Effects of Test Administration Mode on Item Parameter Estimates

Qing Yi
ACT, Inc.

Deborah J. Harris
ACT, Inc.

Tianyou Wang
Independent Consultant

Jae-Chun Ban
Korea Institute of Curriculum & Evaluation

Paper Presented at the Annual Meeting of National Council on Measurement in Education, April 22-25, Chicago, IL.

Abstract

The equivalency of a test administered in a traditional paper-and-pencil (P&P) format and on computer (CBT) needs to be established before the results from the two test administration modes can be used interchangeably. The current study analyzed data collected from two separate comparability studies. In both studies, random equivalent groups of examinees were administered either a mathematics or a reading test from two different testing programs (Testing Programs A and B). The mathematics and reading tests were administered in either a P&P or a linear CBT format.

Item parameters were estimated separately by mode and by pooling for P&P or CBT conditions. The reading tests showed some incomparability between test administration modes while the mathematics tests did not show much incomparability, based on the G-squared statistics (Thissen, Steinberg, & Gerrard, 1986). The implications of combining P&P and CBT data for item estimation versus using data from each mode separately, are discussed.

Effects of Test Administration Mode on Item Parameter Estimates

With the modern development of computer technology, more and more testing companies are considering using, or already have used, the computer for test delivery. A test can be administered on a computer linearly (CBT), that is, a fixed form of the paper-and-pencil (P&P) version of a test is given on the computer; or adaptively (CAT), that is, the next item to be administered is based on an examinee's response to a previously administered item. The current study investigates whether item parameter estimates differ using data collected via P&P and CBT test administration.

There are several advantages in administering a test on computer. Typically cited advantages include flexibility in test scheduling; reduced costs of test production, administration, and scoring; and the possibility of immediate score reporting. However, if the P&P and CBT versions of a test coexist, the equivalency of the test results from the two modes must be established. In particular, before treating test results from the two modes interchangeably, the effects of test administration mode on examinee scores or on item parameter estimates need to be examined. The concerns about the comparability between the P&P and CBT versions of a test are expressed in the American Psychological Association (APA) *Guidelines for Computer-Based Tests and Interpretations* (1986). Guideline 16 states:

> *When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means,*

dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (p.14).

The above guidelines identify criteria to use in the evaluation of the comparability between P&P and CBT versions of a test. The current study used data collected from two separate studies to demonstrate an IRT-based method to assess the comparability of P&P and CBT versions of a test. The implications of combining P&P and CBT data for item parameter estimation, or using the results from the P&P and the CBT test separately, are discussed.

Method

Data were collected from two separate comparability studies using tests of two different testing programs (Testing Programs A and B). The tests of these two testing programs have different content coverage and different reported score scales, although they both include a mathematics and a reading test. To achieve the goal of random equivalent groups design, a spiraling method was used so that examinees were administered either a mathematics or a reading test under either P&P format or on computer (CBT). The Testing Program A mathematics test contains 35 items and the reading test has 40 items. For Testing Program B, there are 30 items in both the mathematics and reading tests. Same exact test was administered under both modes. Number correct score was used to score the tests. Roughly 800 to 1000 examinees took each of the tests.

BILOG (Mislevy & Bock, 1990) was used to obtain item parameter calibrations and ability distributions for the P&P and CBT groups separately. BILOG was also used to obtain the item parameter calibrations and ability distributions when data for the P&P and CBT groups were combined. Three-parameter logistic (3-PL) item response theory (IRT) model was used for

all calibrations. Descriptive statistics of the calibrated item parameters are listed in Table 1. Table 1 also contains the correlation of item parameters between modes for the tests. Figures 1 to 4 present the item parameter calibrations for the Testing Programs A and B mathematics and reading tests between the modes, respectively. Figures 5 to 8 display the test characteristic functions based on the item parameter estimates obtained from separate calibration between the modes, and from combined data for testing Programs A and B, respectively.

The G-squared statistic (Thissen, Steinberg, & Gerrard, 1986) was used to compare the goodness of fit of two competing models. One model, hypothesized that the P&P and the CBT data are comparable, and the item response data can be pooled for calibration to obtain a single set of item parameter calibrations and θ distributions (we call this the single calibration model). The other model, hypothesized that the P&P and the CBT data are not comparable, thus separate calibrations must be performed for each data set (we call this the separate calibration model). The G-squared statistic is a goodness of fit index that is based on a generalized likelihood ratio statistic using a multinomial model. In this context, the response patterns are the cells for the multinomial model. In this generalized likelihood ratio test, the numerator is the likelihood under a constrained model, which is the IRT model with only one set of item parameter calibrations. The denominator is the maximum likelihood with no constraints, thus it is the likelihood with a maximum likelihood solution for the cell probabilities, which is the observed frequency for each cell divided by the sample size. The likelihood ratio can be expressed

$$Likelihood\ Ratio = \frac{\prod_{group}\prod_{cell} P_g(x)^{r_g(x)}}{\prod_{group}\prod_{cell}\left(\frac{r_g(x)}{N_g}\right)^{r_g(x)}} \qquad (1)$$

