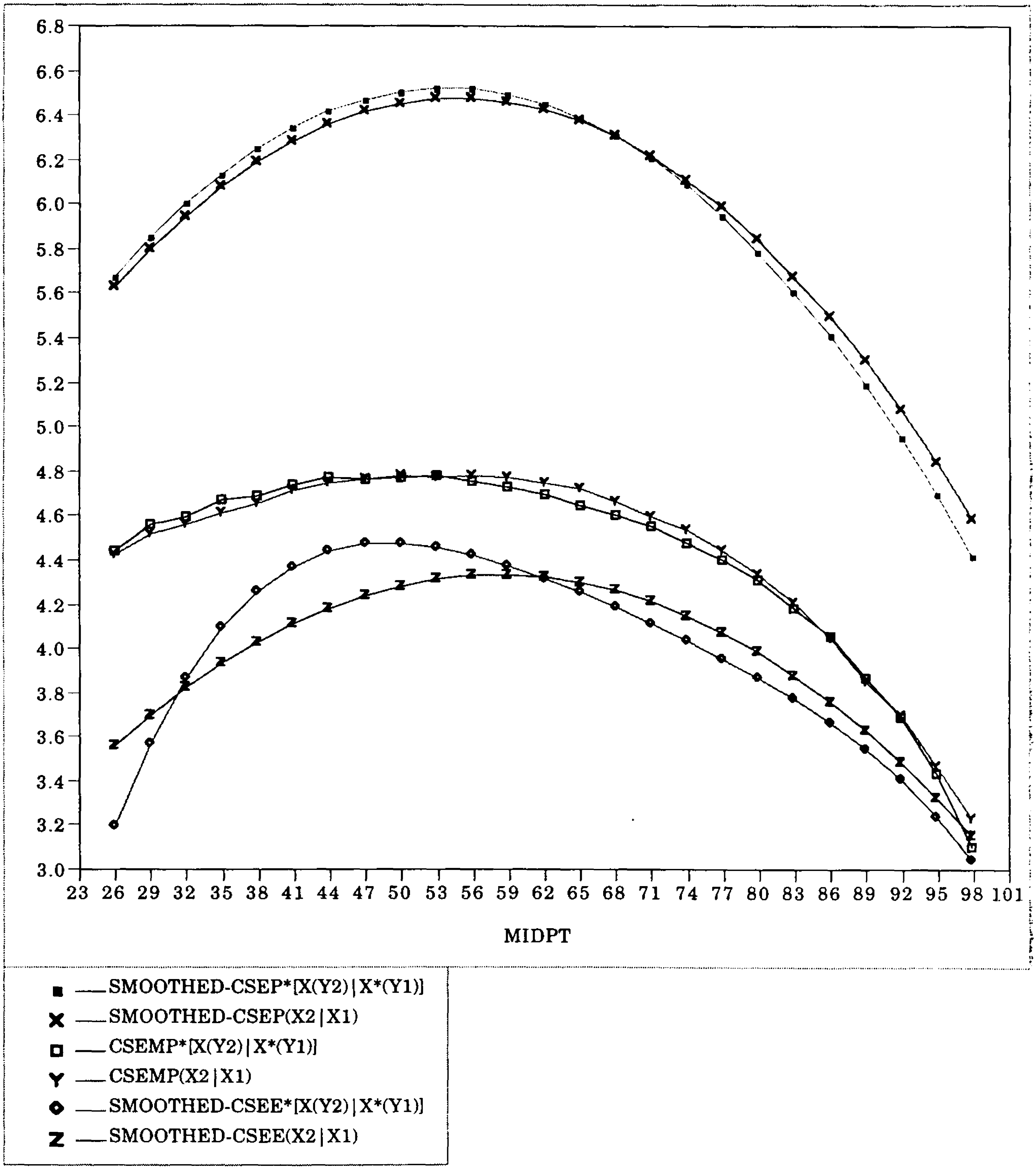

Figure 1b.

Plot of the full-length CSEP(X2 | X1), CSEMP(X2 | X1), and CSEE(X2 | X1) against the stepped-up half--length CSEP*[X(Y2) | X*(Y1)], CSEMP*[X(Y2) | X*(Y1)], and CSEE*[X(Y2) | X*(Y1)] with quadratic polynomial smoothing of the CSEP*, CSEE*, CSEP, and CSEE.

