ED 468 759 TM 034 436

AUTHOR	Reese, Lynda; McKinley, Robert
TITLE	The Effect of COMC on the Stability of LSAT Item Parameter Estimates. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-93-01
PUB DATE	1993-06-00
NOTE	44p.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	College Entrance Examinations; *Estimation (Mathematics); Law Schools; Law Students; *Reliability; Test Items
IDENTIFIERS	Item Characteristic Function; *Item Parameters; LOGIST Computer Program; *Law School Admission Test

ABSTRACT

In item calibration using LOGIST (M. Wingersky, R. Patrick, and F. Lord, 1987), when the program determines that it cannot accurately estimate the c-parameter for a particular item due to insufficient information at the lower levels of ability, an estimate of the c-parameters, called COMC, is obtained by combining all such items. The purpose of this study was to evaluate the effect of using COMC on the stability of item response theory (IRT) parameter estimates. In this study, two Reading Comprehension pretest sections of the Law School Admission Test (LSAT) were administered twice, with samples of 1,00 and 1,030 for the June administration and 1,625 and 1,610 in December. and new parameter estimates were obtained for each administration. Different items were set to COMC for the two administrations due to sample fluctuations. In comparing the item characteristic curves for the different patterns of c-estimation, it seems that the most discrepant item characteristic curves are produced when the c-parameter is fixed for only one of the administrations. When the c-parameter is fixed for both administrations, item characteristic curves are more stable. However, results show that the use of COMC is not an acceptable procedure for dealing with the instability of the c-parameter. A means of deriving a more reasonable estimate of the c-parameter should be identified. (Contains 2 tables, 11 figures, and 2 references.) (SLD)

The Effect of COMC on the Stability of LSAT Item Parameter Estimates

Lynda Reese
Robert McKinley

Law School Admission Council
Statistical Report 93-01
June 1993

LAW Services

A Publication of Law School Admission Council/Law School Admission Services

The Law School Admission Council is a nonprofit association of United States and Canadian law schools. Law School Admission Services administers the Council's programs and provides services to the legal education community.

This study is published and distributed by the Law School Admission Council (LSAC) and Law School Admission Services (LSAS). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of LSAC/LSAS.

Introduction

Purpose

In item calibration using LOGIST (Wingersky, Patrick, & Lord, 1987), when the program determines that it cannot accurately estimate the c-parameter for a particular item due to insufficient information at the lower levels of ability, an estimate of the c-parameter, called COMC, is obtained by combining all such items. The stability criterion for a given item is b - 2/a, where a and b are the estimated parameters for that item. This value represents the ability level for which the proportion of correct responses is only slightly higher than the lower asymptote of the item response function. If there are few test takers whose abilities are lower than this level, then it is not possible to obtain a stable estimate of the c-parameter for the item. A minimum cutoff value called CRITFIXC is set, and if the value of b - 2/a is less than the value of CRITFIXC, the c-parameter is not estimated for that item and the value of COMC determined by combining all such items is assigned. The LOGIST manual suggests that the value of CRITFIXC be adjusted depending upon the sample size, with CRITFIXC being higher for smaller samples (Wingersky, Patrick, & Lord, 1987).

This aspect of the LOGIST program poses an interesting problem for testing programs that employ common item scaling to place different calibrations on a common scale. First, if different values of CRITFIXC are used for two calibrations, some items will have the c-parameter estimated in one calibration but not the other. This may very well produce a different c-parameter estimate for these items for the two calibrations, and since the COMC value tends to fall near the average of the estimated c-values, for high and low c-items, the difference between the two calibrations could be quite large. This appears to violate the IRT assumption of parameter invariance, since in one case the item c-parameter is estimated and in the other case the somewhat poorly defined parameter being estimated is specific to a set of items. Not only has the nature of the parameter being estimated changed, but the parameter being estimated by the COMC estimate is no longer invariant with respect to the other items included on the test, an assumption critical to applications such as adaptive testing. Moreover, this problem may be compounded by the correlated estimation errors among the IRT parameters. Marked changes in the c-parameter estimates such as may be observed when the c-parameter is estimated once but set to COMC in the second calibration may be accompanied by changes in the a- and b-parameter estimates, which perhaps could result in additional scaling errors.

Given these concerns, the purpose of this investigation was to evaluate the effect of using COMC on the stability of IRT parameter estimates. In this study, two LSAT Reading Comprehension pretest sections were administered twice, once in June and again in December of 1991, and new parameter estimates were obtained for each of these administrations. Using the same value of CRITFIXC, different items were set to COMC for the two administrations due to sample fluctuations. Items were categorized as to their pattern of c-estimation, with the three categories being (1) c-parameter estimated at both administrations, (2) c-parameter fixed at one administration, and (3) c- parameter fixed at both administrations. The purpose of this study was to determine the effect of the pattern of c-estimation on the stability of the item parameter estimates obtained using LOGIST.

LSAT Item Calibration

Currently, in the administration of the LSAT, each test item is administered twice before ever appearing on a final test form. First, each item appears in a pretest section. The purpose of this testing stage is to carry out an initial screening of the items. Next, items are administered within the preoperational testing stage which is carried out in order that test forms may be preequated prior to their operational administration. Finally, the intact test form is administered operationally.

Each candidate at an LSAT test administration receives one operational section each of Reading Comprehension and Analytical Reasoning and two operational sections of Logical Reasoning. In addition, each test taker also receives a variable test section which is comprised of either pretest or preoperational items from one of the three question types. Given the sample size at a typical LSAT administration, many different sections of test items may be administered at a single administration within the variable section. This is accomplished through a spiraling plan in which several different variable sections are administered to random samples of test takers. In this way, an entire form of the test may be preequated and several forms may be pretested at every administration.

The LSAT utilizes the IRT 3-parameter logistic model in calibration for which the probability of a test taker at a particular level of ability responding correctly to a given item is represented by the equation

$$P_i(\Theta)=c_i+(1-c_i)\{1+\exp[-1.7a_i(\Theta-b_i)]\}^{-1}, \tag{1}$$

where a_i, b_i, and c_i are the discrimination, difficulty and pseudo-guessing parameters for a given item and Θ is the test taker ability parameter. Items are calibrated using the computer program LOGIST (Wingersky, Patrick, & Lord, 1987). Both the June and December pretests were calibrated simultaneously with their associated final forms. The two calibrations were scaled using the characteristic curve method (Stocking & Lord, 1983) using the December final form, which was preequated in June, as the common item set.

Methodology

Overview

The purpose of this section is to describe both the data that were analyzed for this study and the statistics used to carry out the analyses. All analyses were carried out to compare the calibration results for the two administrations both for the entire set of 40 items (test level) and for each of the c-estimation categories (category level). The analyses were designed to study effects both at the item parameter level and at the more global item characteristic curve level.

Data

A set of four Reading Comprehension passages and the 40 items associated with them were administered within two LSAT pretest sections during both the June and December 1991 administrations of the LSAT. The reason for administering pretest sections twice rather than examining a section that had been administered both at the preoperational and operational testing stages is that these two testing stages differ radically in their sample sizes. By using the pretest stage, a similar sample size in the two administrations was assured. Within each pretest section, all four passages were administered, each followed by seven items. The first four items administered with a passage were common to both sections and appeared in exactly the same order in both sections. The last three items associated with each passage were unique to that pretest section. Thus, 10 items were administered in association with each reading passage at each administration, with four items administered in both sections and six items administered in only one section or the other. The passages were administered in exactly the same order in the two administrations, and the noncommon items were administered in exactly the same order for both administrations. Table 1 presents the sample sizes for the items in each section for both administrations.

Table 1
Sample Size by Administration

Section	June 1991	December 1991
1	1020	1625
2	1030	1610

The scheme developed for numbering the items reflects the format in which the items were administered. The first digit is a letter reflecting the item set, A, B, C, or D, to which the item belongs. Next, a number from 1 to 10 was assigned to reflect the position of the individual item. Items numbered 1 to 4 are common to both pretest sections and appeared in the first four positions of the set. Items 5, 6, and 7 appeared in the first pretest section, while items 8, 9, and 10 appeared in the second pretest section.

Analyses

Categorization based on pattern of C-estimation. At the outset, items were categorized as to the pattern of c-estimation for the two administrations in the manner described in the Introduction section. For the June 1991 administration, the value of COMC was .144, while the value of COMC for the December 1991 administration was .122. The value of CRITFIXC was the same for both administrations, with this value being set at -3.5.

Summary statistics, correlations, and scatterplots. Initially, summary statistics of the item parameter estimates were studied for the two administrations. The number of items, minimum value, maximum value, and mean and standard deviation are compared and discussed both at the test level and at the category level. Next, both at the test level and category level, correlations and scatterplots among the items are presented and discussed.

Root mean squared difference. The root mean squared difference at each value of Θ ($RMSD_\Theta$) between test characteristic curves for the two administrations are presented and discussed at both the test level and category level. In calculating this statistic, the difference between the probability of a correct response for an item at each of 17 levels of Θ (-4 to 4 with an increment of .5) was taken, and these differences were squared. At each level of Θ, the mean of the squared differences was calculated across items. Finally, the square root of this value was taken at each level of Θ . The results of this analysis are presented in graph form. The equation for the $RMSD_\Theta$ may be written as

$$RMSD_\Theta = \{1/n \sum_{i=1}^{n} [P(\Theta)_J - P(\Theta)_D]_i^2\}^{1/2}, \qquad (2)$$

where $RMSD_\Theta$ is the RMSD at a given value of Θ, $P(\Theta)_J$ represents the probability of a correct response to an item for the June 1991 administration, $P(\Theta)_D$ represents the probability of a correct response to an item for the December 1991 administration, and n is the number of items being analyzed.

Bias statistic. In calculating the bias statistic, again, the difference between the probability of a correct response at each administration was calculated for each item at each of 17 levels of Θ, and the mean of this difference at each Θ-level was taken. Again, the results of this analysis both at the test level and the category level are presented in graph form. The equation for the bias statistic is presented in Equation 3.

$$Bias_\Theta = 1/n \sum_{i=1}^{n} [P(\Theta)_J - P(\Theta)_D]_i \qquad (3)$$

Item characteristic curve overlay plots. Overlay plots of the item characteristic curves derived at each of the two administrations were studied in order to more carefully evaluate the effect of the different patterns of c-estimation. In order to evaluate the extent to which these differences are represented at the test level, an overlay plot of the test characteristic curves for the two administrations was also presented and analyzed.

Results

Overview

In this section, the results of the analyses that have been carried out will be described and discussed. All of the results reported reveal quite consistent patterns. Overall, the greatest agreement between the two calibrations is found for items having the c-parameter fixed for both administrations, while the greatest discrepancies are found between items for which the c-parameter was fixed for one administration but not the other. In general, differences at the category level do not present themselves at the test level.

Pattern of C-estimation

Items were categorized as to whether or not the c-parameters were fixed for the two administrations. For the majority of the items (n = 25, 62.5%), the c-parameter was not fixed for either administration. For six items (15%), the c-parameter was fixed for one or the other administration, but not both. Finally, for nine items (22.5%), the c-parameter was fixed for both administrations.

It was of interest to know if the distribution of items among the three patterns of c-estimation was representative of what would be expected to occur at other pairs of administrations. In order to make this determination, comparisons between the preoperational and operational stages for the total test were available for Forms 11, 12, and 13 of the LSAT. For these forms, the c-parameter was estimated for both stages for 66.3%, 49.5%, and 59.4% of the items for Forms 11, 12, and 13, respectively. With the exception of Form 12, these percentages are very close to the 62.5% observed for the two pretests being evaluated here. Items were assigned the value of COMC for only one stage for 12.9%, 14.9%, and 19.8% of the items on Forms 11 through 13, and this again is very close to the 15% observed for the two pretest sections being evaluated here. Finally, 20.8%, 35.6%, and 20.8% of the items for Forms 11 through 13 were set to COMC for both administrations. While the percentage for Form 12 is a bit higher, the percentages for Forms 11 and 13 are very similar to the 22.5% observed for the two pretests under study.

Summary Statistics

Category level. The summary statistics for the item parameter estimates are presented in Table 2. The statistics for the category level reveal that when the c-parameter is estimated for both administrations, both the mean and standard deviations of all of the item parameter estimates are quite similar. This is not the case when the c-parameter is fixed for only one administration. The mean a-, b-, and c-parameters for the two administrations are very different under this pattern of c-estimation, while the standard deviations for all of the item parameters in this category are quite similar. Finally, the summary statistics for the a- and b-parameters when the c-parameter was not estimated for either administration are more similar than any observed for the other patterns of c-estimation. This implies that the a- and b-parameters are most stable when the c-parameter is fixed for both administrations.

Test level. In studying the summary statistics of the item parameter estimates at the test level, no extreme differences appear to be present between the two administrations. For the a- and c-parameter estimates, all summary statistics are nearly identical. Slightly more difference may be observed between the b-values for the two administrations, with a slightly higher mean and standard deviation observed for the June administration (.307 and .941 for June as opposed to .191 and .901 for December). However, these differences do not appear to be substantial. When compared to the summary statistics at the category level, the test level analyses reveal more agreement than that observed when the c-parameter was fixed for only one administration, but less agreement than was observed when the c-parameter was fixed for both administrations.

Table 2
Summary Statistics by Estimation Category

		IRT Parameter/Administration					
Estimation		a		b		c	
Category[1]	Statistic	J	D	J	D	J	D
Both							
		25	25	25	25	25	25
	N	0.411	0.547	(0.595)	(0.233)	0.00	0.00
	Min	1.403	1.360	1.915	1.605	0.368	0.321
	Max	0.825	0.807	0.777	0.710	0.186	0.188
	S	0.219	0.179	0.656	0.594	0.119	0.086
One							
	N	6	6	6	6	6	6
	Min	0.483	0.303	(0.579)	(0.819)	0.078	0.122
	Max	1.023	0.747	1.592	0.782	0.508	0.421
	Mean	0.714	0.543	0.236	(0.287)	0.328	0.172
	S	0.174	0.177	0.734	0.599	0.176	0.122
Neither							
	N	9	9	9	9		
	Min	0.426	0.451	(1.807)	(1.873)		
	Max	0.785	0.727	(0.258)	(0.189)		
	Mean	0.573	0.575	(0.952)	(0.930)		
	S	0.107	0.104	0.460	0.495		
Total							
	N	40	40	40	40	40	40
	Min	0.411	0.303	(1.807)	(1.873)	0.00	0.00
	Max	1.403	1.360	1.915	1.605	0.508	0.421
	Mean	0.752	0.715	0.307	0.191	0.198	0.171
	S	0.216	0.201	0.941	0.901	0.127	0.085

[1] Both = c-parameter estimated at both administrations, One = c-parameter estimated at only one administration, Neither = c-parameter estimated at neither administration, and Total = all 40 items

Correlations and Scatterplots

Category level. The correlations among the item parameters are presented in Table 3, and the graphs of these relationships for the a-, b-, and c-parameters are presented in Figures 1, 2, and 3, respectively. The correlations among the a- and b-parameters when the c-parameter was estimated for both administrations are moderate, with a correlation of .620 for the a-parameters and a correlation of .912 for the b-parameters. The graphs of these relationships clearly depict the degree of association between the a- and b-parameter estimates. The c-parameters also show a moderate relationship, with a correlation of .693. This moderate, positive relationship is clearly depicted in the graph presented in Figure 3. For the June administration, the correlation between the a- and b-parameters for this categorization is .284, while for the December administration, the correlation between these parameters is .424. The relationship between the a- and b-parameters for the two administrations are fairly similar for this c-estimation category.

The correlations among the item parameter estimates for the items for which the c-parameter was estimated for only one administration should be interpreted cautiously due to the small number of items in this classification. This pattern of estimation had a clear effect upon the parameter estimates. For the a- and c-parameters, the relationship between the estimates for the two administrations is negative, with correlations of -.750 for the a-parameter estimates and -.513 for the c-parameter estimates. With respect to the c-parameter estimates, the graph of this relationship shows clearly that the one item at the top of the graph for which the c-parameter was fixed for only December is having a very strong effect upon the correlation, and is primarily responsible for this weak relationship. For the b-parameter estimates, the correlation of .328 is quite weak. The graph of this relationship indicates that the outlier at the top of this graph is greatly responsible for this weak relationship. With only six observations, an outlier of this extreme would greatly reduce the correlation. The correlations between the a- and b-parameters are very different for the two administrations for this categorization with a correlation of .725 for June and a correlation of .290 for December. These values are more discrepant than observed for any other categorization.

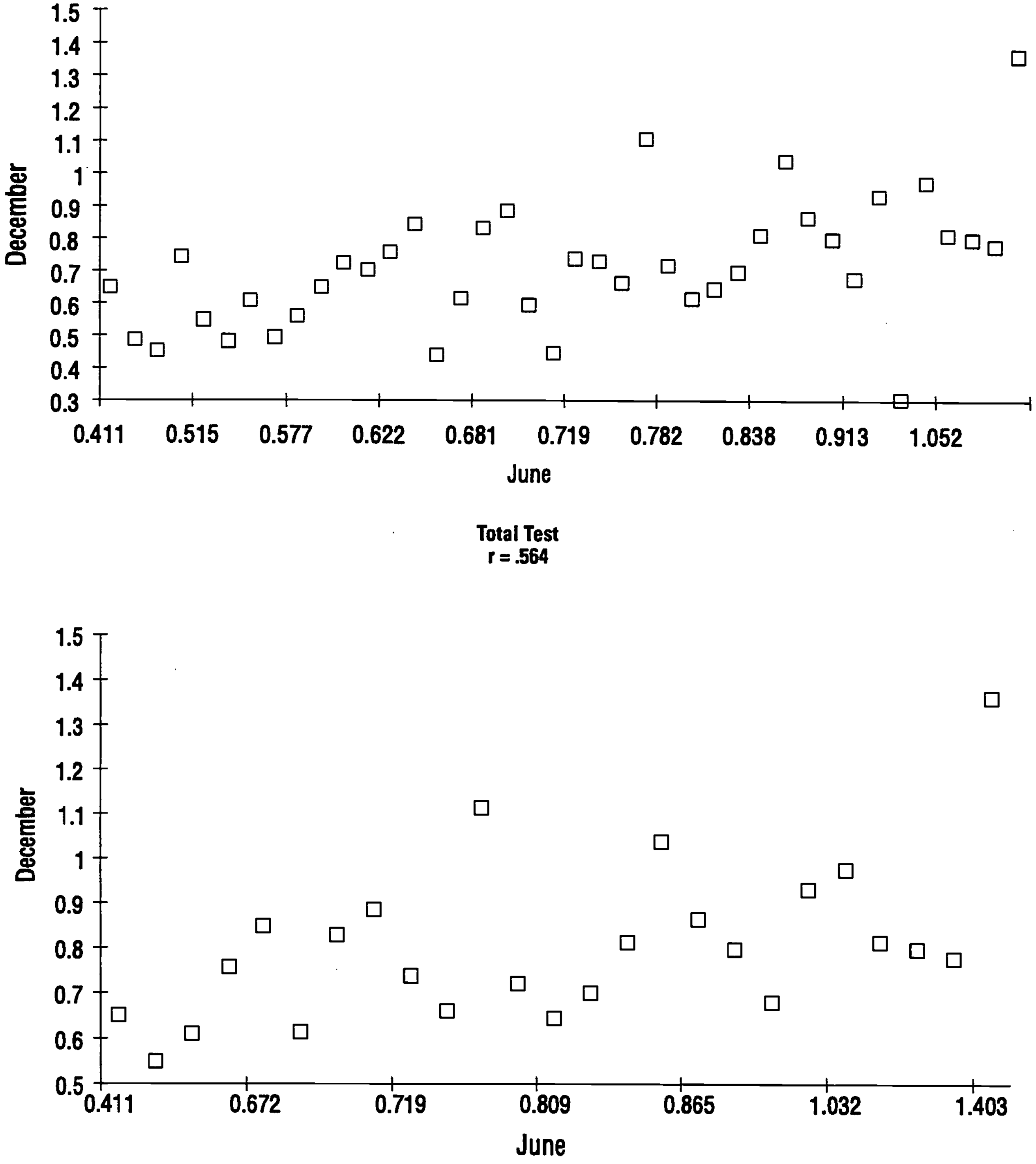
Finally, the correlations between the a- and b-parameters for the items having the c-parameter fixed for both administrations are the highest observed with a correlation of .648 for the a-parameter estimates and a correlation of .955 for the b-parameter estimates. In comparing the graphs of these relationships to the others discussed, the stronger relationship among the item parameters for this categorization is clear. In terms of the relationship between the a- and b-parameters for this categorization, greater similarity was also observed between the two administrations than was observed for any of the other categories. These correlations are quite low and negative, with a correlation of -.075 for June and a correlation of -.072 for December.