where x is the cell (i.e., response pattern), g is the group index (i.e., data were collected from P&P, CBT, or combined group), $r_g(x)$ is the observed frequency for the cell x, N_g is the sample size for group g, and $P_g(x)$ is the marginal probability of getting response pattern x under the IRT model, which can be expressed

$$P_g(x) = \int \prod_j P_{jg}(x_{jg} \mid \theta)\, \phi(\theta) d\theta \qquad (2)$$

where j indexes item, $P_{jg}(x_{jg} \mid \theta)$ is the conditional probability of getting response x_{jg}, which in turn is given by the 3-PL IRT model used in the calibration, and $\phi(\theta)$ is the θ distribution that is estimated using the BILOG program with 30 quadrature points. The integration in Equation 2 can be computed using the numerical integration method with the quadrature points and weights obtained from running the BILOG program. The G-squared statistic is the logarithm of the likelihood ratio in Equation 1 multiplied by -2, and thus can be expressed

$$G^2 = -2 \sum_{group} \sum_{cell} r_g(x) \ln\left[N_g P_g(x) / r_g(x)\right] \qquad (3)$$

The G-squared statistic is asymptotically distributed as a χ^2 distribution with degrees of freedom equal to [(the number of possible response patterns minus the number of item parameters) minus one)].

In this study, the G-squared statistic was computed for the single and separate calibration models for the Testing Programs A and B mathematics and reading tests, respectively. With the single calibration model, the same $P_g(x)$ is used for the P&P and CBT data in Equation 3. To compute $P_g(x)$, the single set of item parameter calibrations and $\phi(\theta)$ obtained by pooling the item response data across the two modes was used. In the separate calibration model, the item parameter calibrations and $\phi(\theta)$ obtained from the separate calibrations were used for each of the

two groups in computing $P_g(x)$. The difference of the G-squared statistics of these two competing models also asymptotically has a χ^2 distribution, with degrees of freedom equal to the difference of the degrees of freedom under the two models.

As the numbers of items increase, the numbers of possible response patterns increases dramatically. The sample size requirements for having enough observed frequencies in the cells also increases dramatically. Because of the limited sample sizes in this study, it is not feasible to compute the G-squared statistics for the entire tests. Therefore, the G-squared statistics were computed for item sets. The items were arbitrary divided into five consecutive sets with each set containing seven items for the Testing Program A and six items for the Testing Program B mathematics tests. For the reading test, four sets of 10 items each for the Testing Program A and five sets of six items each for the Testing Program B were obtained.

Results

Figures 1 to 4 present the item parameter estimates between the modes for the Testing Programs A and B mathematics and reading tests, respectively. Figures 1 and 2 show that for the Testing Program A, item parameter estimates for the mathematics test have a relative linear relationship, especially the *b*-parameter estimates, while the item parameter estimates for the reading test are more scattered between the modes. Figures 3 and 4 display the item parameter estimates for the Testing Program B tests. The *b*-parameter estimates for the mathematics test again show a very linear relationship, while the other item parameter estimates do not have a strong linear correlation between the modes. The correlations of item parameters between modes for the tests are listed in Table 1.

Based on the item parameter estimates obtained from the separate calibration between the modes and from the combined data, a test characteristic function (TCF) for each test was

computed. Figures 5 and 6 present the TCFs of Testing Program A's mathematics and reading tests. The differences among the TCFs for separate calibration between the modes and for the combined data are very small. Figures 7 and 8 have the TCFs of Testing Program B's mathematics and reading tests. The TCFs obtained from the CBT mode and from combined data look very similar, but are different from the TCF computed based on the item parameter estimates from the P&P mode for the mathematics test. For the reading test, the differences among the three TCFs are small for the high end of the ability scale, but differ elsewhere.

Table 2 contains the G-squared statistics for each of the two competing models, and the differences of the G-squared statistics. The differences that are statistically significant at a .05 α level are marked with an *. A statistically significant χ^2 difference indicates the null hypothesis of the P&P and CBT data being comparable is rejected. The G-squared statistics show that the mathematics tests for both Testing Programs display few statistically significant differences. Although the differences for the first item set of the Testing Program A mathematics test is statistically significant, the χ^2 value is only slightly above the critical value. For the Testing Program A reading test, the first two item sets do not show statistically significant differences, but the last two sets do show statistically significant differences. All five sets of items of the Testing Program B reading test show statistically significant differences.

The G-squared statistics seem to indicate that the item sets in the mathematics tests do not display statistically significant incomparability, and the reading tests do display statistically significant incomparability, for both Testing Programs A and B.

Discussion

This study examined the effects of test administration mode on item parameter estimates using data collected from two separate comparability studies. The results of the G-squared

statistics show that both Testing Programs A and B reading tests have some item sets that display statistically significant χ^2 differences between the modes, while the mathematics tests do not.

The comparisons of items administered in the P&P and CBT modes are based on the assumption that the groups of examinees in each mode are randomly equivalent. In addition, data used for this study are not from operational test administrations, but from two separate comparability studies. Therefore, it cannot be certain that the examinees who have participated in the studies behaved in the same way they would behave in a real test administration; they also were likely not as motivated as they would be in an operational testing.