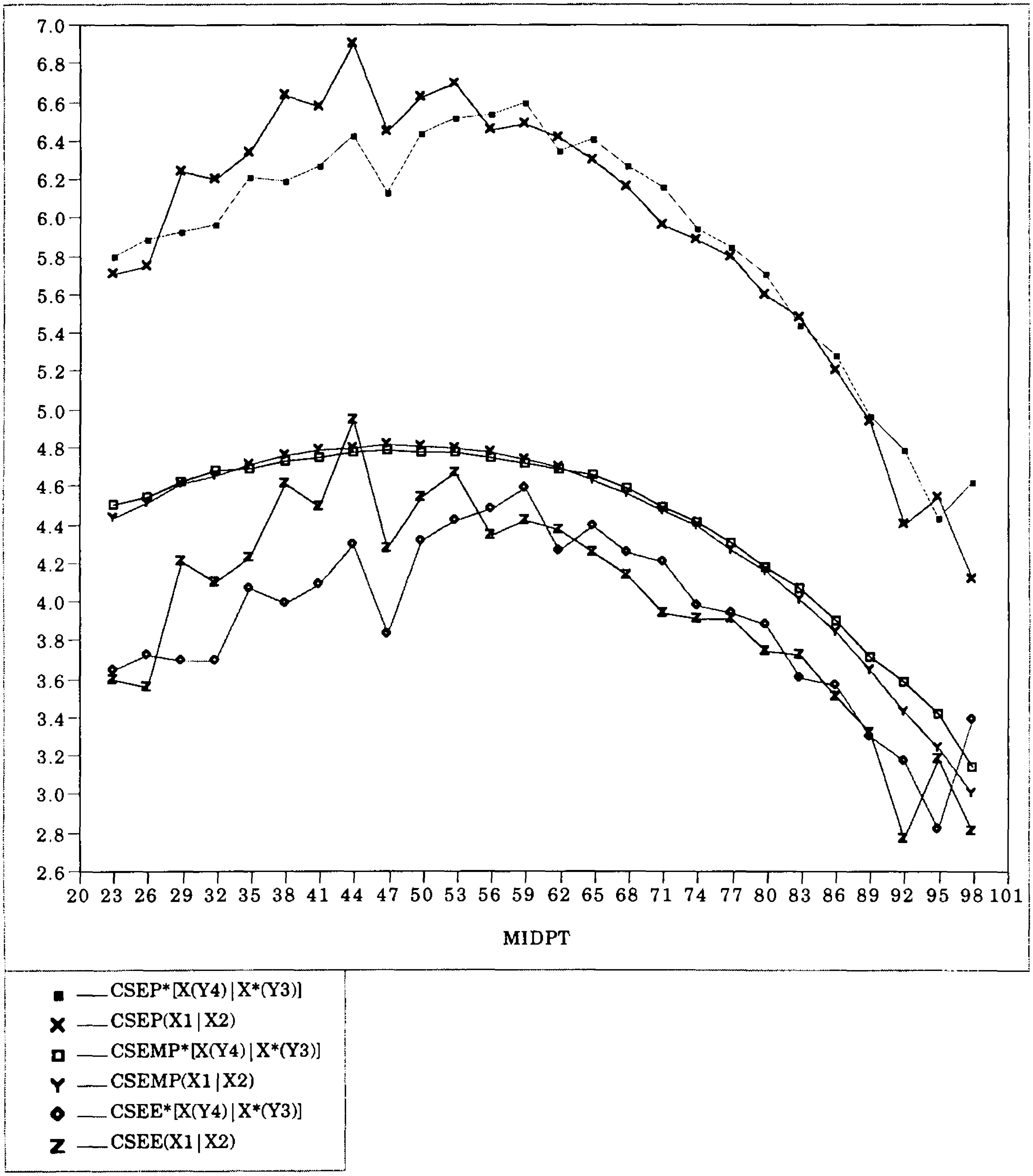

Figure 2a.

Plot of the full-length CSEP(X1|X2), CSEMP(X1|X2), and CSEE(X1|X2) against the stepped-up half-length CSEP*[X(Y4)|X*(Y3)], CSEMP*[X(Y4)|X*(Y3)], and CSEE*[X(Y4)|X*(Y3)].