Test level. At the test level, a correlation of .564 between the a-parameters is indicative of a moderate positive association. The b-parameters for the two administrations show the strongest relationship, with a correlation of .908, and the c-parameters show the weakest relationship with a correlation of .323. In comparing the scatterplots of the b- and c-parameters for the two administrations, the strong relationship between the b-parameters is clearly apparent. For the c-parameters, there is more scatter present in the graph, but the weaker relationship between these two sets of parameters appears to be to a great extent attributable to the four items lining up to the right of the graph. These items appear to all have a fixed value of the c-parameter for the December administration, but different values of the c-parameter for the June administration. In terms of the relationship between the a- and b-parameters for the two administrations, the correlations of .543 for June and .609 for December are moderate and very similar. For all of the item parameters, the correlations observed at the test level are stronger than those observed when the c-parameter was fixed for only one administration, but weaker than those observed when the c-estimation was consistent across the two administrations. Similarly, the correlations between the a- and b-parameters are less consistent than those observed with the c-parameter fixed for both administrations, but more similar than for the other two estimation categories.

Table 3
Correlations by Estimation Category

		Administration/Parameter					
Estimation	Admin/	June			December		
Category	Param	a	b	c	a	b	c
Both							
	June b	.284	1.00	.543	.354	.912	.375
	June c	.496	.543	1.00	.025	.356	.693
	Dec a	.620	.354	.025	1.00	.424	.221
	Dec b	.172	.912	.356	.424	1.00	.432
	Dec c	.306	.375	.693	.221	.432	1.00
One							
	June b	.725	1.00	.716	(.697)	.328	(.093)
	June c	.594	.716	1.00	(.851)	(.273)	(.513)
	Dec a	(.750)	(.697)	(.851)	1.00	.290	.562
	Dec b	(.227)	.328	(.273)	.290	1.00	.875
	Dec c	(.651)	(.093)	(.513)	.562	.875	1.00
Neither							
	June b	(.075)	1.00		(.225)	.995	
	Dec a	.648	(.225)		1.00	(.072)	
	Dec b	(.124)	.955		(.072)	1.00	
Total							
	June b	.543	1.00	.409	.433	.908	.391
	June c	.409	.409	1.00	(.214)	.108	.323
	Dec a	.564	.433	(.214)	1.00	.609	.363
	Dec b	.424	.908	.108	.609	1.00	.523
	Dec c	.250	.391	.323	.363	.523	1.00

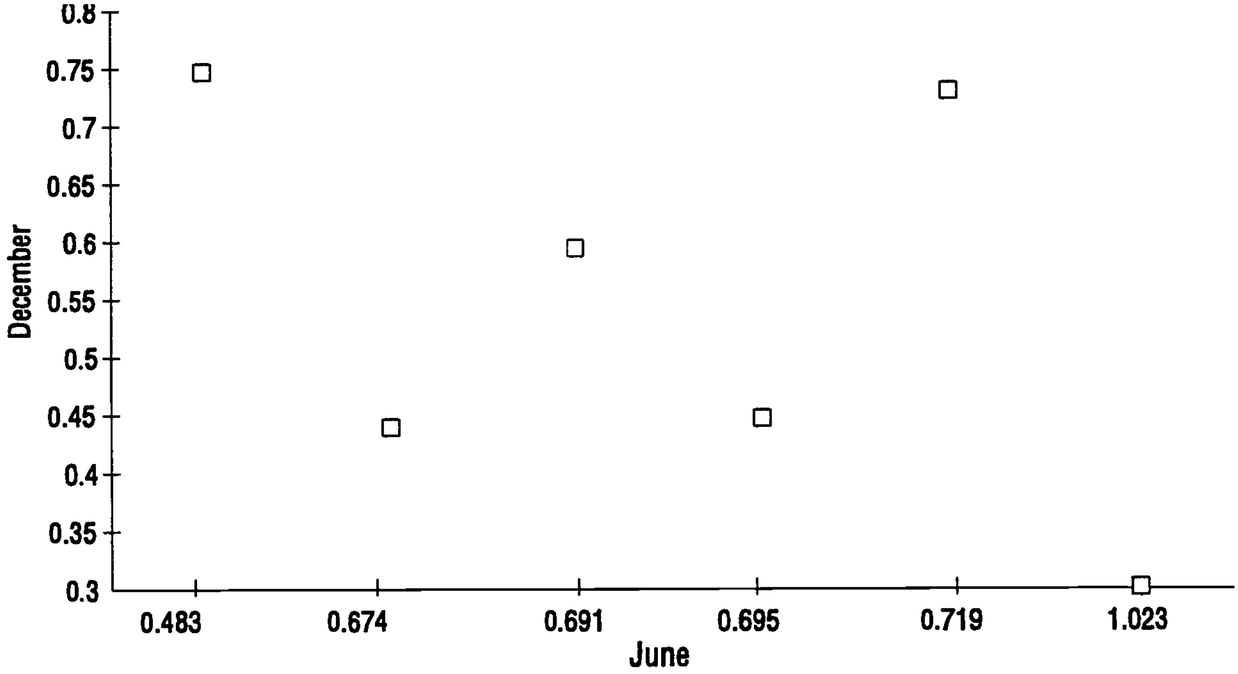
Figure 1
Scatterplots of the A-Parameters



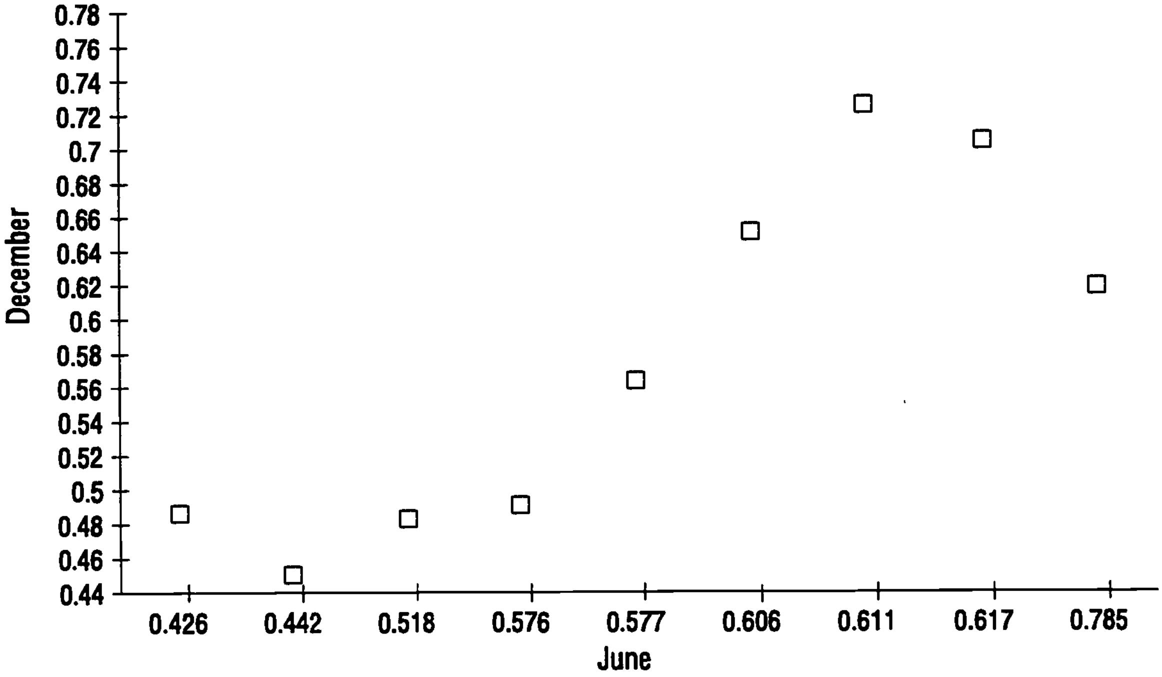
Total Test
r = .564

Both
r = .620

Figure 1 (continued)

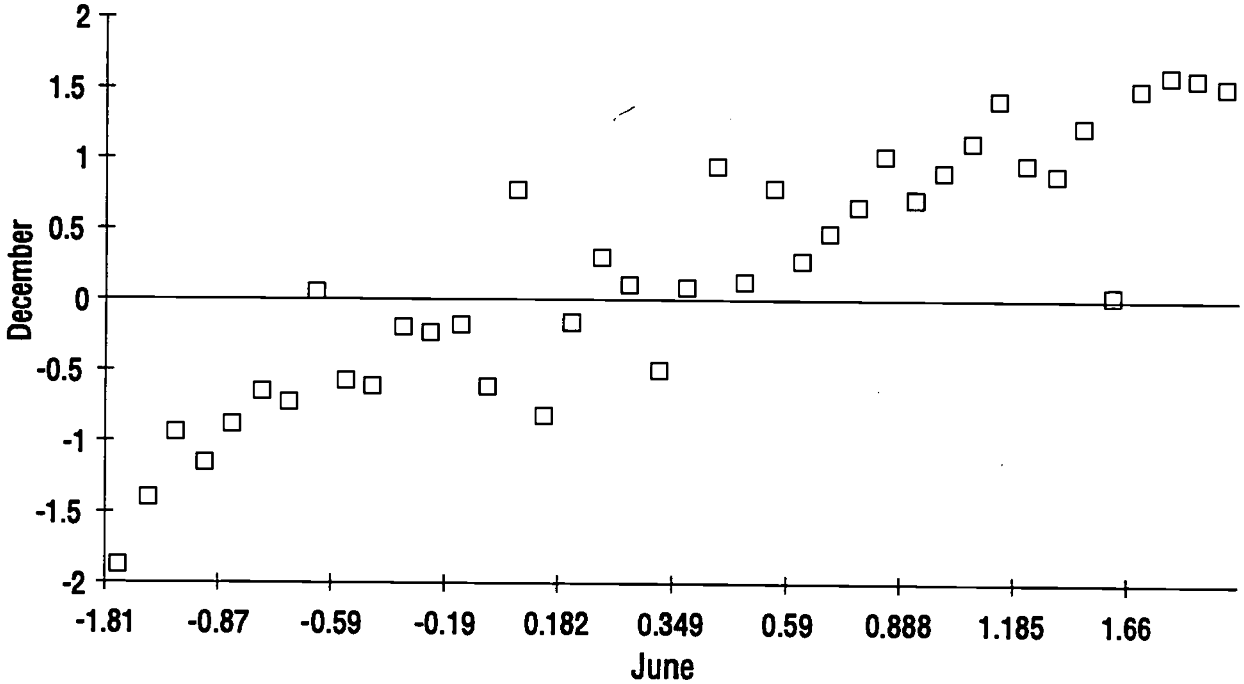


One
r = -.750

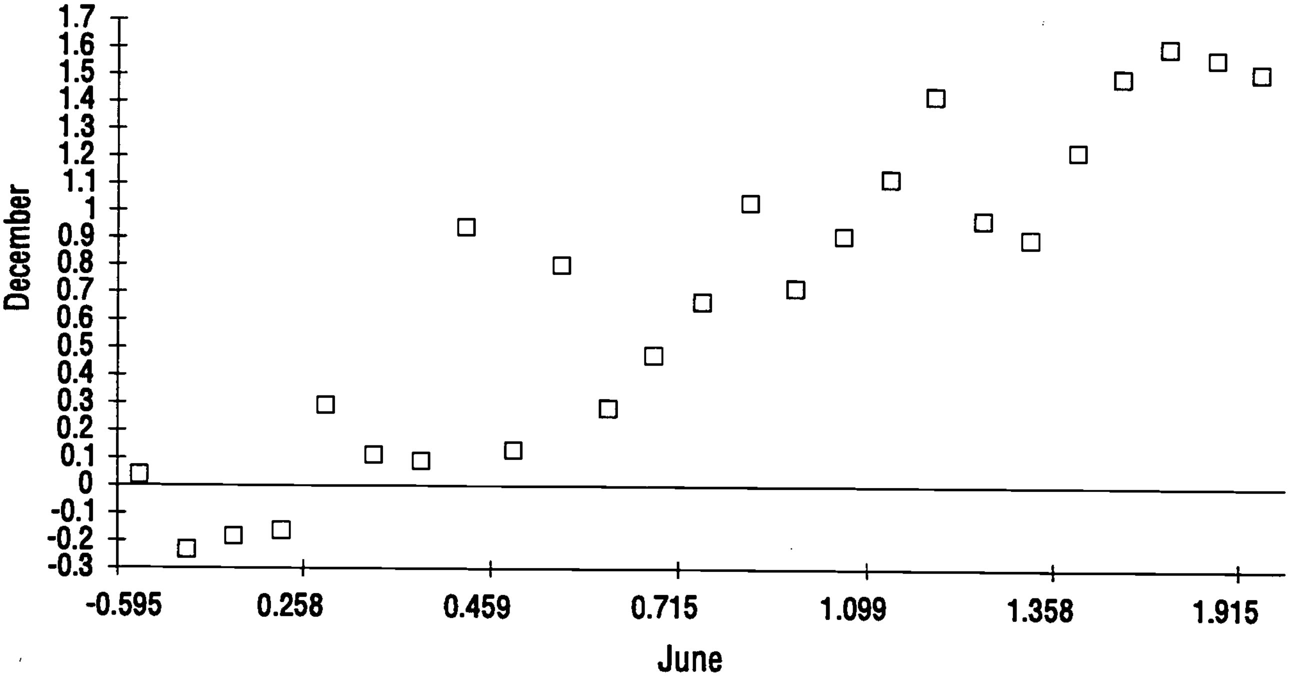


Neither
r = .648

Figure 2
Scatterplots of the B-Parameters

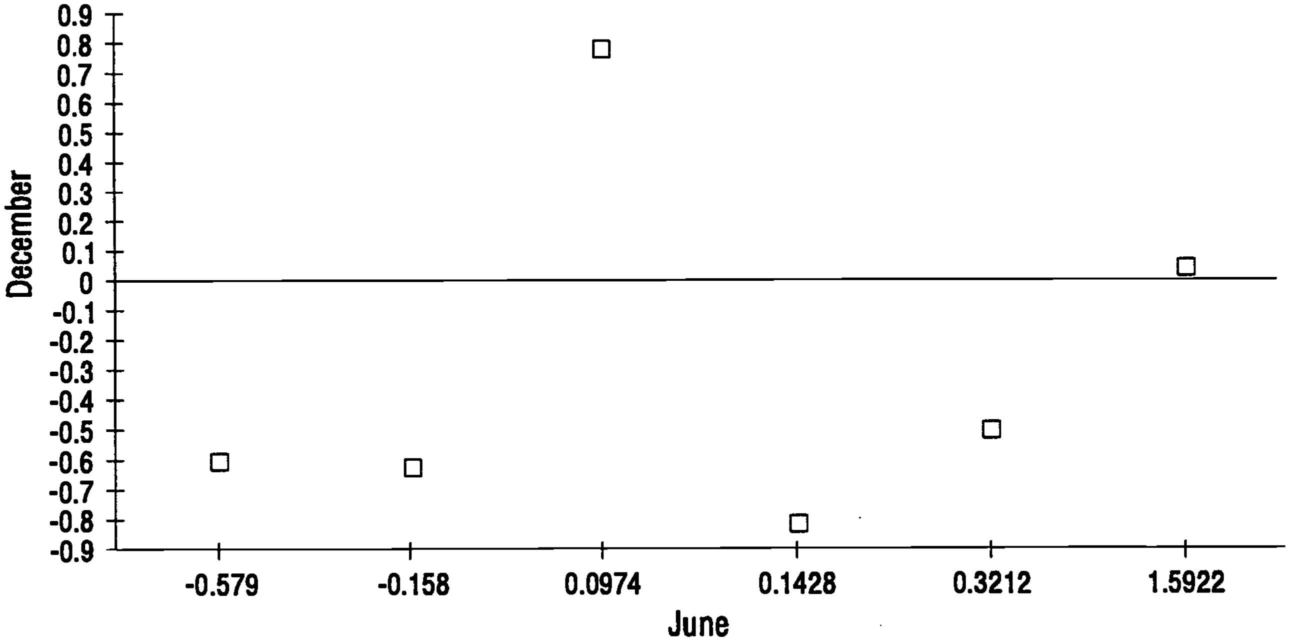


Total Test
r = .908

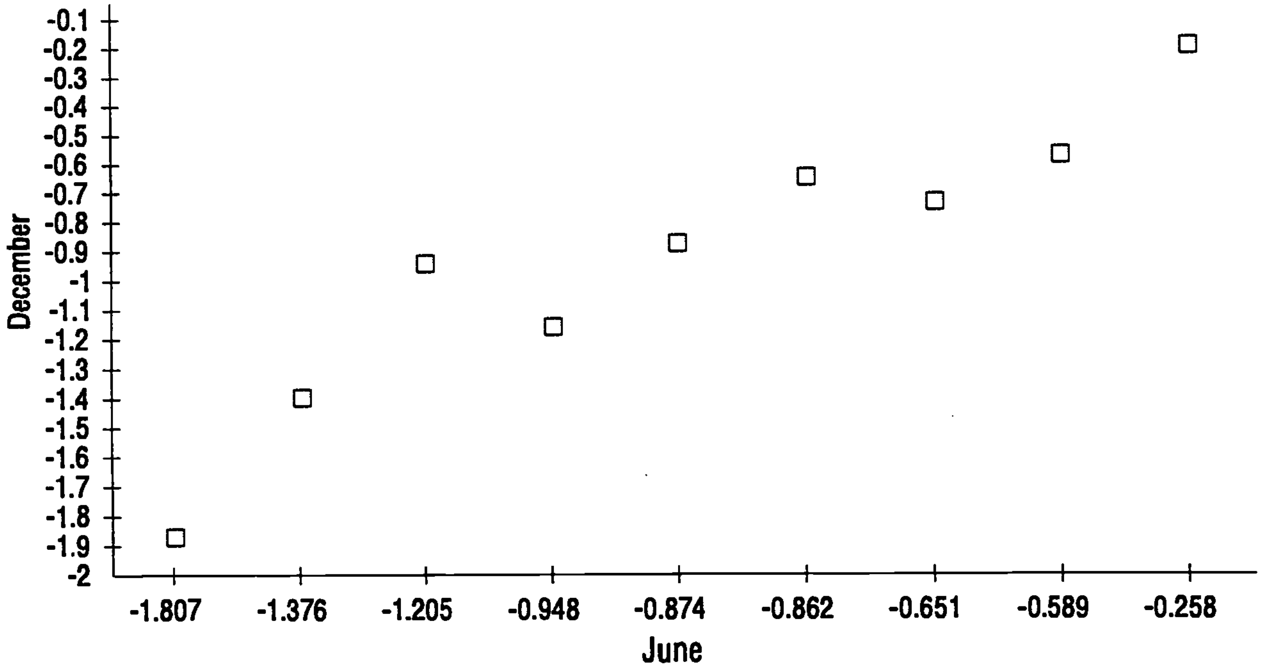


Both
r = .912

Figure 2 (continued)