It is recommended that the reading item sets with statistically significant differences be examined by content experts to see if there are some features of the sets or items in terms of their content or presentations that might have affected the comparability of the two modes. Another approach to try to understand the differences between an item administered in P&P mode and the same item administered in CBT mode would be to ask examinees what they are thinking during the test administrations (e.g., the "think aloud" method). When one wishes to use both the P&P and CBT modes to administer a test, and one wishes to use the obtained scores interchangeably, one needs to ensure item parameter estimates are stable across the two modes (assuming, as is the case in this study, that IRT item parameters are of interest). There are various approaches one may take: deleting items that show a mode effect; discerning what causes a mode effect and addressing the cause; refraining from using scores from the two modes interchangeably (e.g., set separate passing scores per mode); or attempting to compensate at the high level of interest (e.g., ensure score level comparability, even if there is not item level comparability).

The last point is important from a practical standpoint. In some of our CBT studies, we have discovered mode effects at the item level. However, when analyses were examined at the

total score level, on which actual decisions would be made, there was no effect (e.g., some items became easier, some became harder, and over all items the effect was negligible enough not to impact total scores).

This study looked at mode effects across two content areas: mathematics and reading tests, and two distinct types of tests within each content (i.e., Testing Programs A and B). For each test, item parameter estimates from the P&P and CBT modes were compared using the G-squared statistics. This methodology has not been used much in mode effects studies, but appears to be well suited for analyses such as those conducted in the present study.

The finding of mode effects for both Testing Programs, in reading but not in mathematics, is not easy to explain, as the items within type of test (e.g., Testing Program A mathematics and reading tests) are more similar than are items within content area, across Testing Programs. The fact that mode effects were not predicable should raise concern from practitioners, and encourage them to examine mode effects in various contexts before assuming item parameter estimates across different modes are interchangeable. This means, for example, that caution is needed when using items calibrated using P&P administrations to launch CBT tests operationally. The use of inappropriate item parameter estimates could increase the error in examinee scores, and result in, for instance, increased Type I and Type II error rates when pass/fail decisions are made.

References

American Psychological Association (1986). *Guidelines for computer-based tests and interpretations.* Washington, D. C.: American Psychological Association.

Mazzeo, J., & Harvey, A. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature.* College Board Report No. 88-8. College Entrance Examination Board: New York.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* [Computer program]. Chicago, IL: Scientific Software.

Thissen, D., Steinberg, L., and Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118-128.

Table 1. Correlation and Descriptive Statistics of Item Parameter Estimates between Modes and Tests

				Item Parameter Calibration		
				P&P	CBT	Together
Test	Parameter	Correlation	Statistics			
Testing Program A						
Mathematics	*a*	0.796	Mean	1.332	1.277	1.305
			SD	0.443	0.498	0.488
	b	0.984	Mean	0.230	0.330	0.274
			SD	0.892	0.916	0.905
	c	0.792	Mean	0.185	0.186	0.183
			SD	0.067	0.066	0.071
Reading	*a*	0.833	Mean	0.922	0.795	0.851
			SD	0.328	0.284	0.310
	b	0.965	Mean	0.265	0.285	0.264
			SD	0.967	1.055	1.017
	c	0.514	Mean	0.200	0.159	0.174
			SD	0.057	0.048	0.058
Testing Program B						
Mathematics	*a*	0.543	Mean	1.223	1.249	1.225
			SD	0.219	0.274	0.228
	b	0.992	Mean	-0.385	-0.295	-0.347
			SD	1.399	1.220	1.324
	c	0.699	Mean	0.178	0.174	0.168
			SD	0.058	0.053	0.062
Reading	*a*	0.759	Mean	0.880	0.839	0.857
			SD	0.297	0.308	0.316
	b	0.938	Mean	-0.569	-0.397	-0.468
			SD	1.826	1.627	1.814
	c	0.645	Mean	0.155	0.133	0.126
			SD	0.050	0.044	0.061

Table 2. G-squared Statistics.

Item Set	Testing Program A χ^2	Testing Program A DF	Testing Program B χ^2	Testing Program B DF
Mathematics Test				
Set One				
Single Calibration	311.64	107	117.05	46
Separate Calibration	277.78	86	109.31	28
Difference	33.86*	81	7.75	18
Set Two				
Single Calibration	268.74	107	178.56	46
Separate Calibration	247.59	86	164.14	28
Difference	21.15	21	14.42	18
Set Three				
Single Calibration	285.49	107	145.92	46
Separate Calibration	267.08	86	123.36	28
Difference	18.42	21	22.56	18
Set Four				
Single Calibration	264.77	107	172.08	46
Separate Calibration	238.75	86	155.82	28
Difference	26.02	21	16.26	18
Set Five				
Single Calibration	248.59	107	128.54	46
Separate Calibration	243.83	86	103.43	28
Difference	4.76	21	25.11	18
Reading Test				
Set One				
Single Calibration	1750.05	994	250.24	46
Separate Calibration	1712.55	964	145.87	28
Difference	37.50	30	104.37*	18
Set Two				
Single Calibration	1757.82	994	248.27	46
Separate Calibration	1717.69	964	149.95	28
Difference	40.12	30	98.32*	18
Set Three				
Single Calibration	1408.28	994	252.61	46
Separate Calibration	1347.33	964	169.72	28
Difference	60.95*	30	82.89*	18
Set Four				
Single Calibration	2598.86	994	267.04	46
Separate Calibration	2536.45	964	131.37	28
Difference	62.41*	30	135.66*	18
Set Five				
Single Calibration			208.49	46
Separate Calibration			134.38	28
Difference			74.12*	18