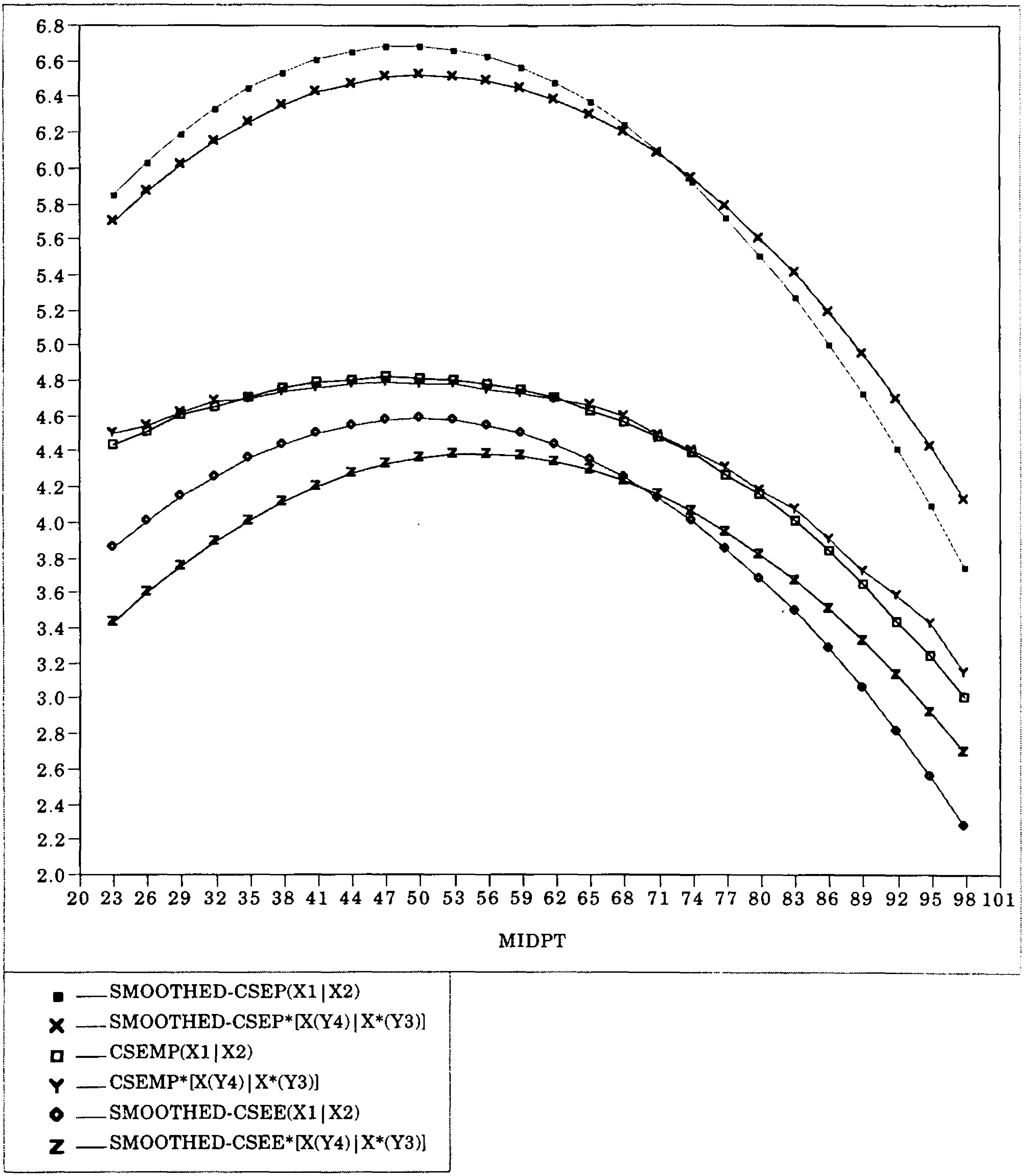

Figure 2b.

Plot of the full-length CSEP(X1|X2), CSEMP(X1|X2), and CSEE(X1|X2) against the stepped-up half-length CSEP*[X(Y4)|X*(Y3)], CSEMP*[X(Y4)|X*(Y3)], and CSEE*[X(Y4)|X*(Y3)] with quadratic polynomial smoothing of the CSEP*, CSEE*, CSEP, and CSEE.