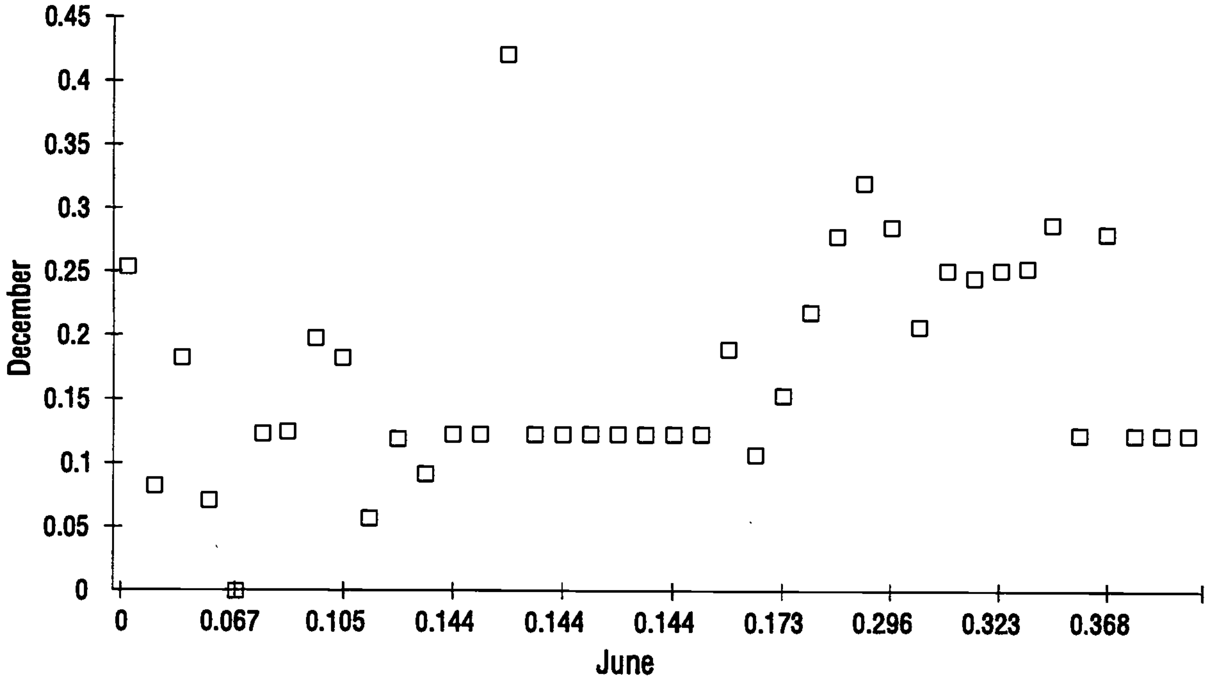


One
r = .328

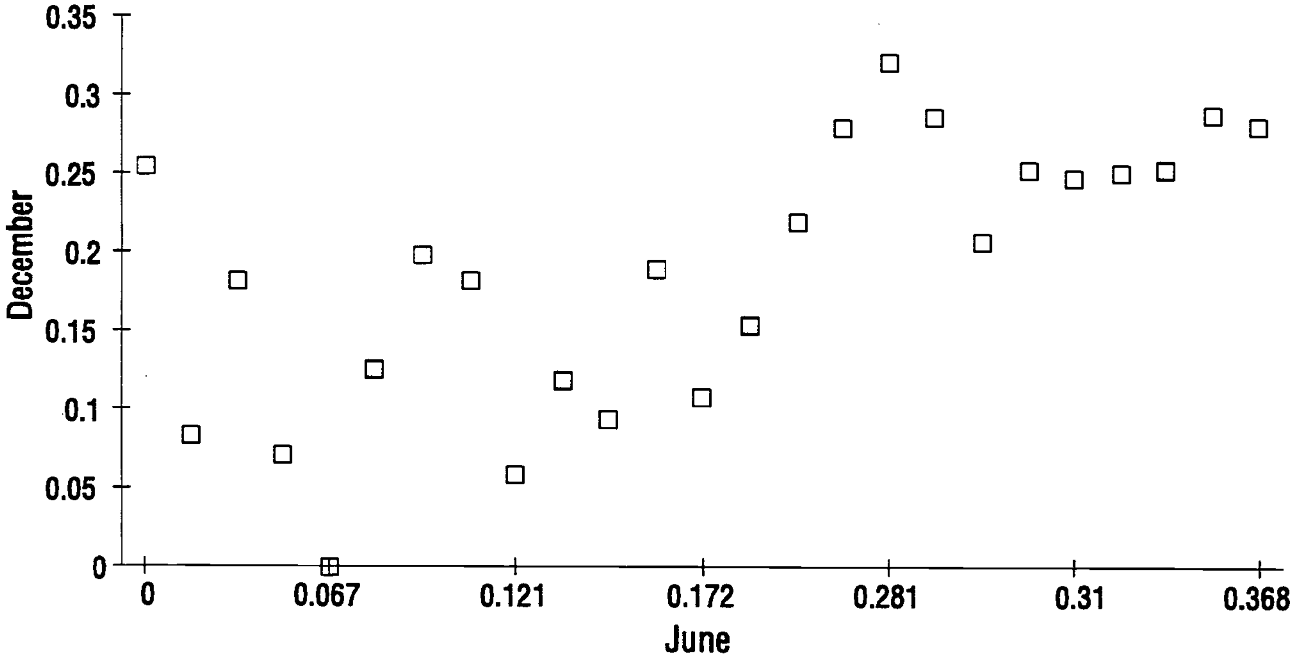


Neither
r = .955

Figure 3
Scatterplots of the C-Parameters

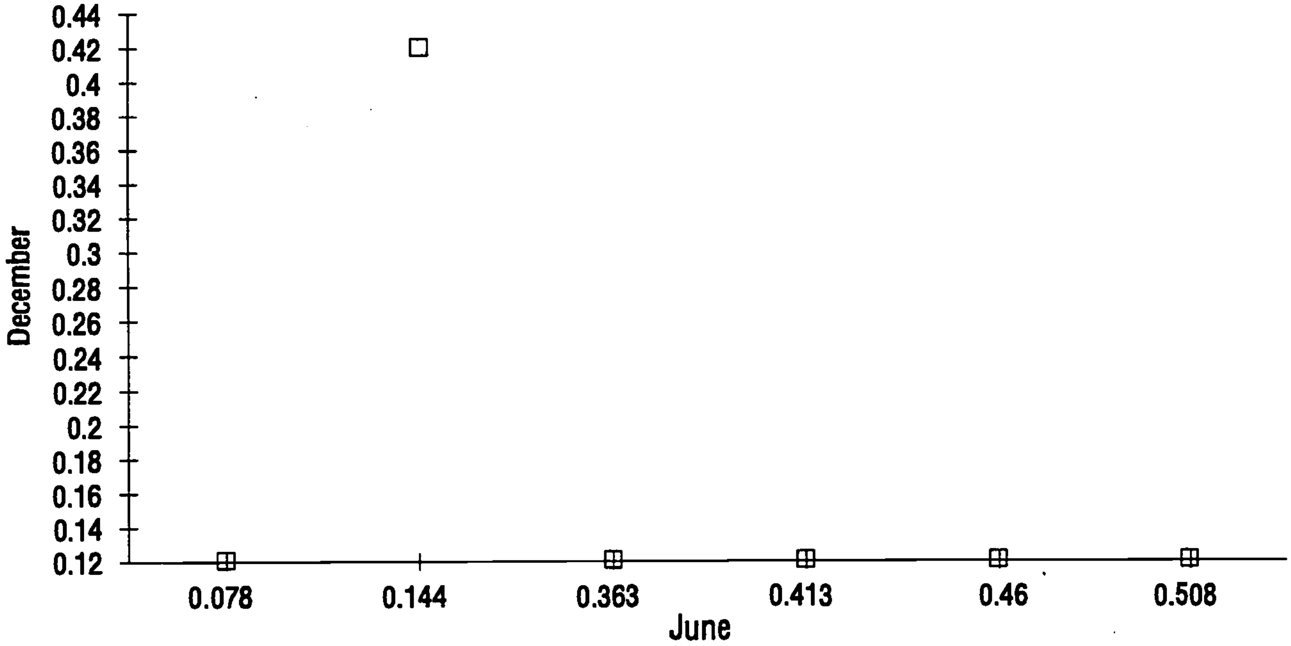


Total Test
r = .323



Both
r = .693

Figure 3 (continued)

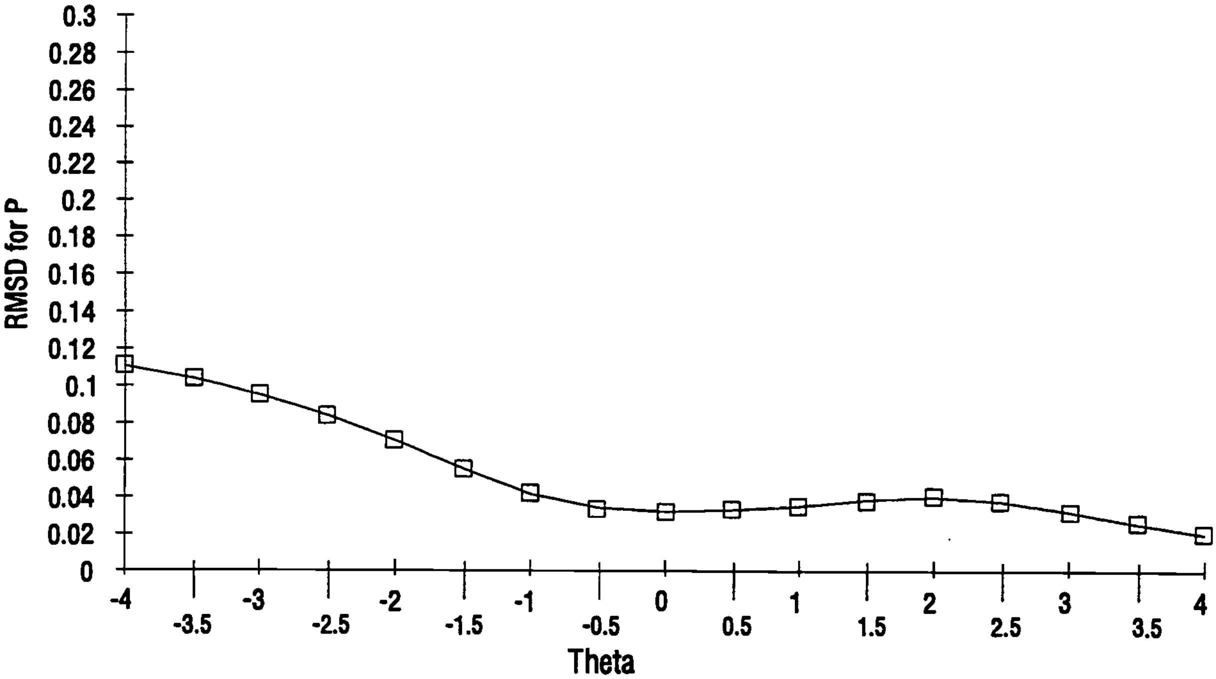


One
r = -.513

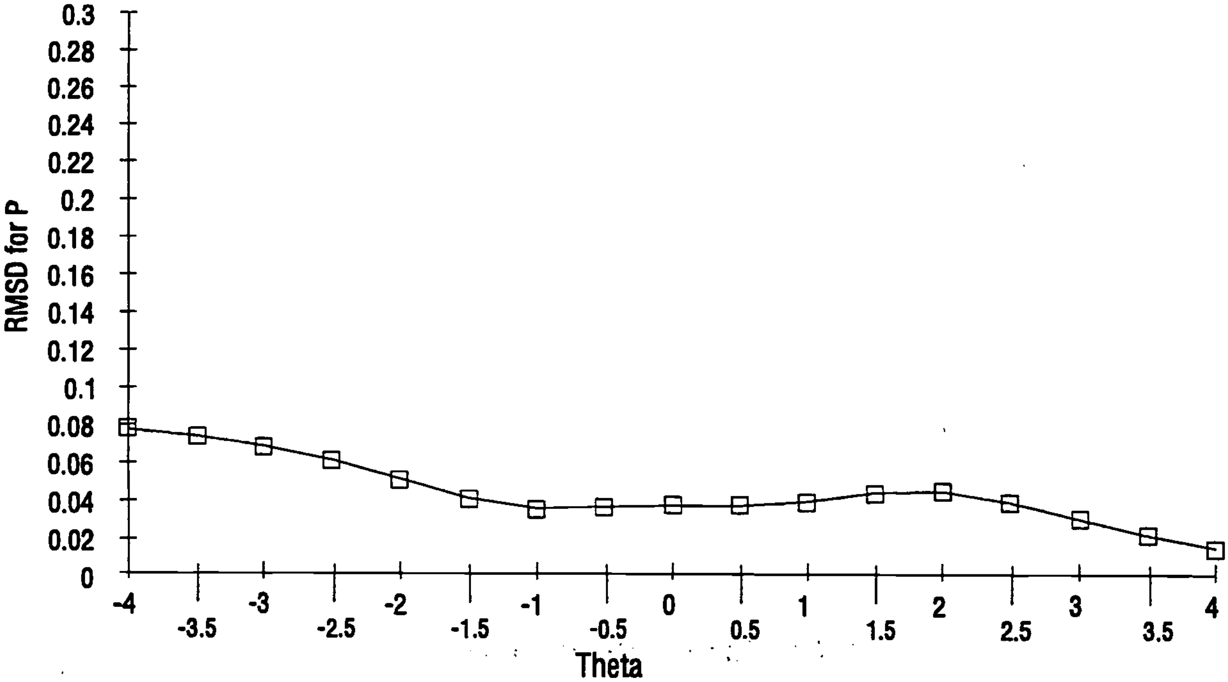
Root Mean Squared Difference and Bias

Graphs of the RMSDΘ at the test level and category level are presented in Figure 4. In comparing these graphs, it is clear that the greatest degree of difference is incurred when the c-parameter is fixed at one administration but not the other. In the graph for this pattern, the RMSDΘ reaches approximately .24 at the lowest end of the Θ scale and then steadily drops off to approximately .03 in the middle of the scale. The values rise slightly from the middle to the high end of the Θ scale, but not to a significant degree. The least difference is incurred when the c-parameter is fixed for both administrations. The values of this statistic for this pattern are at their highest at the lower end of the Θ scale, but only reach a value of approximately .05. At the high end of the Θ scale, this statistic drops to almost 0. The graph of the RMSDΘ with the c-parameter estimated at both administrations falls in between the values for the other two patterns of c-estimation.