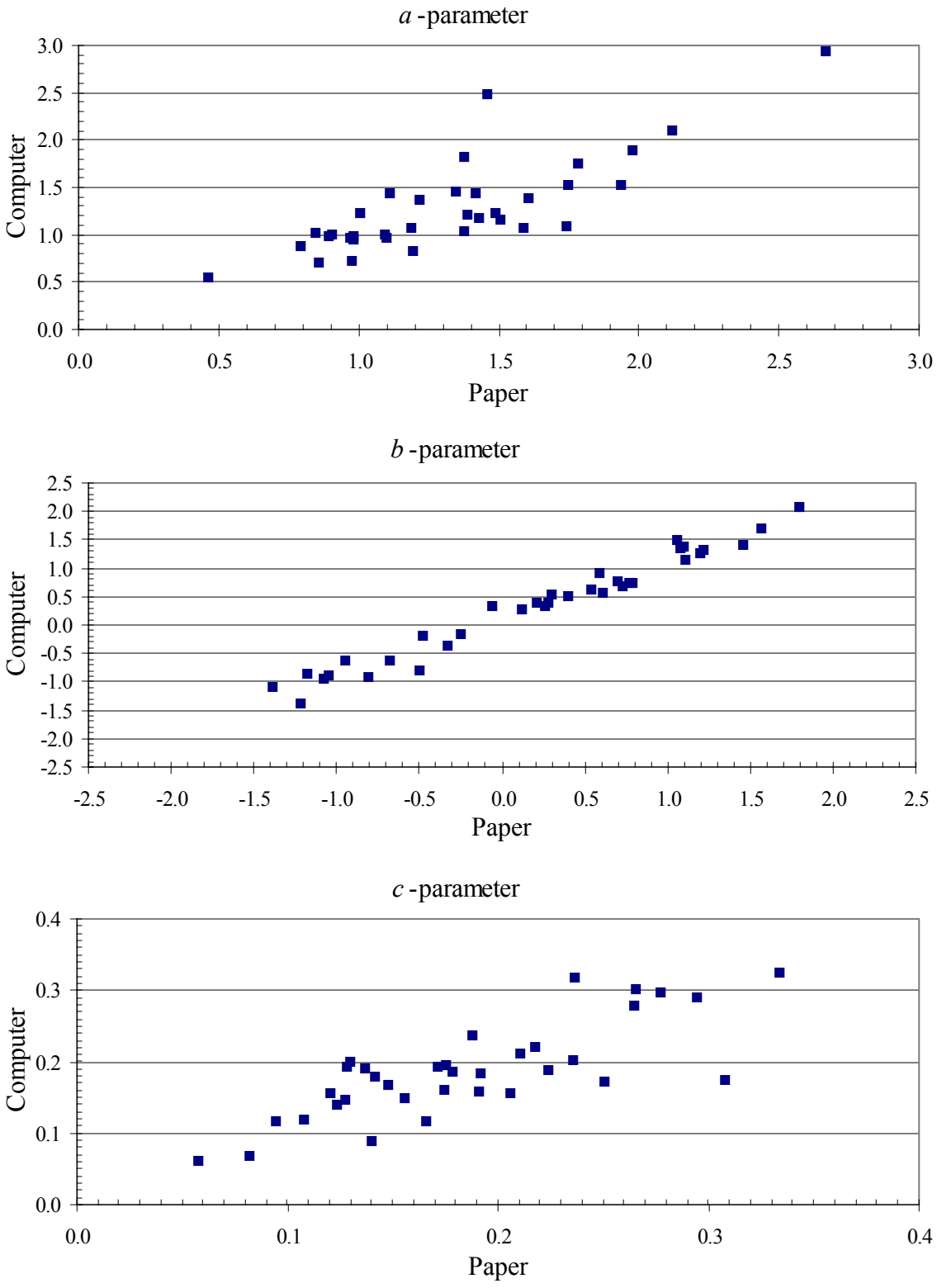

Figure 1. Item parameter calibration for Testing Program A mathematics test between modes.