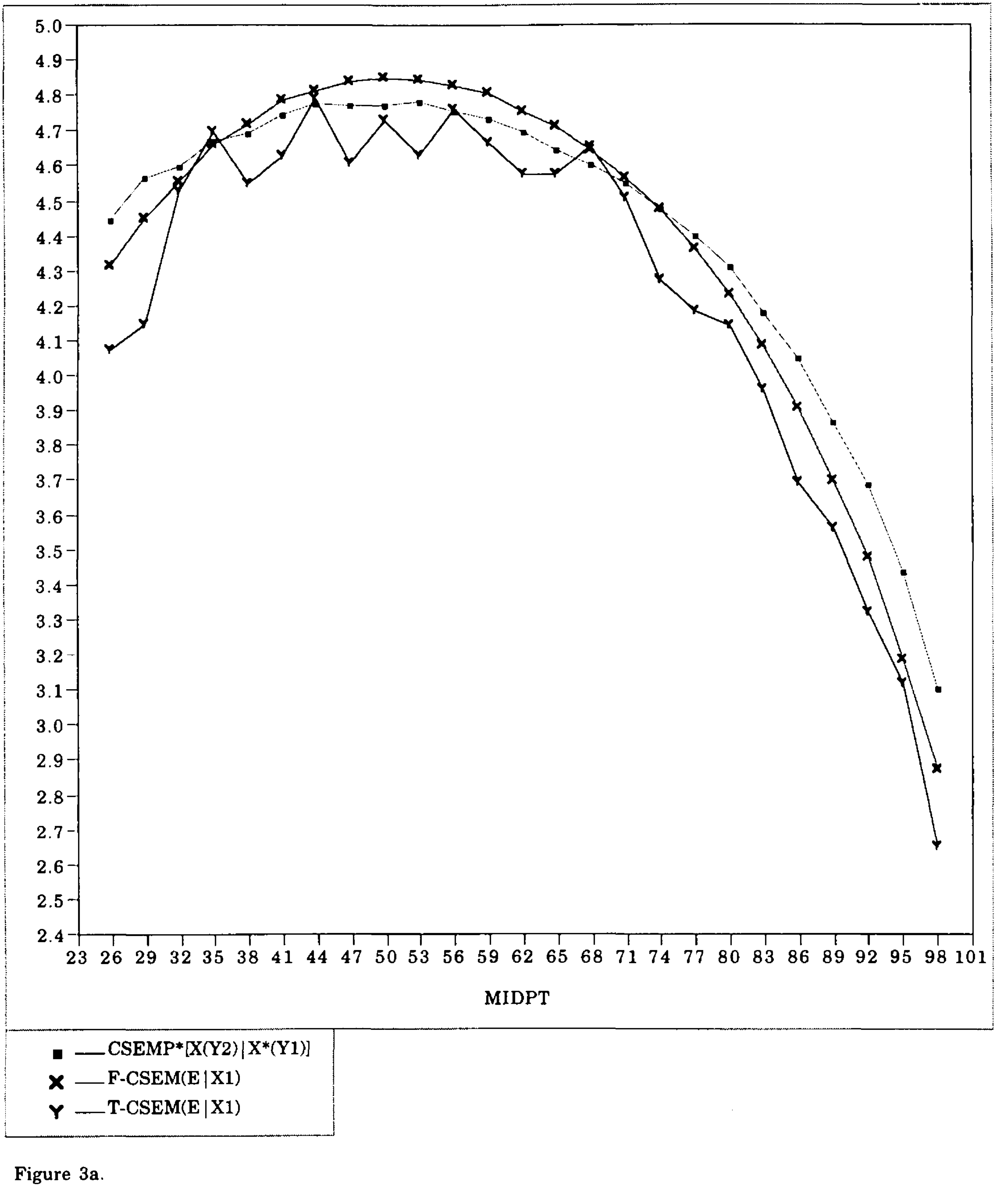

Figure 3a.

Plot of the stepped-up half-length CSEMP*[X(Y2)|X*(Y1)] against the Feldt et. al. (1985) ANOVA method estimate, F-CSEM(E|X1), and the Thorndike (1951) difference method estimate, T-CSEM(E|X1).