Figure 4
RMSD between the Values of P at Θ

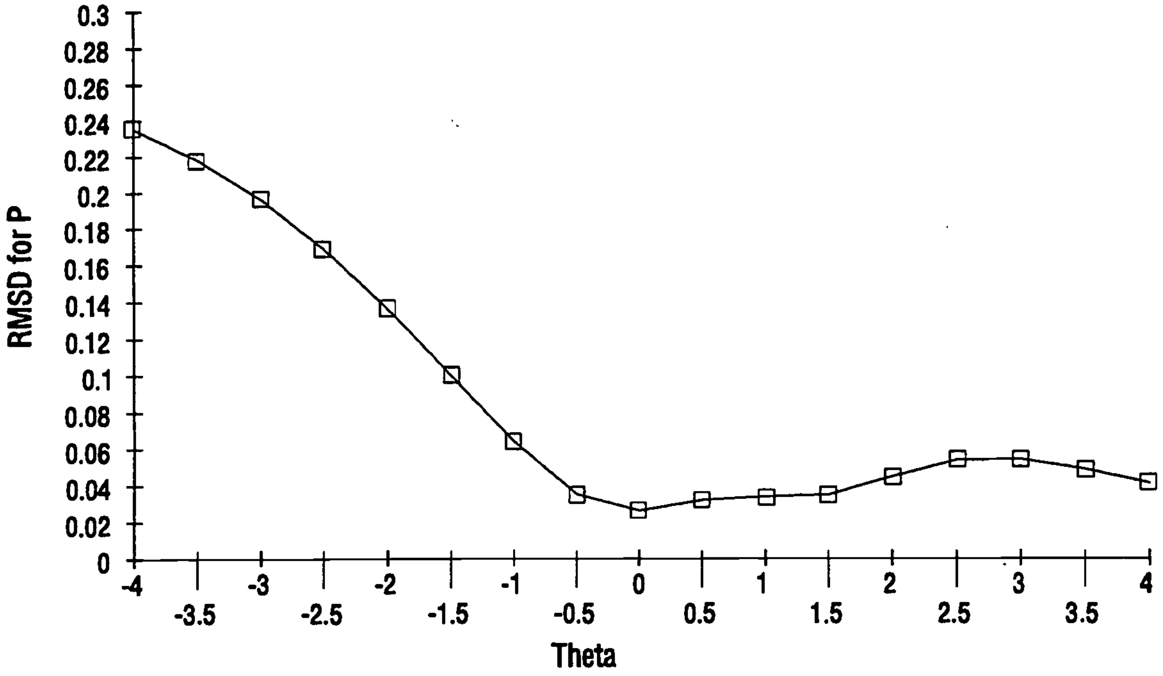


Total Test

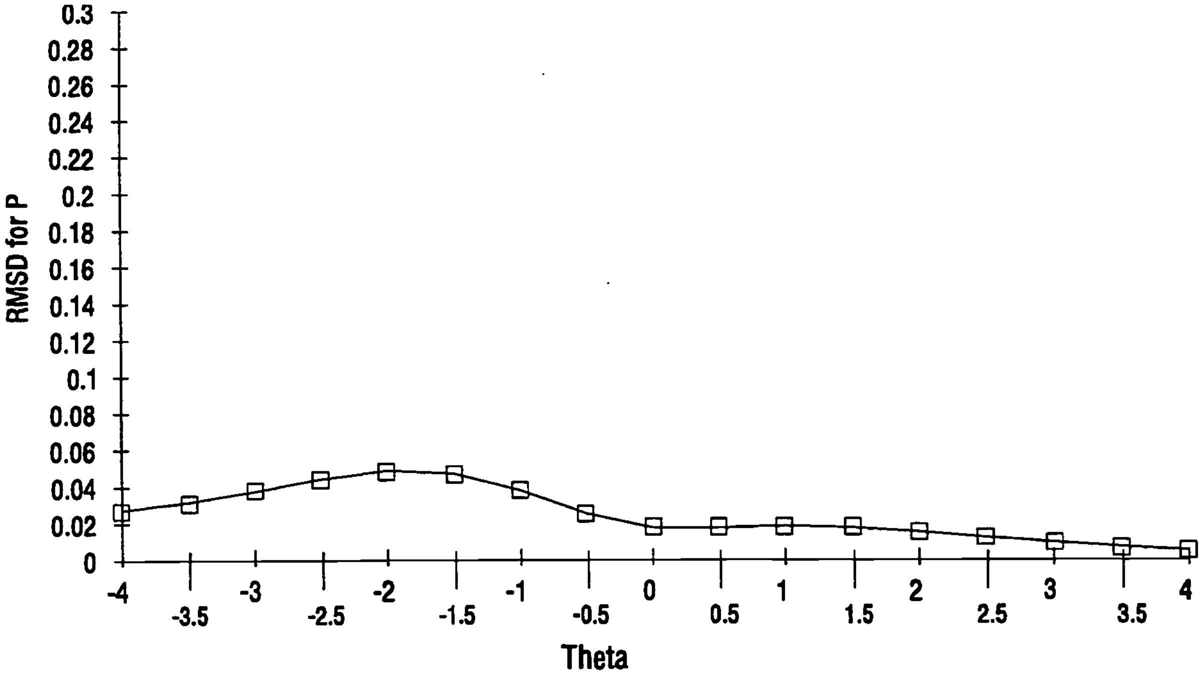


Both

Figure 4 (continued)



One



Neither

Figure 5 represents the bias$_\Theta$ statistic at the test level and the category level. The graphs for the test level, for the c-parameter estimated at both administrations, and for the c-parameter fixed at both administrations are extremely similar with values hovering around 0 for most of the Θ scale. For both the total test and for items having the c-parameter fixed for both administrations, the bias$_\Theta$ statistic is slightly positive at the low end of the scale, reflecting slightly higher values of P(Θ) for the June administration. Conversely, the values of the bias statistic for items having the c-parameter estimated at both administrations dip slightly below 0 at the higher end of the scale, representing higher values of P(Θ) for the December administration. However, this dip appears to be negligible.

Clearly, the highest degree of bias is incurred for items for which the c-parameter is fixed at only one administration. This statistic reaches approximately .125 at the lower end of the Θ scale, drops to 0 at the middle of the scale, and increases slightly to approximately .025 at the higher end of the scale. This result indicates that when the c-parameter was fixed at only one administration, the values of P(Θ) tended to be higher for the June administration.

Item Characteristic Curve Overlay Plots

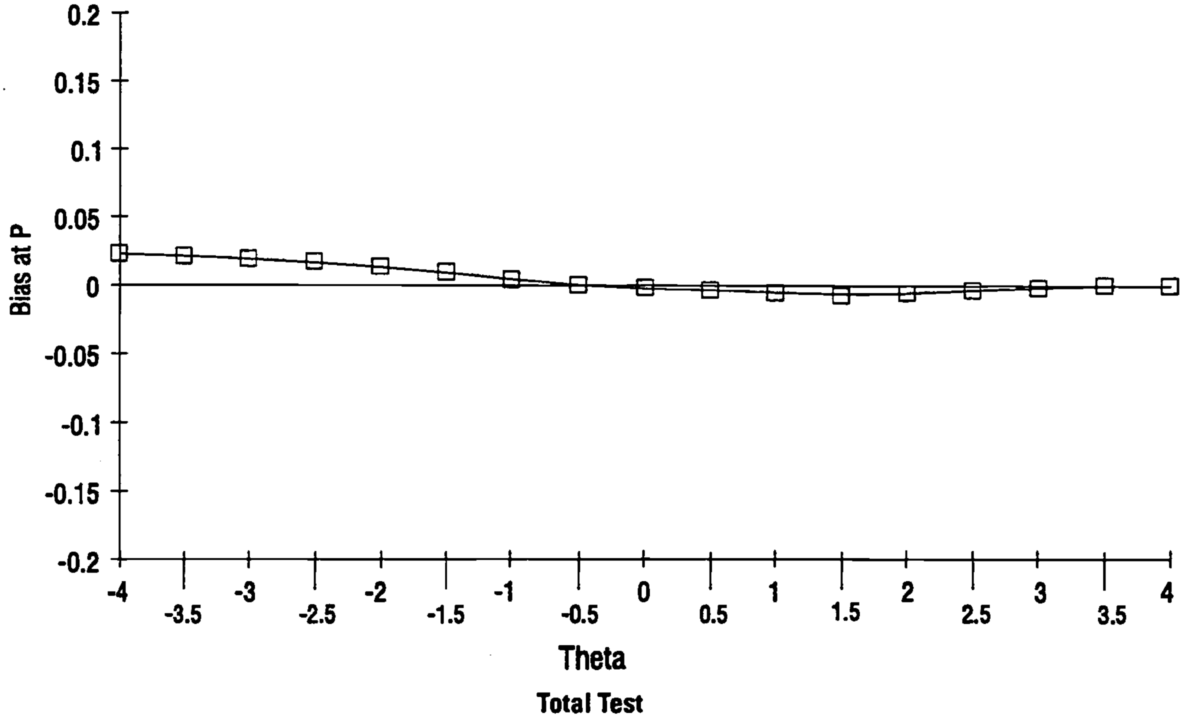
In order to further explore the effect of these various patterns of c-estimation, item characteristic curves were compared for each item. These plots are presented in Figures 6 to 10. The numbers assigned to these items follow the coding scheme described in the Methodology section, with the letter representing the item set and the number representing the position of the item within the set. An attempt was made to detect any patterns that may exist among the items within a particular pattern of c-estimation. These results are summarized as follows.

C estimated at both administrations. For the 25 items for which the c-parameter was estimated for both administrations, the c-values were similar for four items, numbers A5, A6, A10, and C8 presented in Figure 6. For these four items, the a- and b-parameter estimates behaved similarly, with similar a- and b-values for both administrations for item A10, higher June a- and b-values for item A6, and higher December a- and b-values for items A5 and C8.

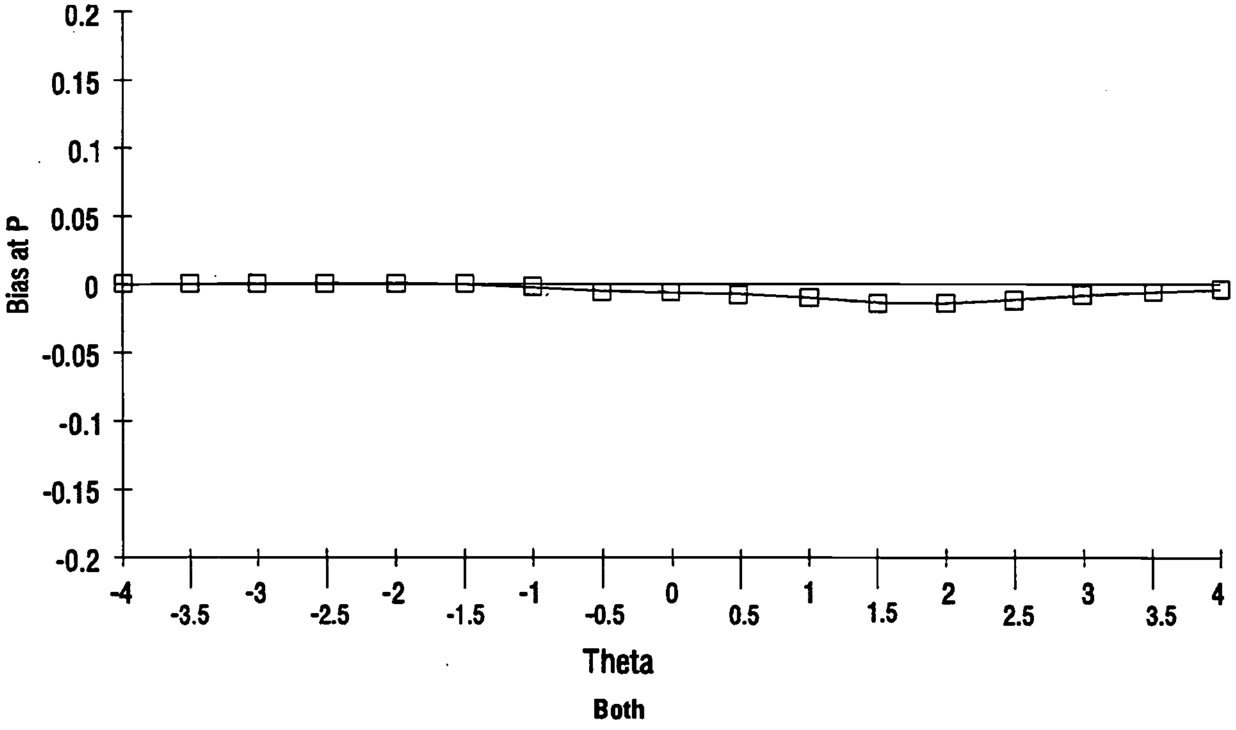
For this same pattern of c-estimation, the June c-values were higher for 12 items, numbers A3, A8, B5, B7, B8, C4, C5, C10, D1, D3, D4, and D7 presented in Figure 7. With the exception of items A3 and D7 for which the a-values were similar for both administrations, the a-values for these items were higher for June than they were for December. Again, the b-parameter estimates behaved in a manner similar to the a-parameter estimates, but showed slightly more stability. The b-values were similar for four items, numbers A8, C4, C5, and D1, but the June b-values were higher for the other eight items.

Finally, for 9 items, the December c-values were higher. This was observed for items B1, B3, B4, B6, B9, C3, C6, C9, and D5 presented in Figure 8. Similar to the items for which the June c-values were higher, the a-parameter estimates for these nine items were similar for three of the items, numbers B4, B6 and C3, but higher in December for the remainder of the items. Again the b-parameter estimates are somewhat more stable, with similar b-values for items B1, B4, B6, and C3, a higher June b-value for item C6, and the remainder of the items having a higher b-value for the December administration. These observations seem to indicate that when the c-parameter is estimated, a higher value of the c-parameter is indicative of a higher value of the a-parameter. A similar effect is observed for the b-parameter estimates, but the effect upon this parameter is not quite as strong.

Figure 5
Bias between the Values of P at Θ

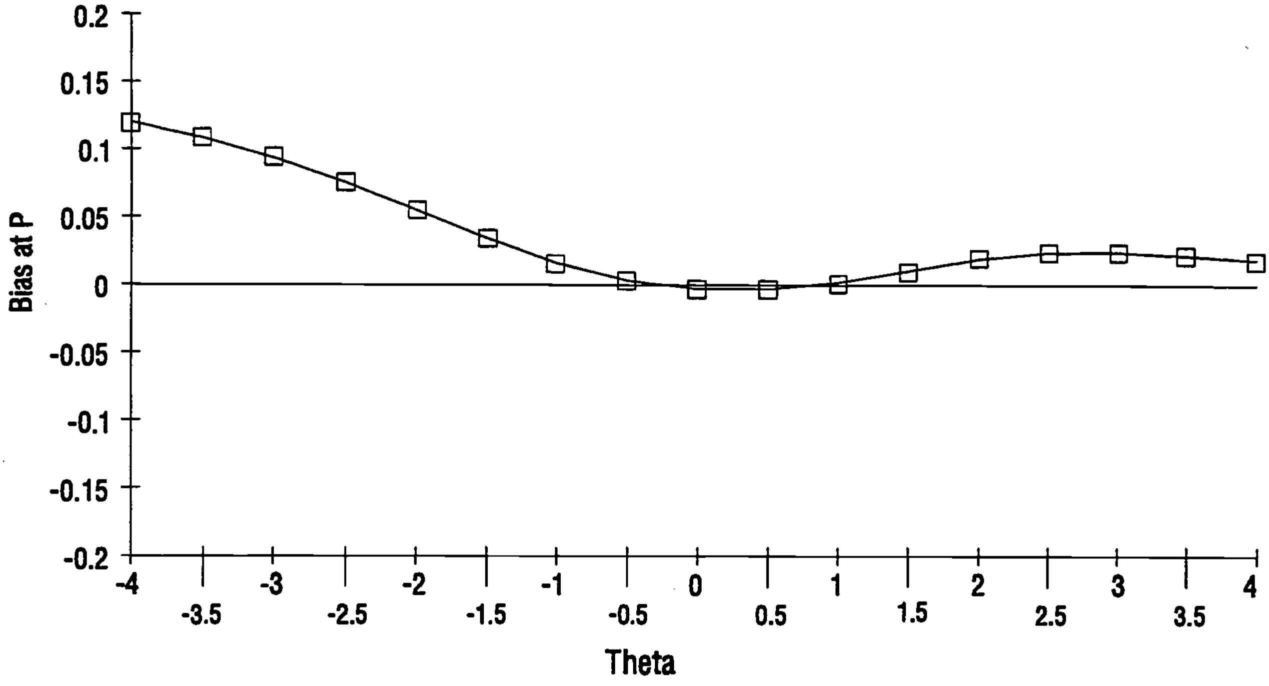


Total Test

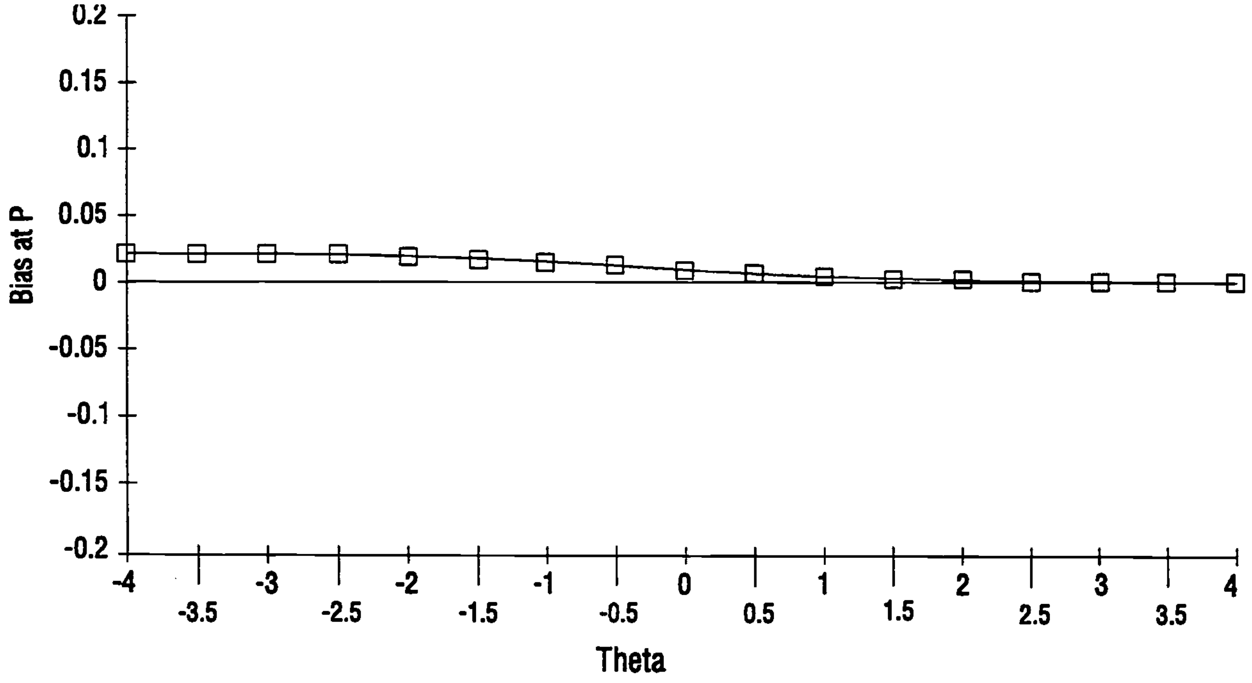


Both

Figure 5 (continued)