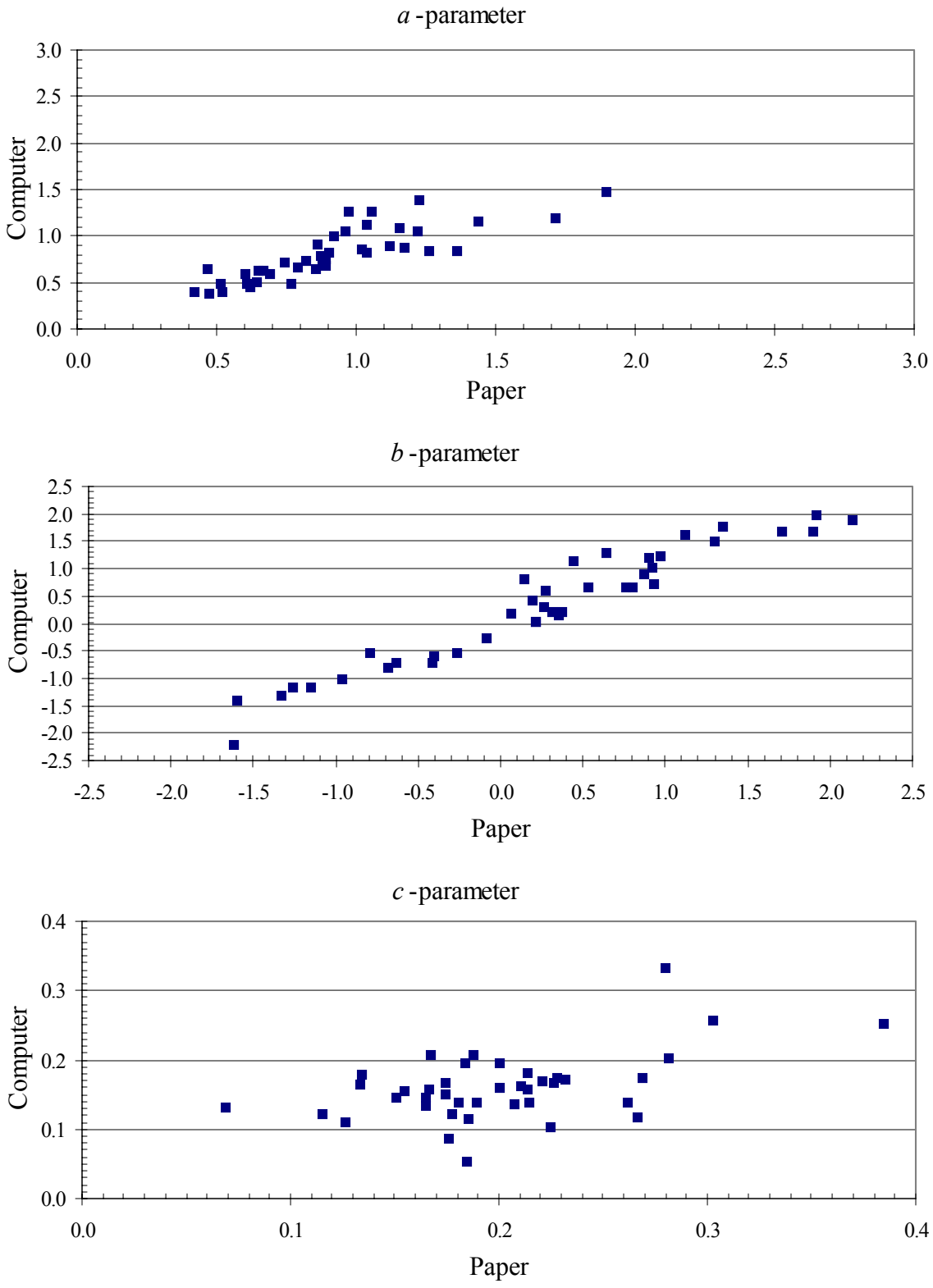

Figure 2. Item parameter calibration for Testing Program A reading test between modes.