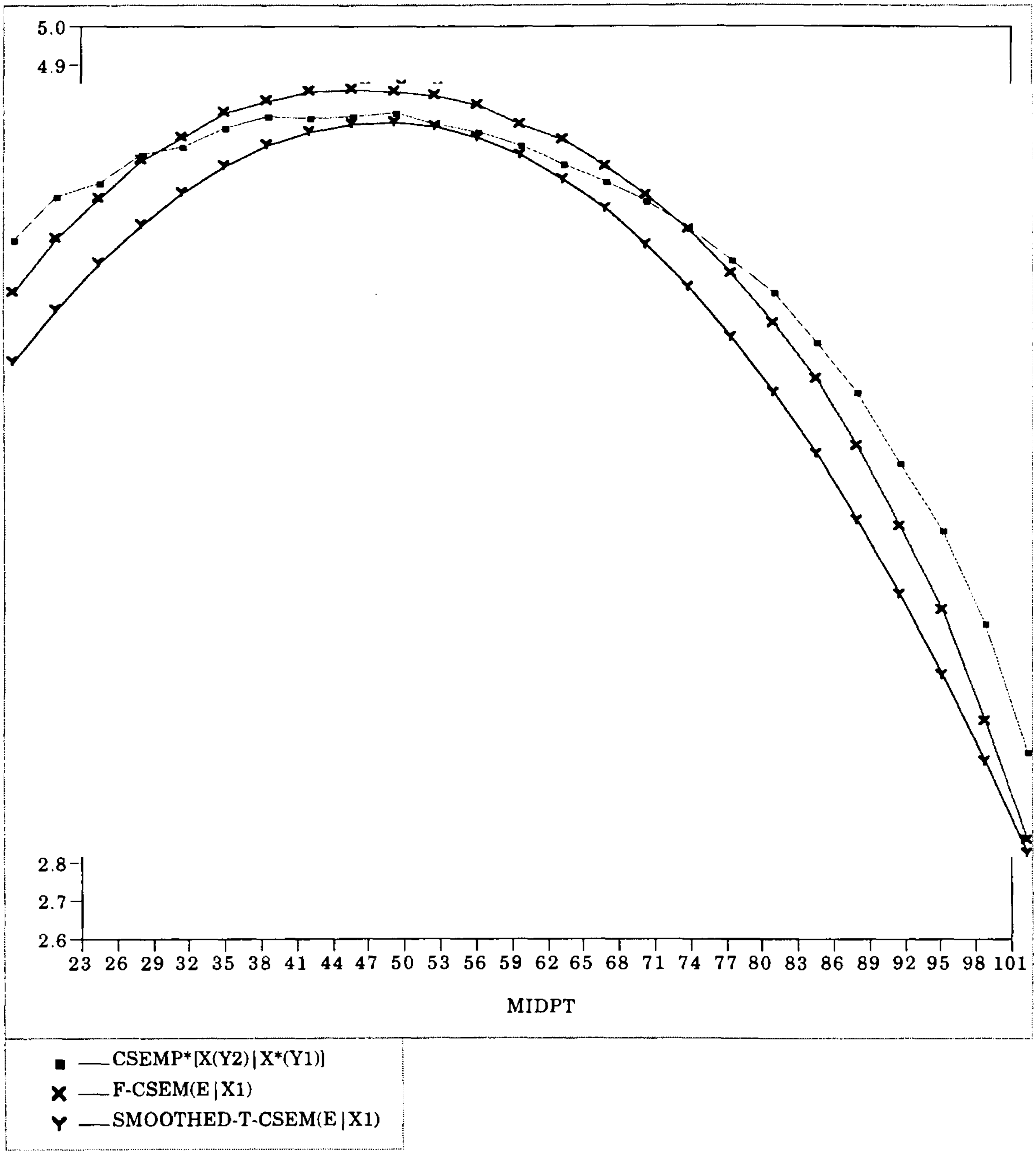

Figure 3b.

Plot of the stepped-up half-length CSEMP*[X(Y2)|X*(Y1)] against the Feldt et. al. (1985) ANOVA method F-CSEM(E|X1) and the Thorndike (1951) difference method T-CSEM(E|X1) with quadratic polynomial smoothing of the latter.