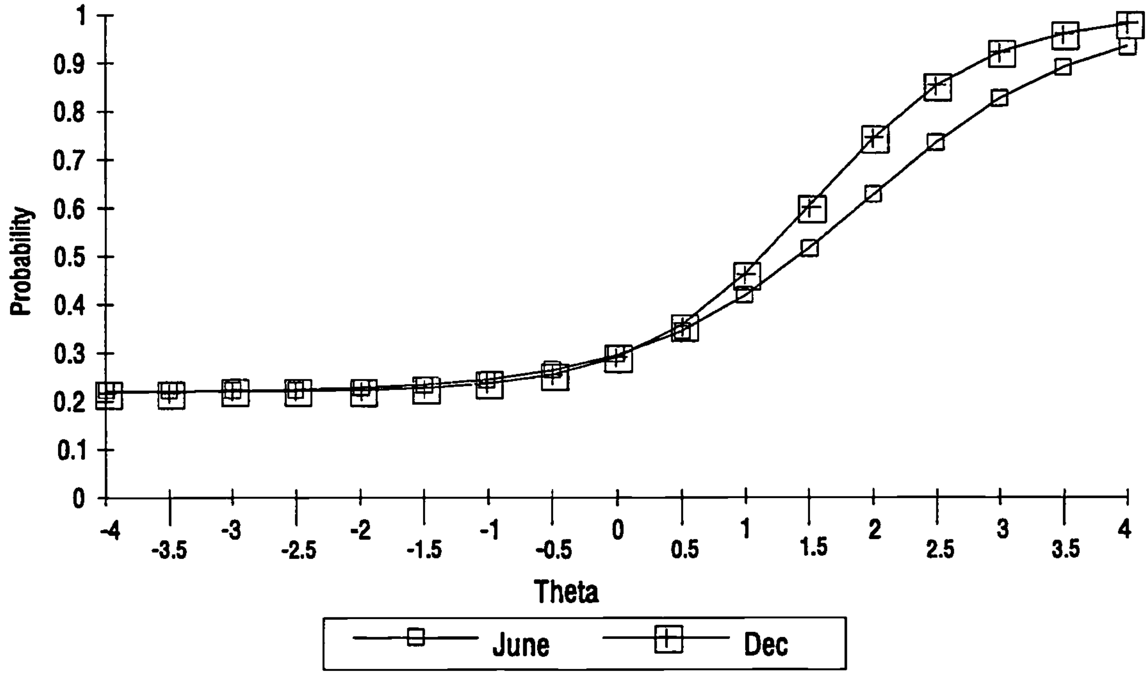


One

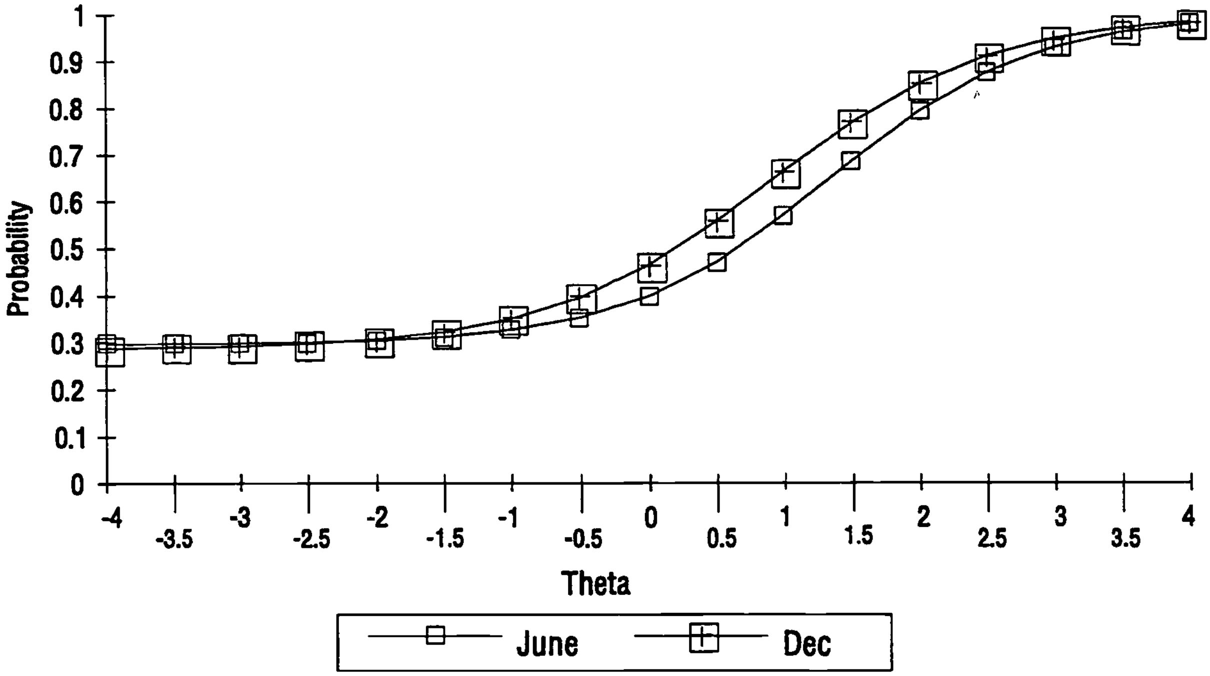


Neither

Figure 6
C Estimated at Both Administrations—Similar C-Values

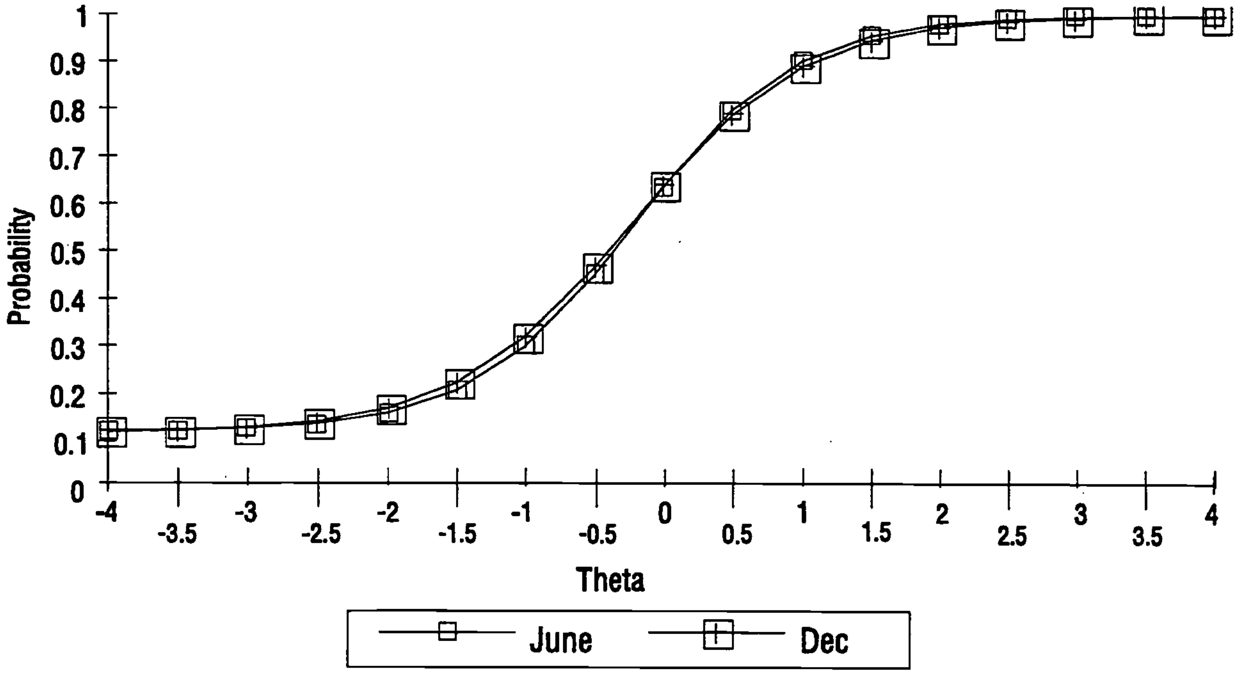


Item A5

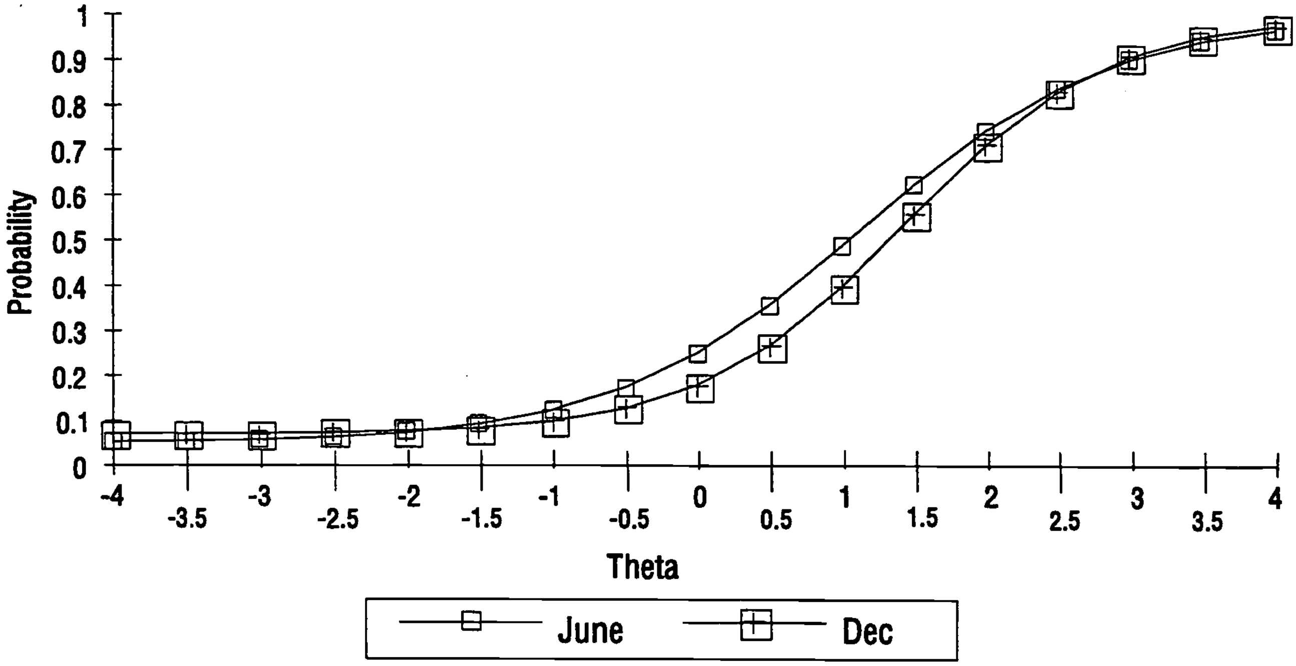


Item A6

Figure 6 (continued)

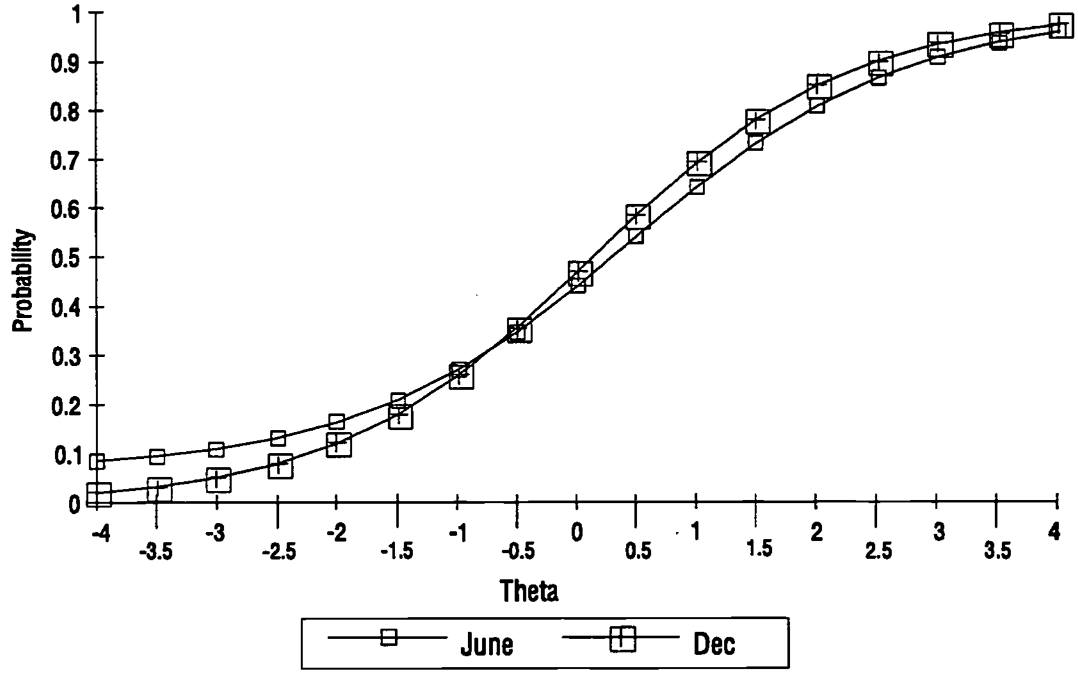


Item A10

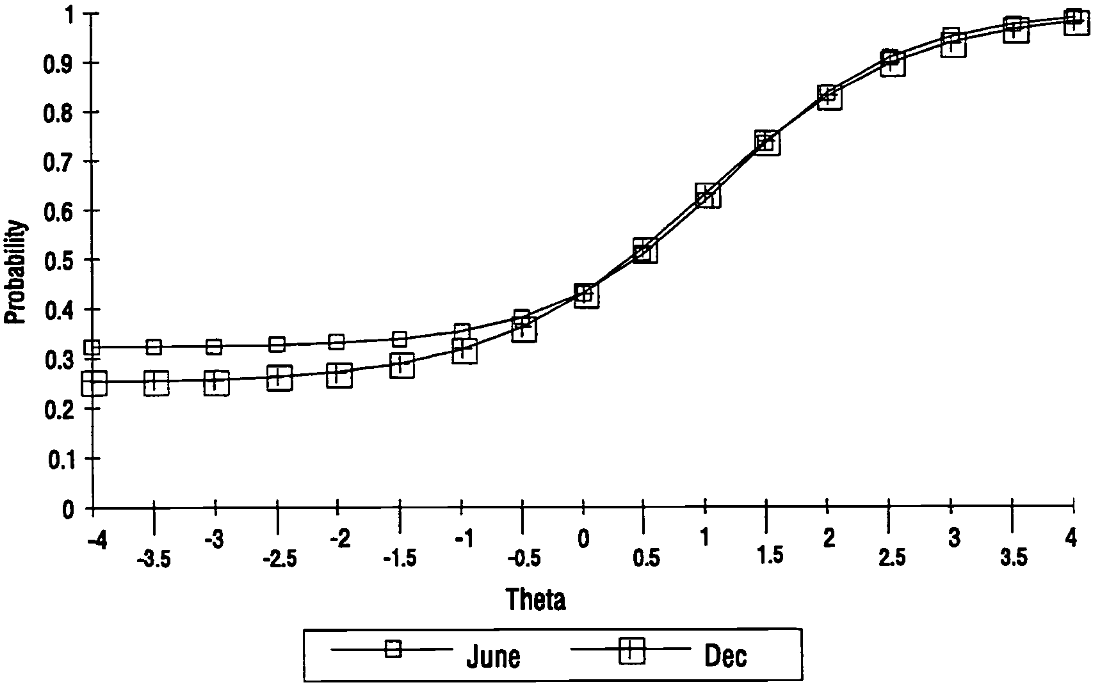


Item C8

Figure 7
C Estimated for Both Administrations—June C-Value Higher

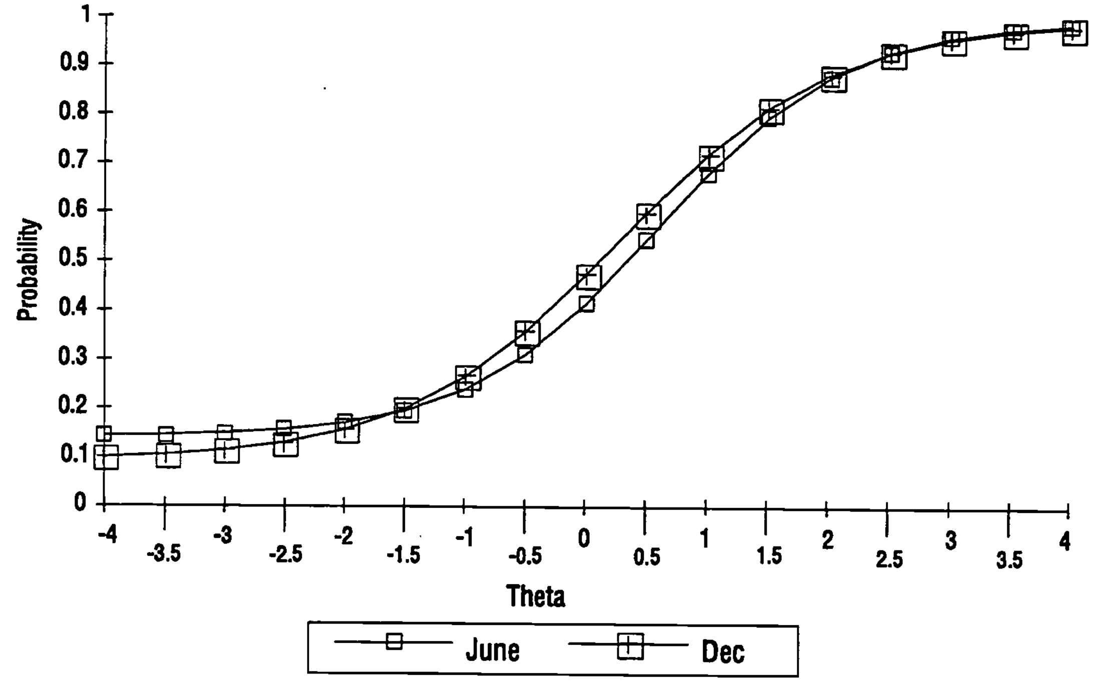


Item A3

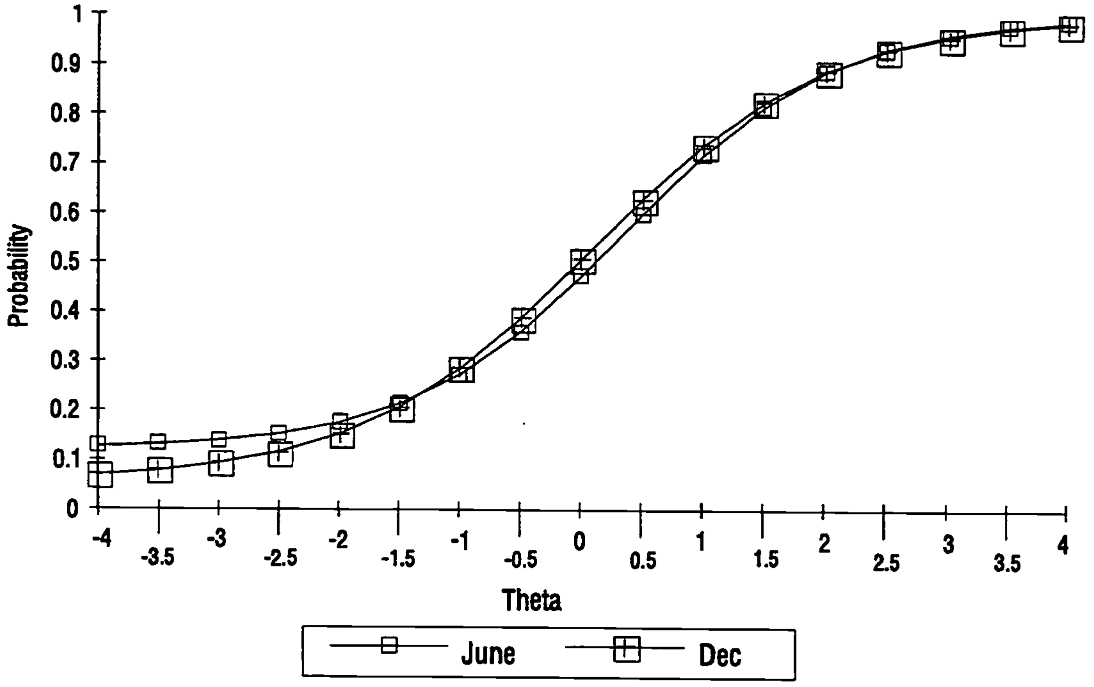


Item A8

Figure 7 (continued)