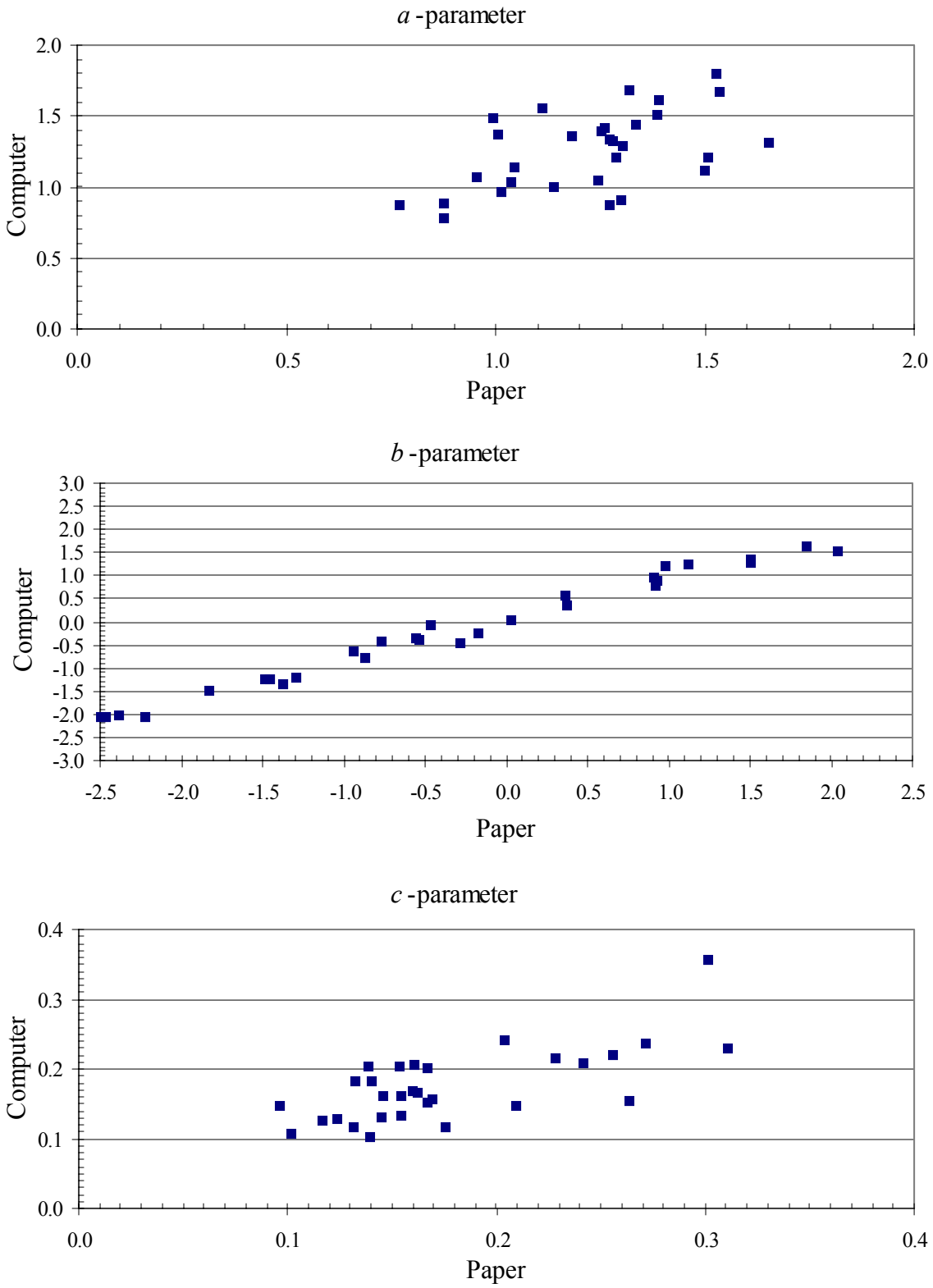

Figure 3. Item parameter calibration for Testing Program B mathematics test between modes.

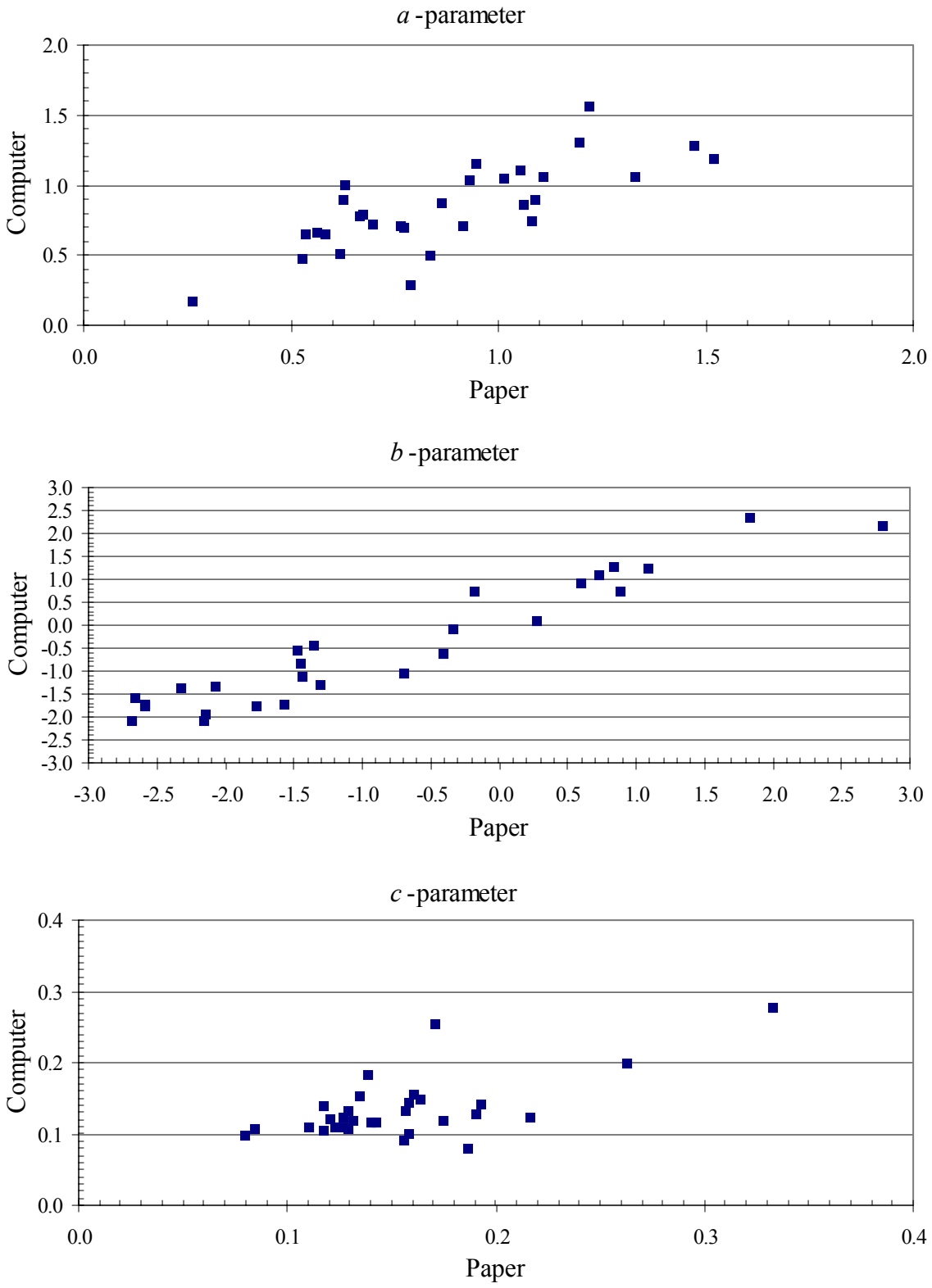

Figure 4. Item parameter calibration for Testing Program B reading test between modes.