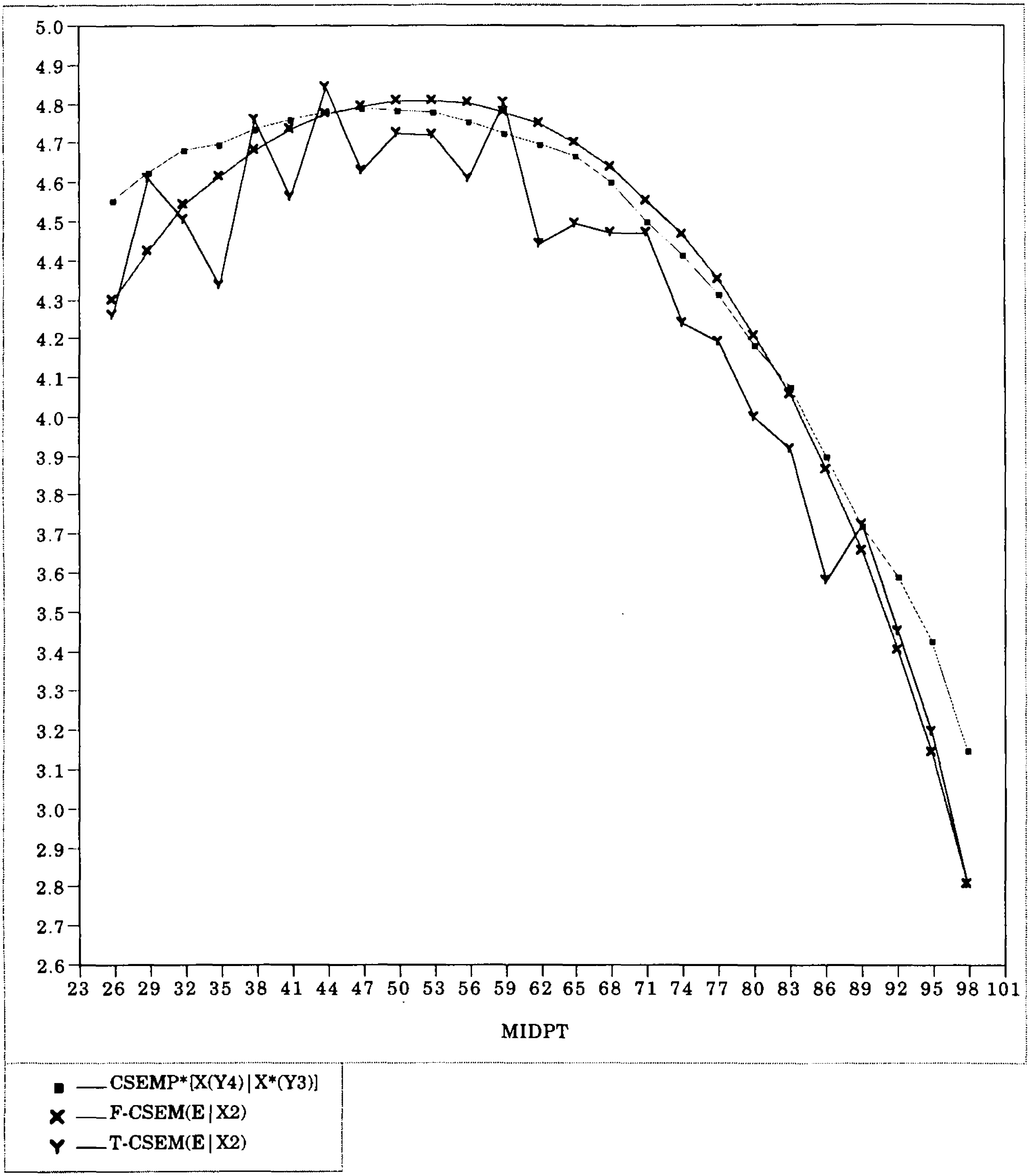

Figure 4a.

Plot of the stepped-up half-length CSEMP*[X(Y4)|X*(Y3)] against the Feldt et. al. (1985) ANOVA method estimate, F-CSEM(E|X2), and the Thorndike (1951) difference method estimate, T-CSEM(E|X2).