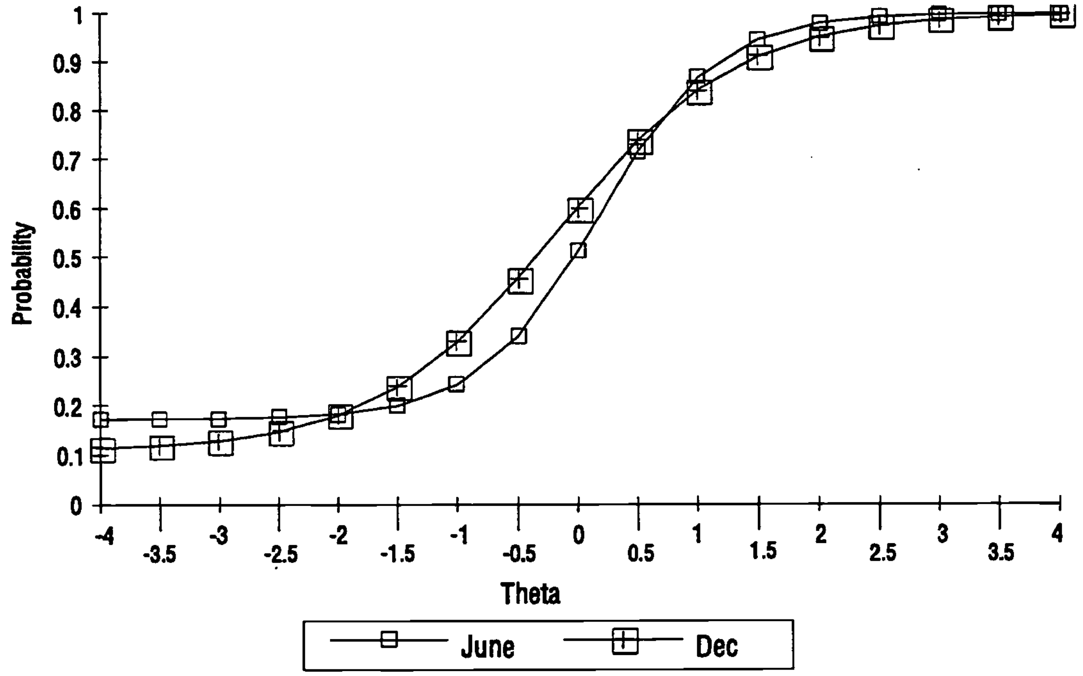


Item B5

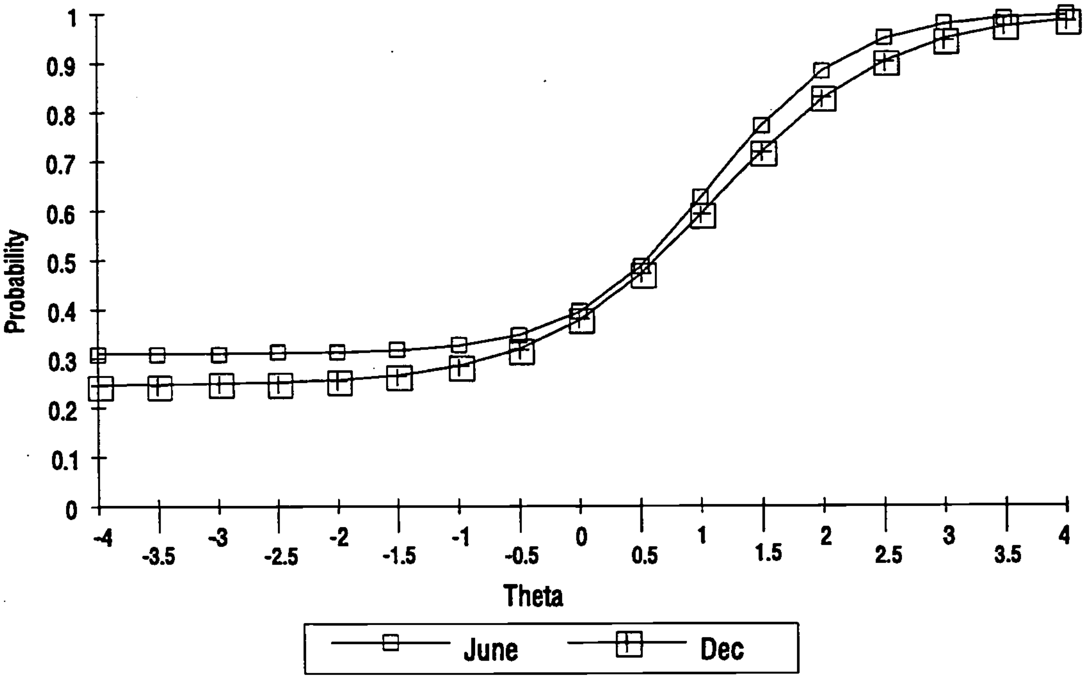


Item B7

Figure 7 (continued)

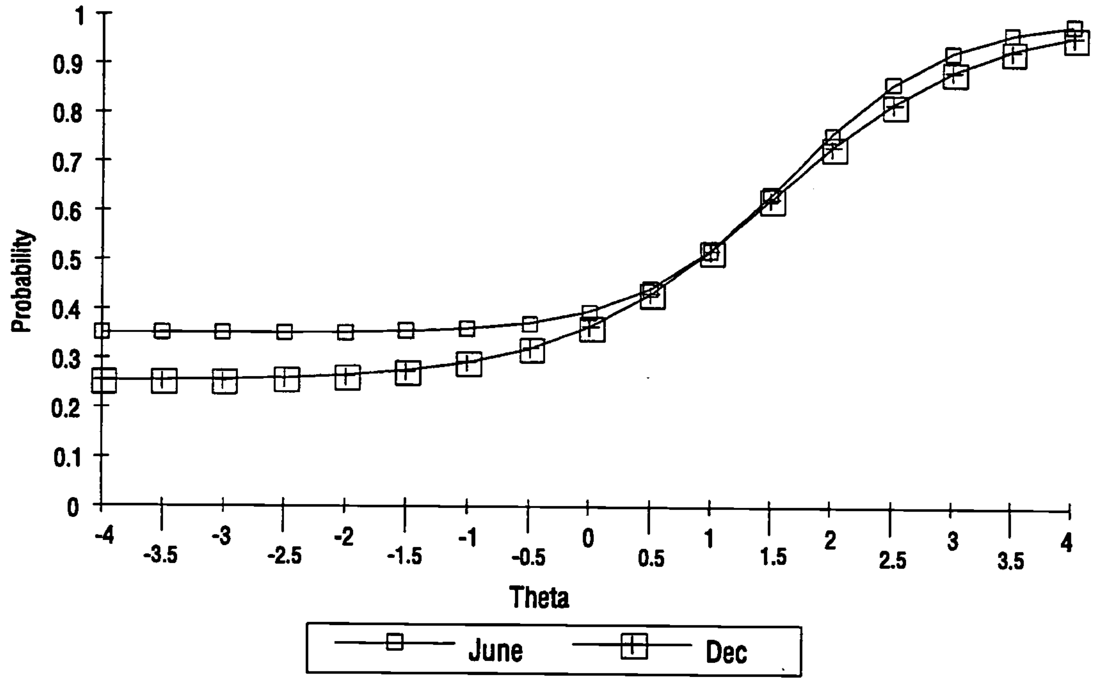


Item B8

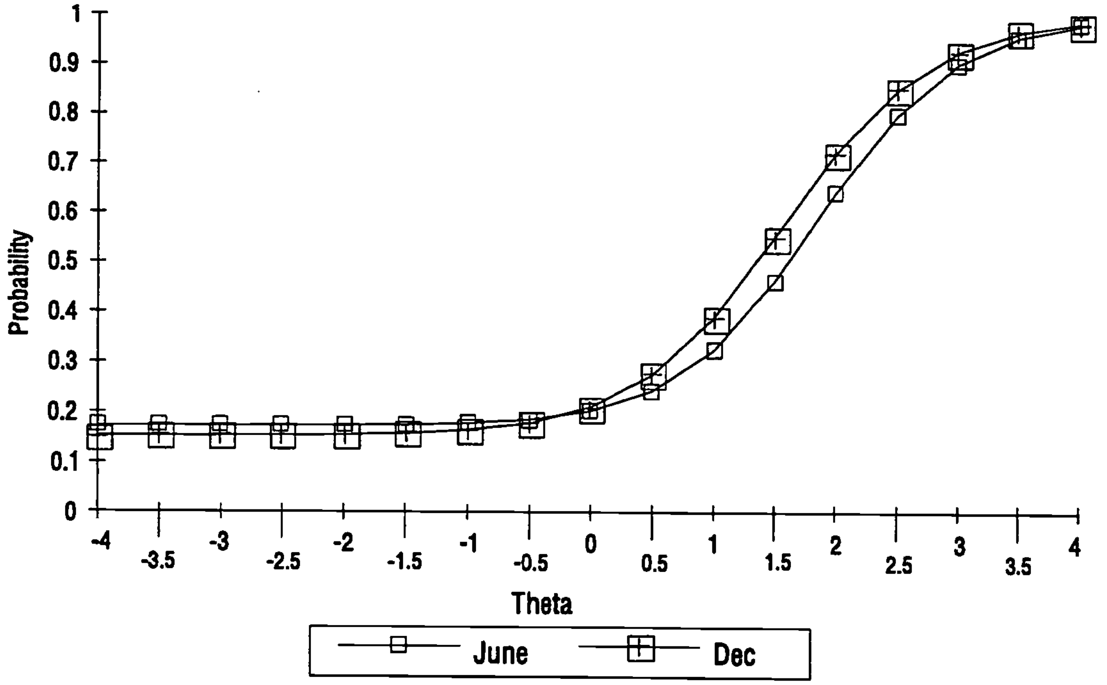


Item C4

Figure 7 (continued)

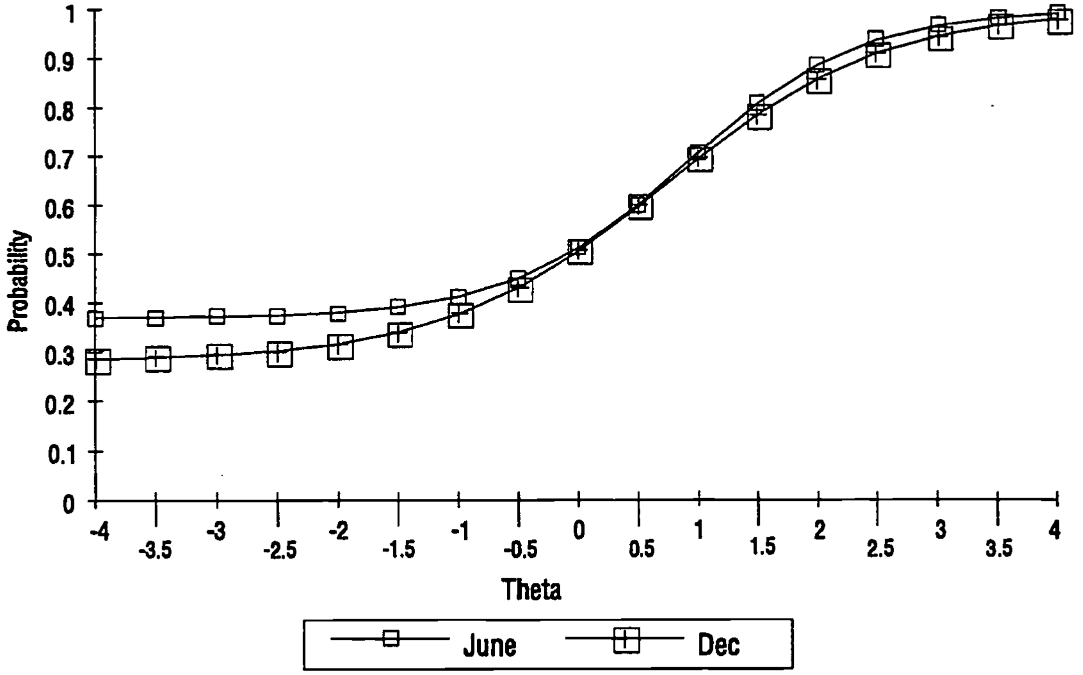


Item C5

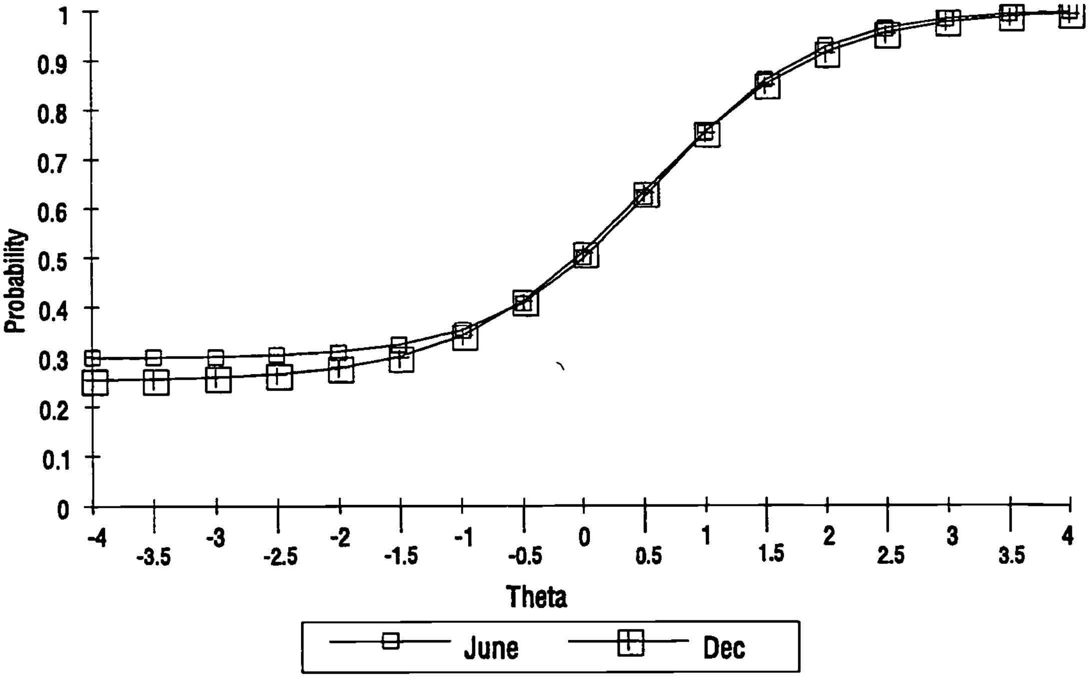


Item C10

Figure 7 (continued)

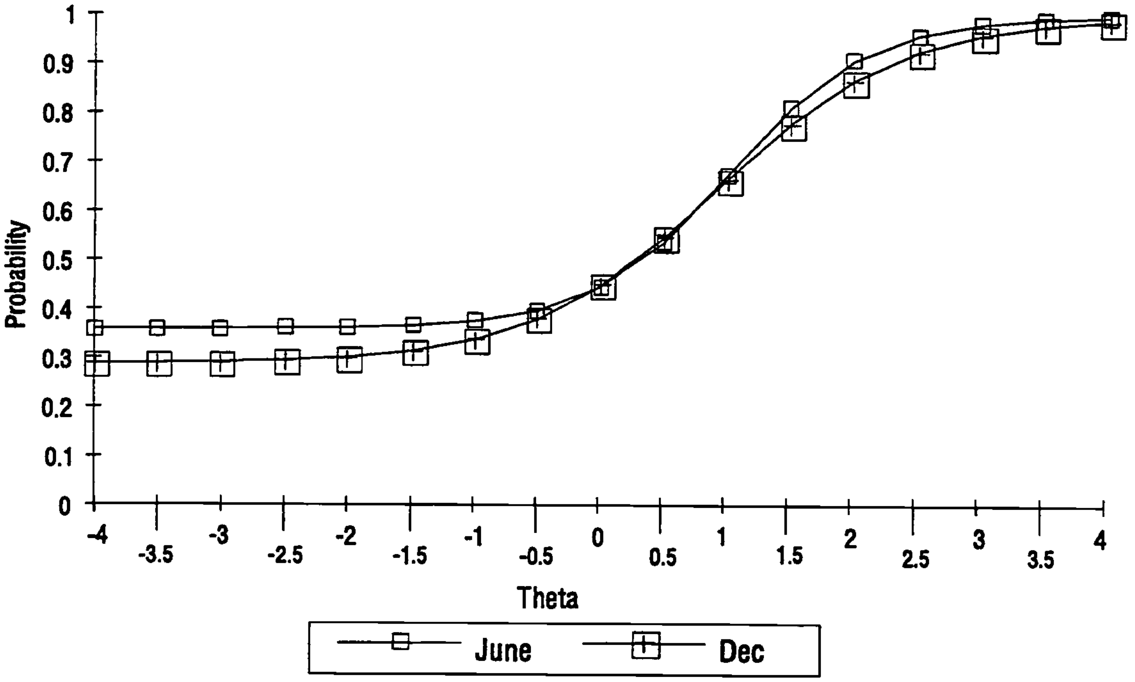


Item D1

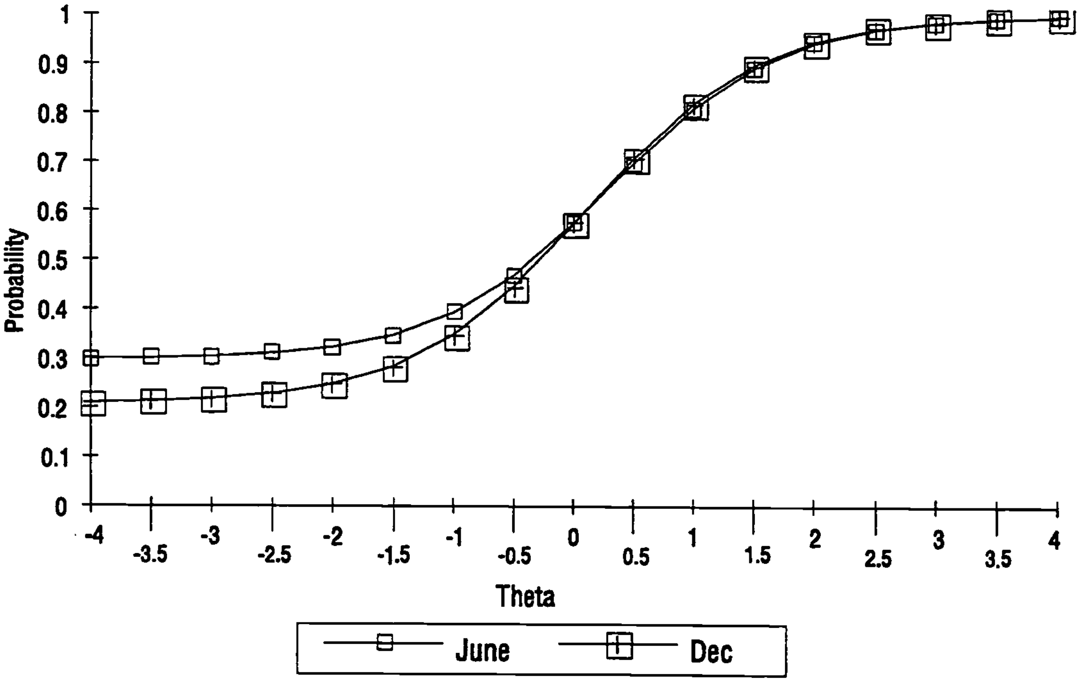


Item D3

Figure 7 (continued)