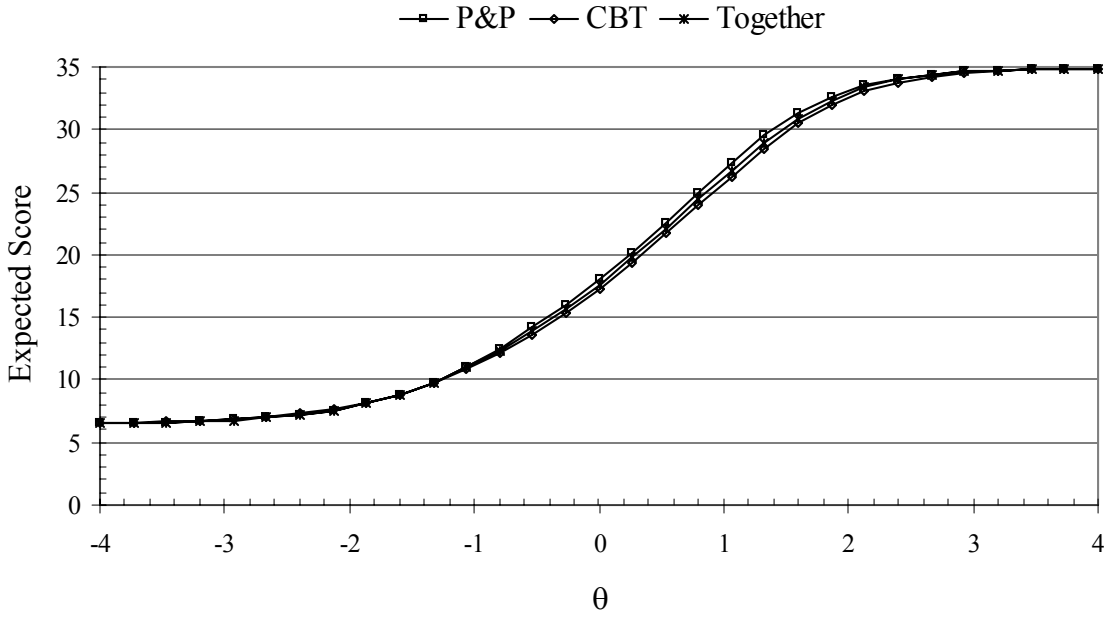

Figure 5. TCF for Testing Program A mathematics test obtained when item parameter calibrated separately between modes and calibrated together.

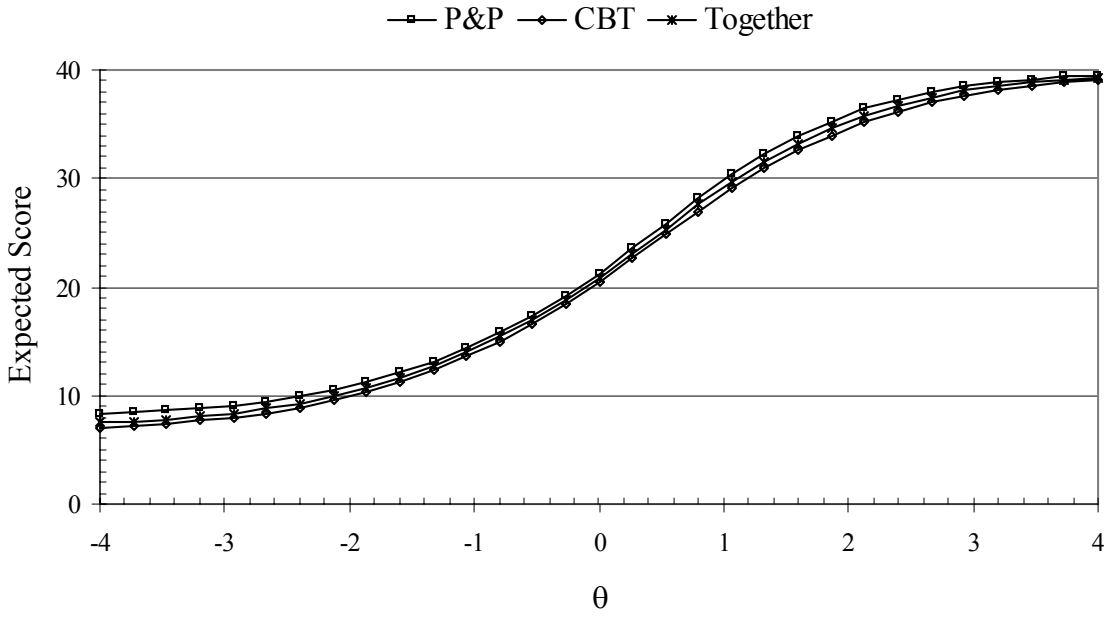

Figure 6. TCF for Testing Program A reading test obtained when item parameter calibrated separately between modes and calibrated together.