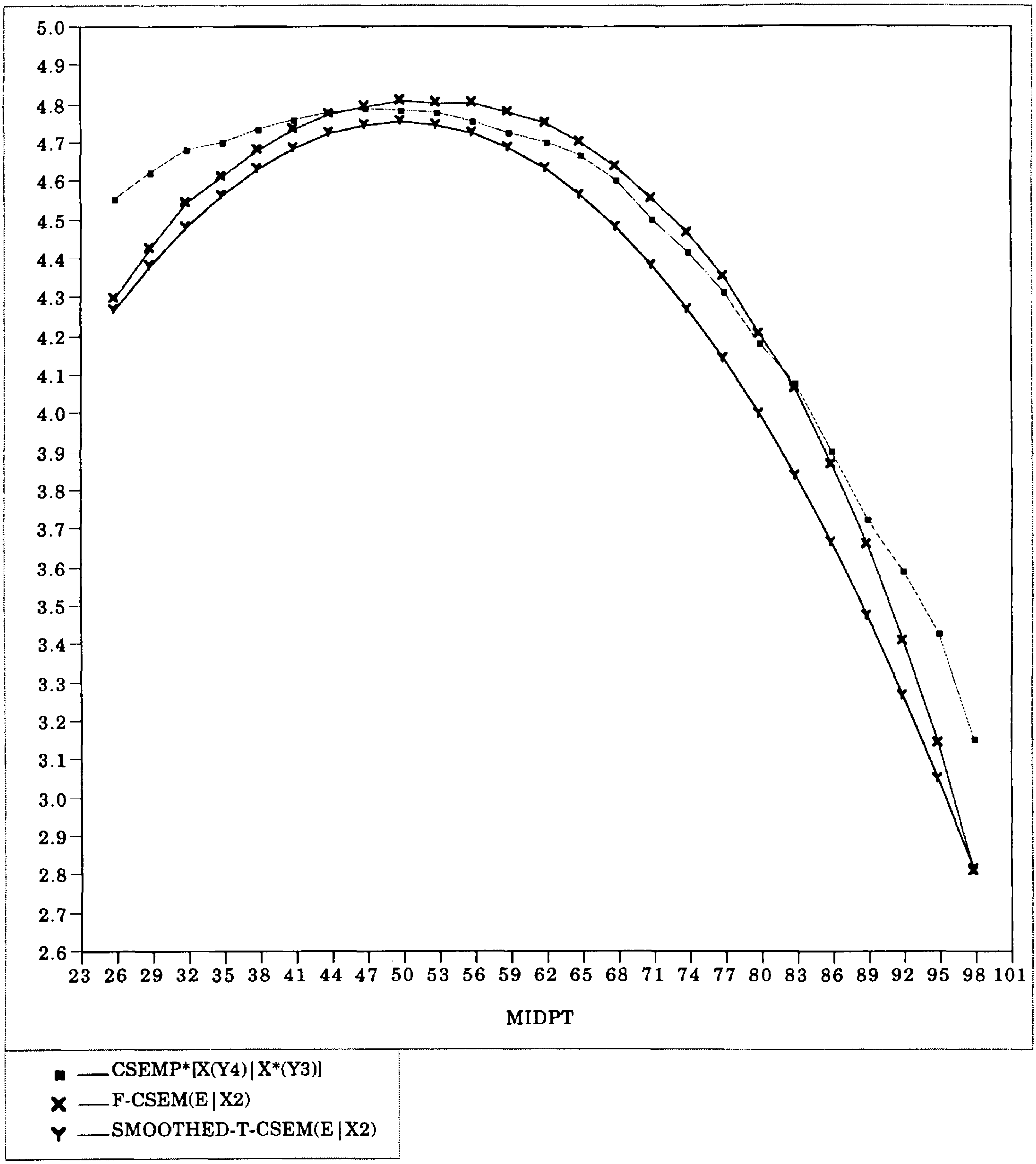

Figure 4b.

Plot of the stepped-up half-length CSEMP*[X(Y4)|X*(Y3)] against the Feldt et. al. (1985) ANOVA method F-CSEM(E|X2) and the Thorndike (1951) difference method T-CSEM(E|X2) with quadratic ploynomial smoothing of the latter.