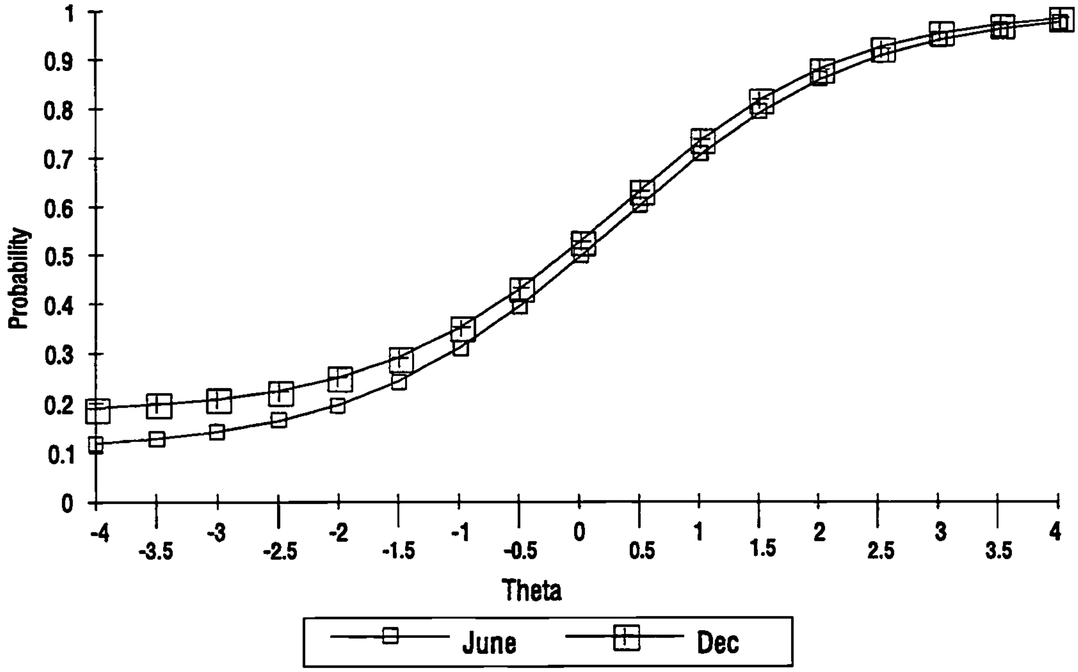


Item D4

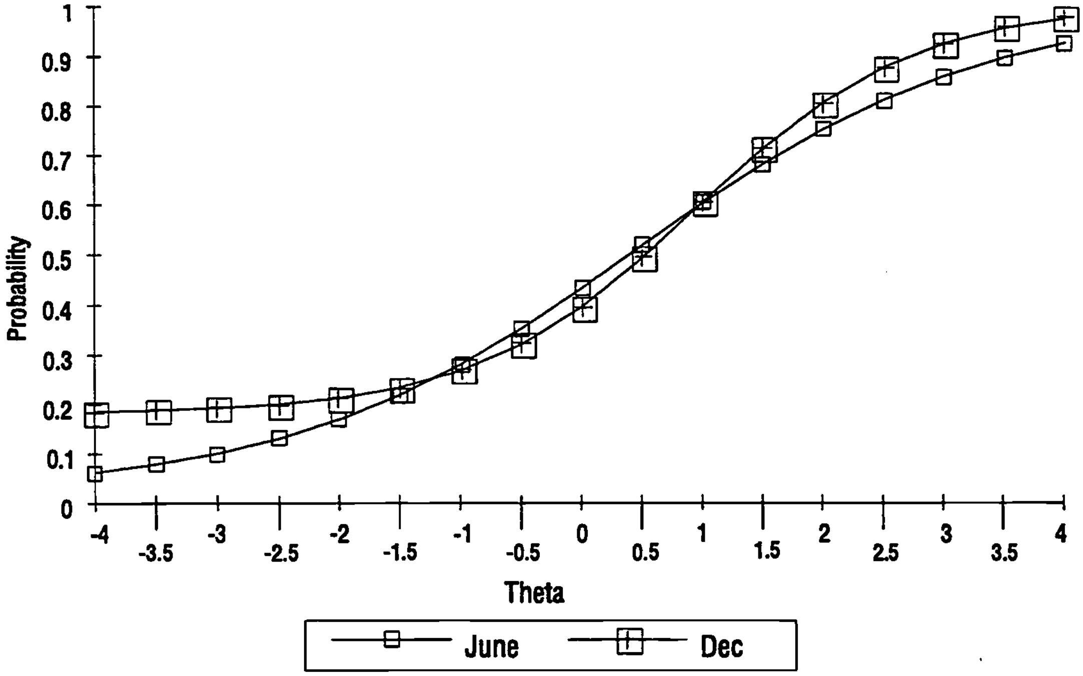


Item D7

Figure 8
C Estimated for Both Administrations—December C-Value Higher

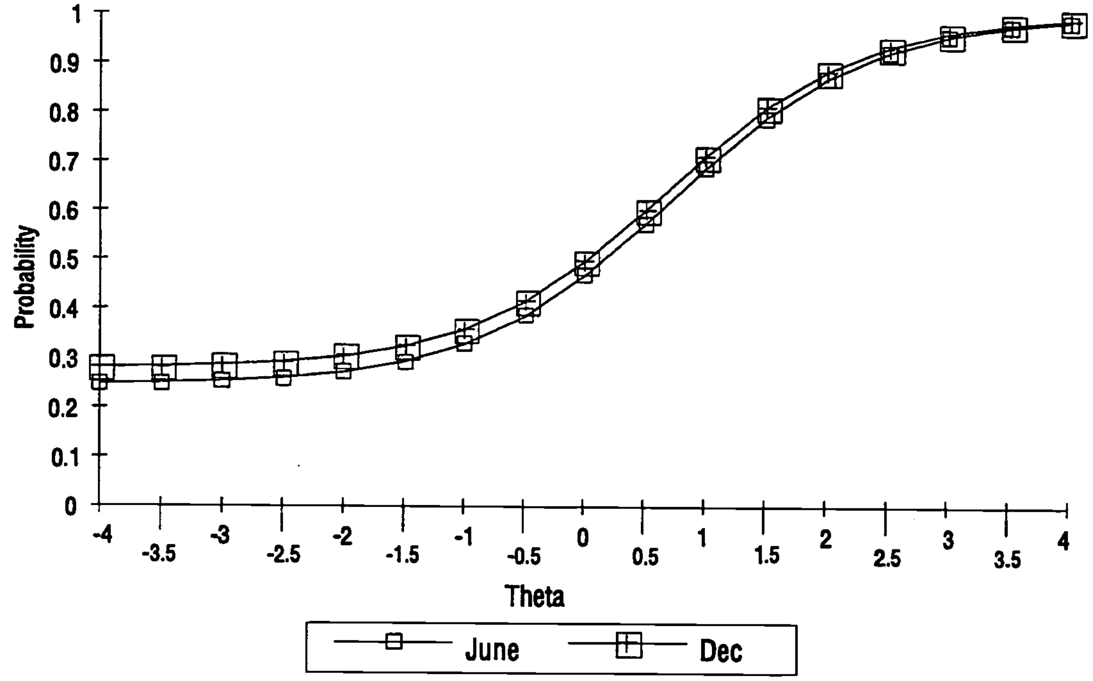


Item B1

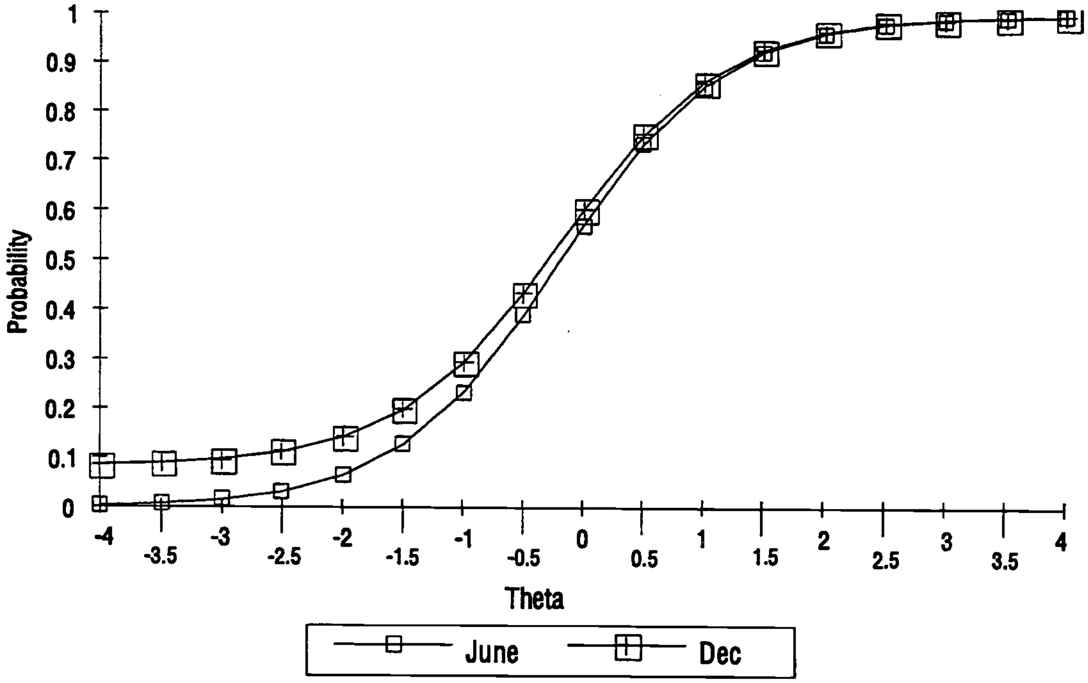


Item B3

Figure 8 (continued)

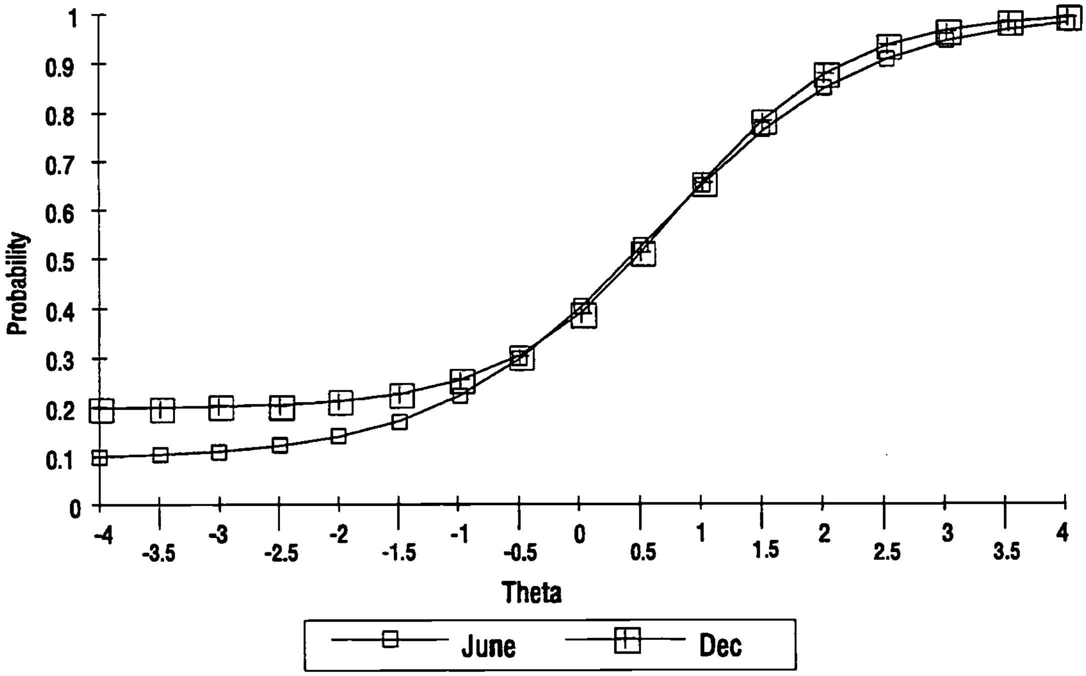


Item B4

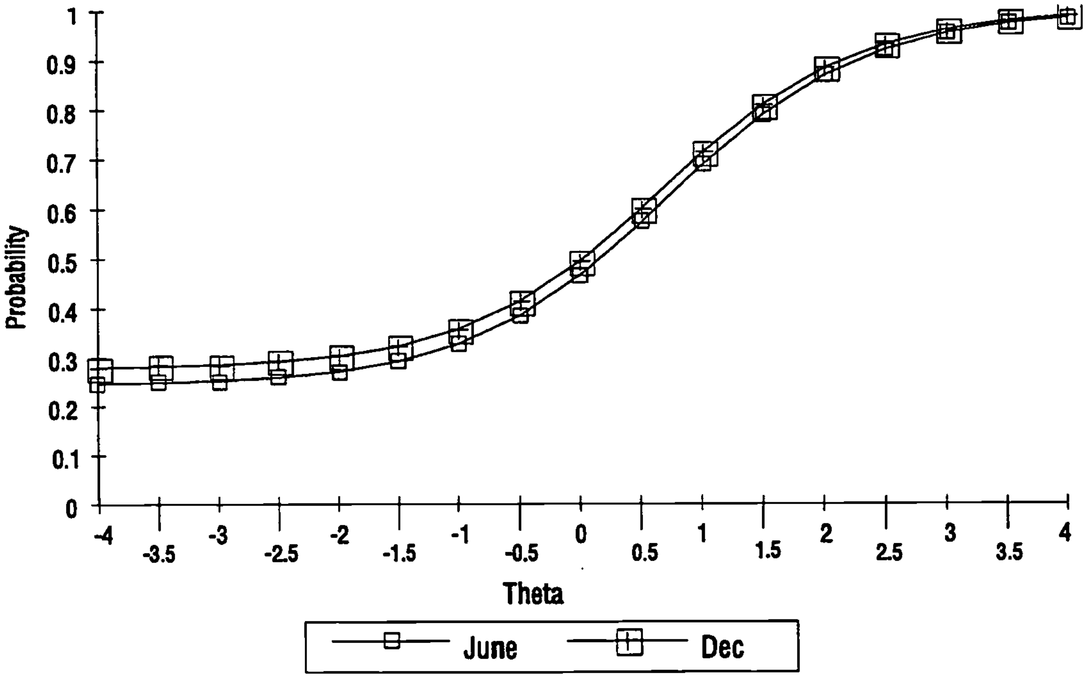


Item B6

Figure 8 (continued)

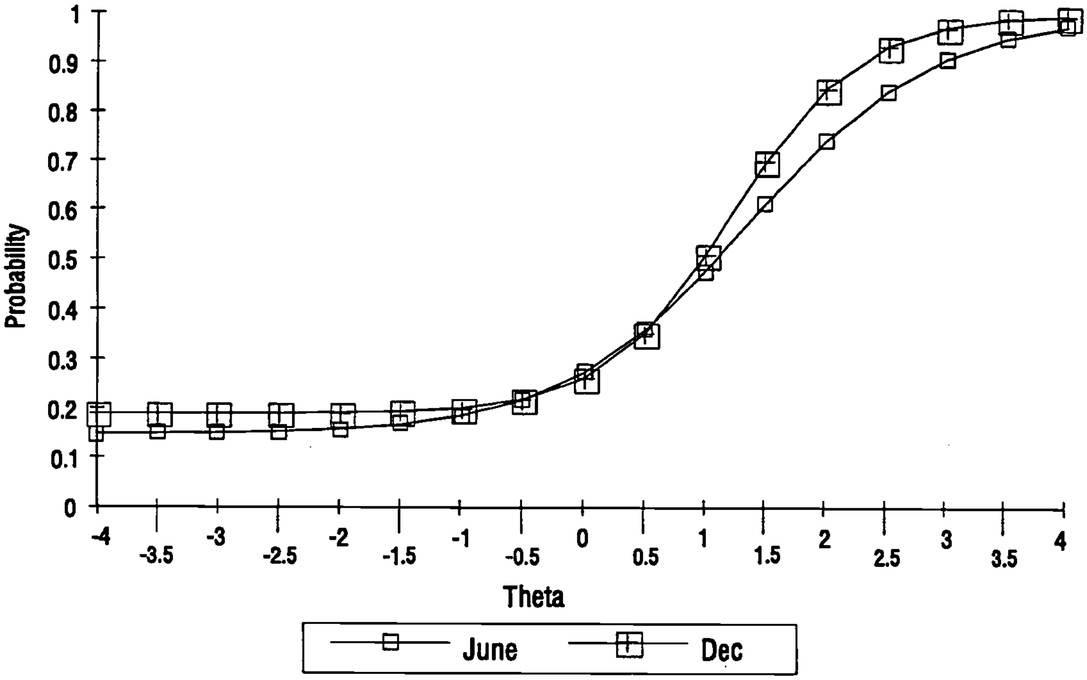


Item B9

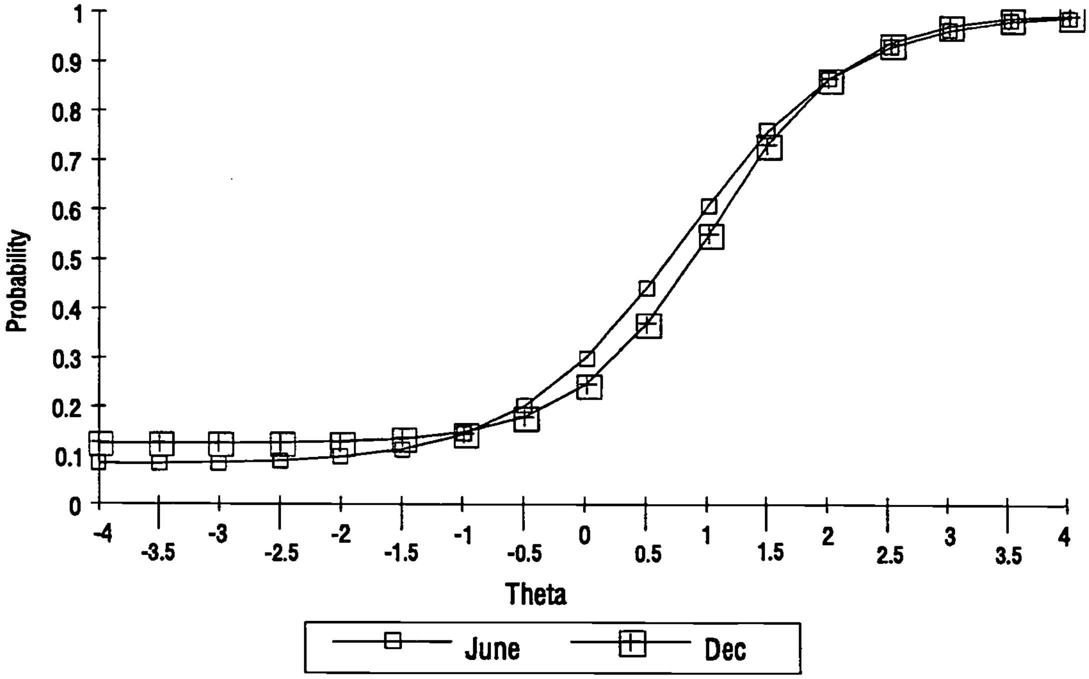


Item C3

Figure 8 (continued)