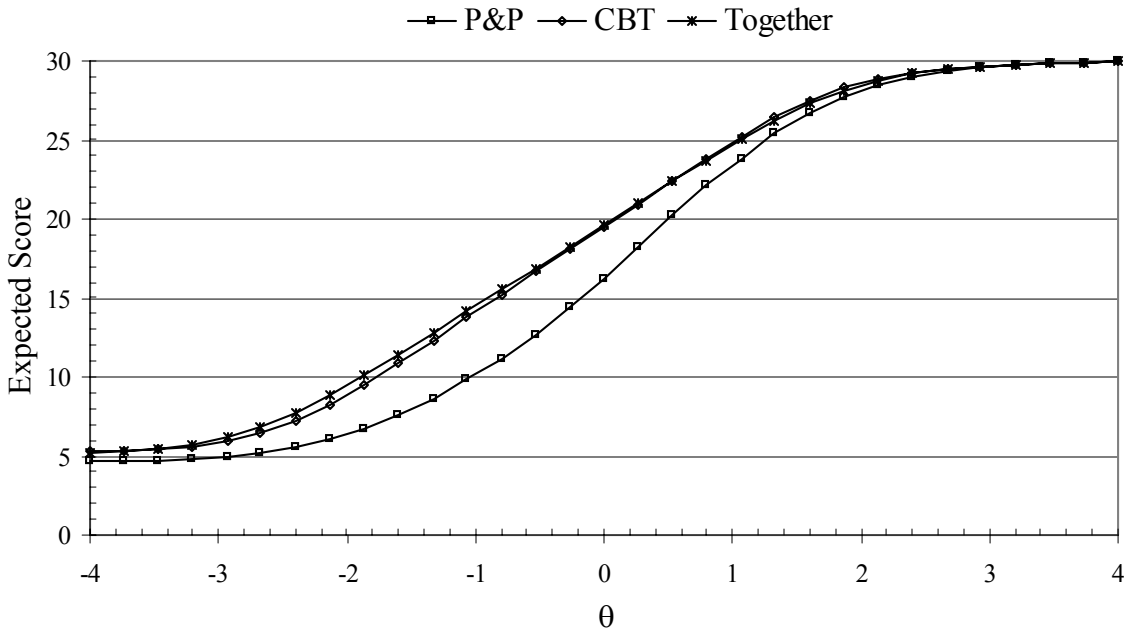

Figure 7. TCF for Testing Program B mathematics test obtained when item parameter calibrated separately between modes and calibrated together.

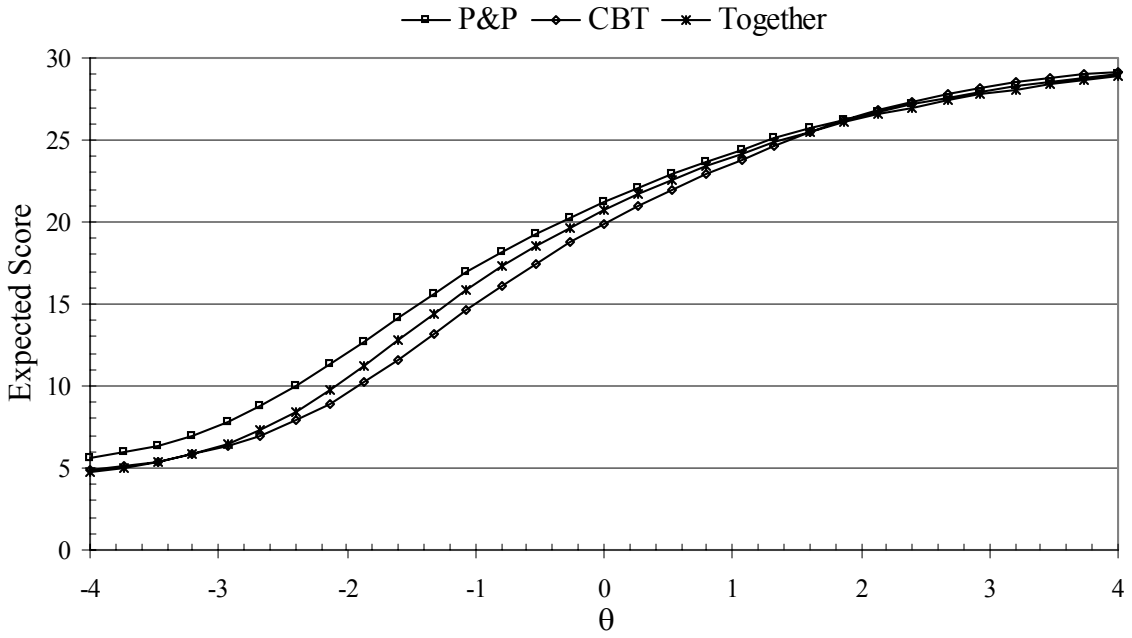

Figure 8. TCF for Testing Program B reading test obtained when item parameter calibrated separately between modes and calibrated together.