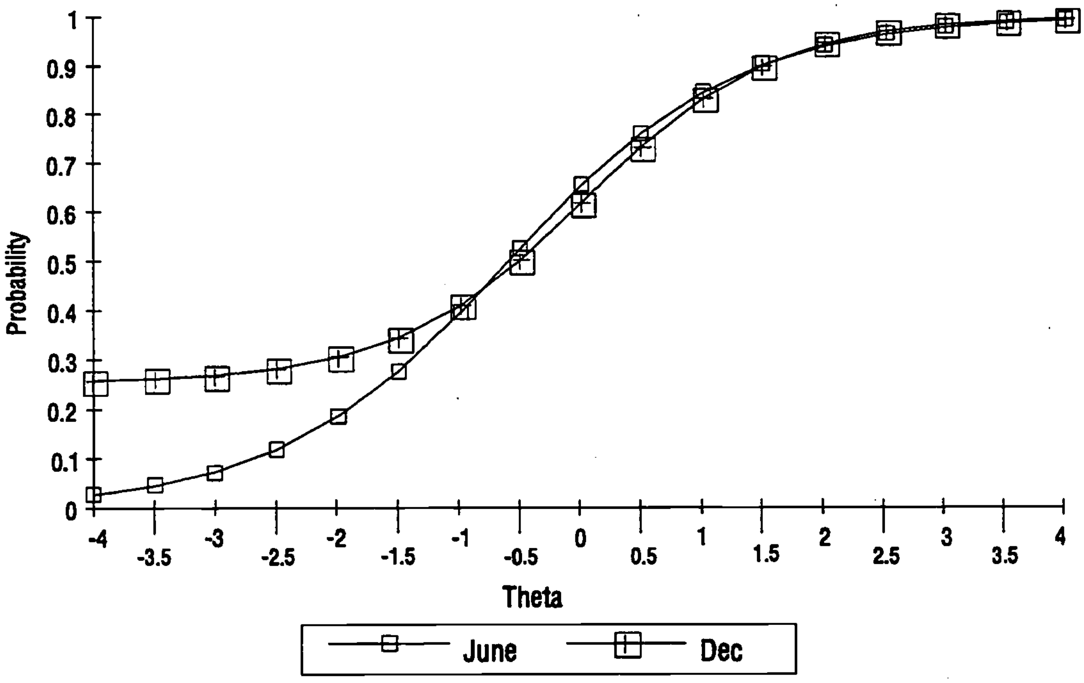


Item C6



Item C9

Figure 8 (continued)



Item D5

C fixed at one administration. The 6 items for which the c-parameter was fixed in one administration but not the other are presented in Figure 9. Here, only item A4 had a fixed c-value in June but not in December. For this item, the a-, b- and c-values were all higher for December than for June.

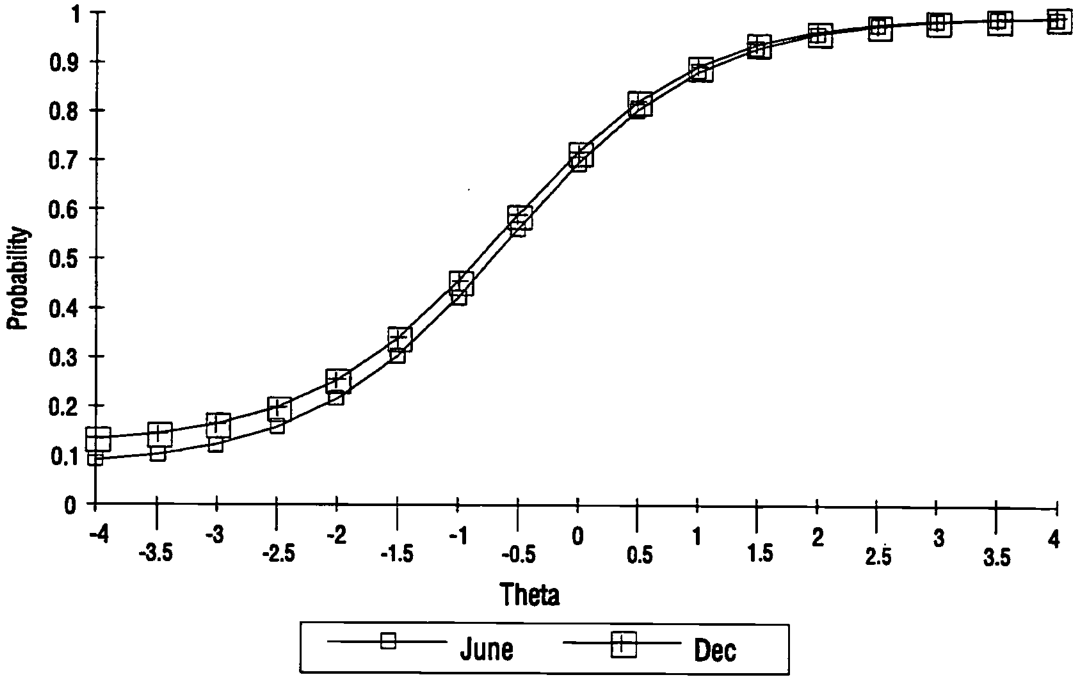
For the 5 items having a fixed c-value in only the December administration, only one item, number A2, had a higher c-value for December than for June. For this item, the a- and b-values were similar for both administrations. For the remaining items, numbers C7, D6, D8, and D9, the June a-, b- and c-values were higher. These observations seem to imply that if the c-parameter is fixed in one administration, the a-, b- and c-values in the other administration all have a tendency to be higher.

C fixed for both administrations. Finally, the 9 items for which the c- parameter was fixed for both administrations are items A1, A7, A9, B2, B10, C1, C2, D2, and D10, presented in Figure 10. For the majority of these items, numbers A7, A9, B2, C2, and D2, the a-values were similar for the two administrations. For two items, A1 and D10, the June a-values were higher and for two items, C1 and B10, the December a-values were higher. Interestingly, the b-parameter estimates followed the same pattern as the a-parameter estimates, with the exception of item D10 for which the b-parameter estimates were similar, but the June a-values were higher. Overall, it seems clear that the greatest agreement between the item parameters is achieved when the c-values are fixed for both administrations.

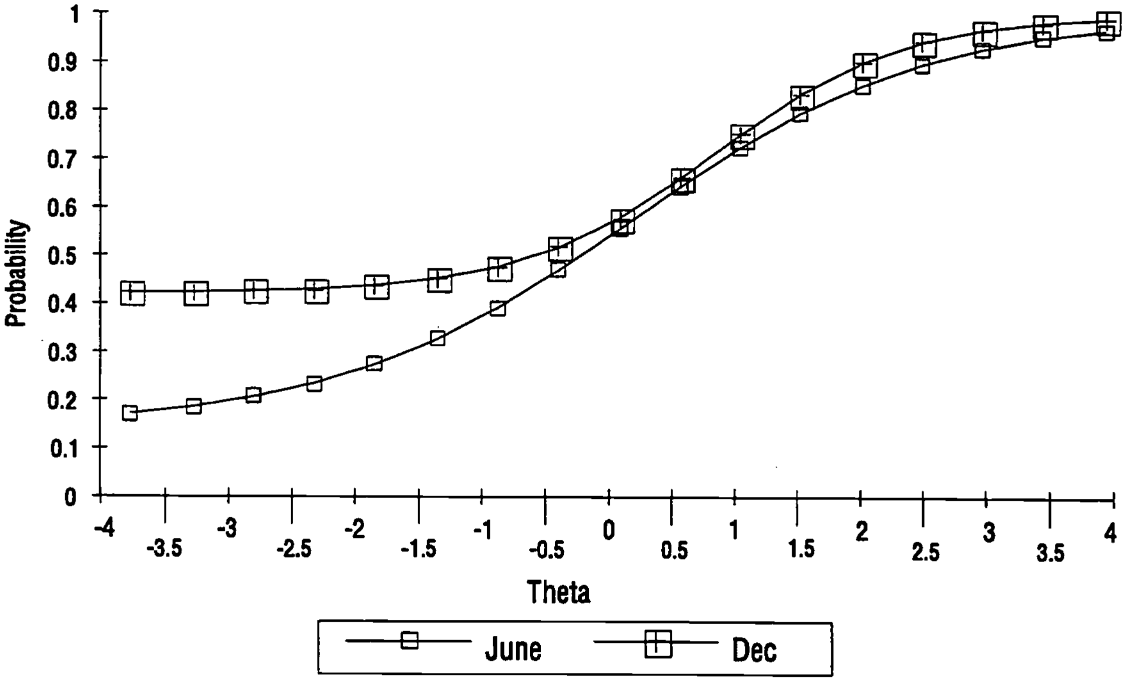
Test Characteristic Curves

The test characteristic curves for the two administrations, presented in Figure 11, are encouraging. These graphs imply that the differences between the a- and b-parameters at the item level average out at the test level. The c-parameter estimates do not average out to the same degree, but the differences observed at the item level are greatly minimized at the test level.

Figure 9
C Estimtated for Only One Administration

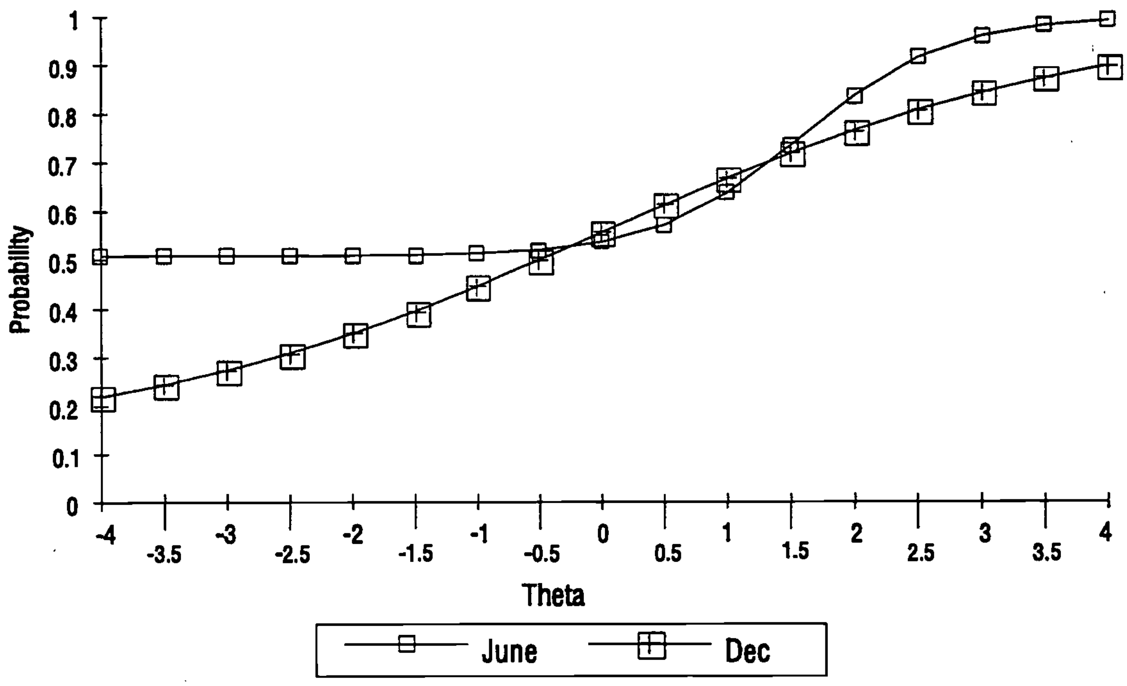


Item A2

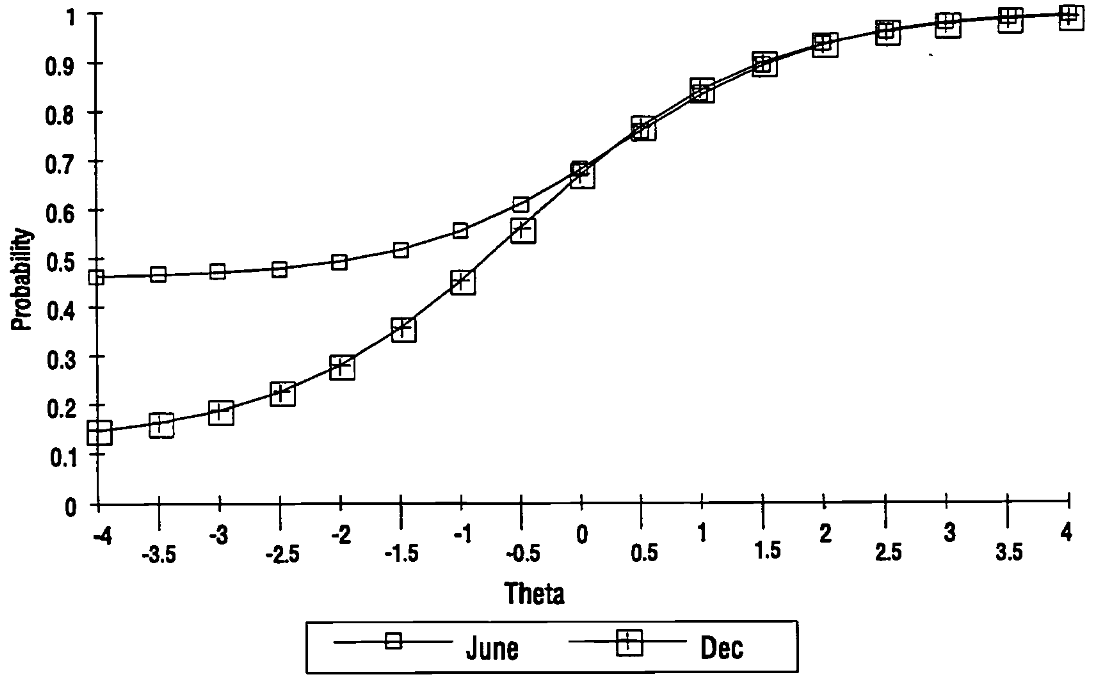


Item A4

Figure 9 (continued)

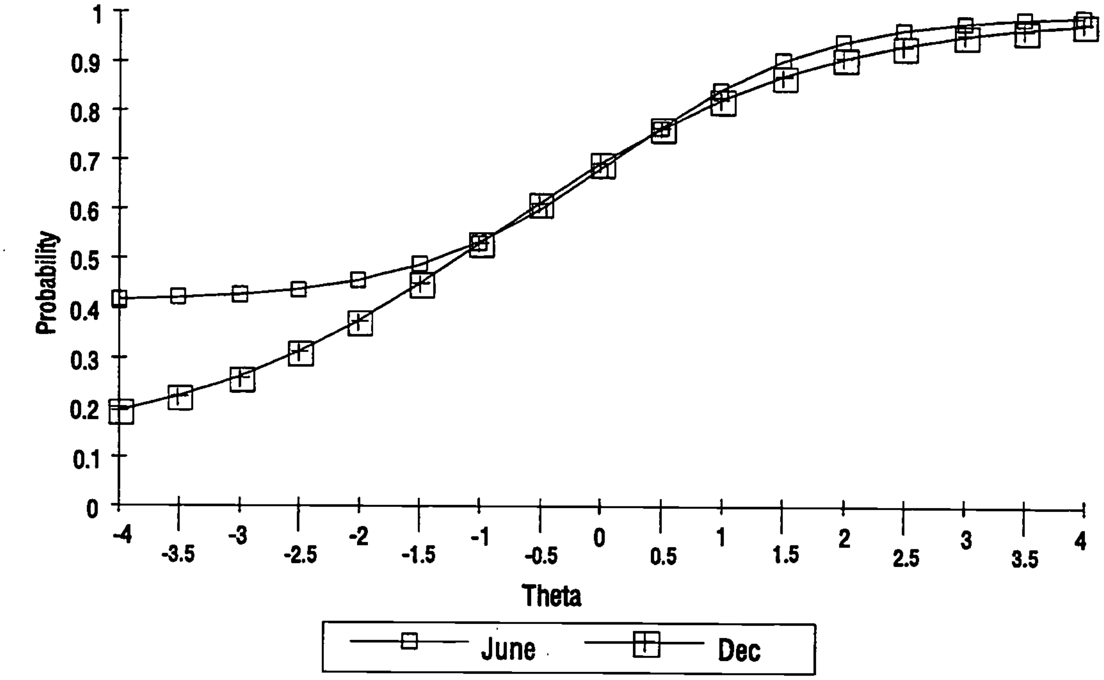


Item C7

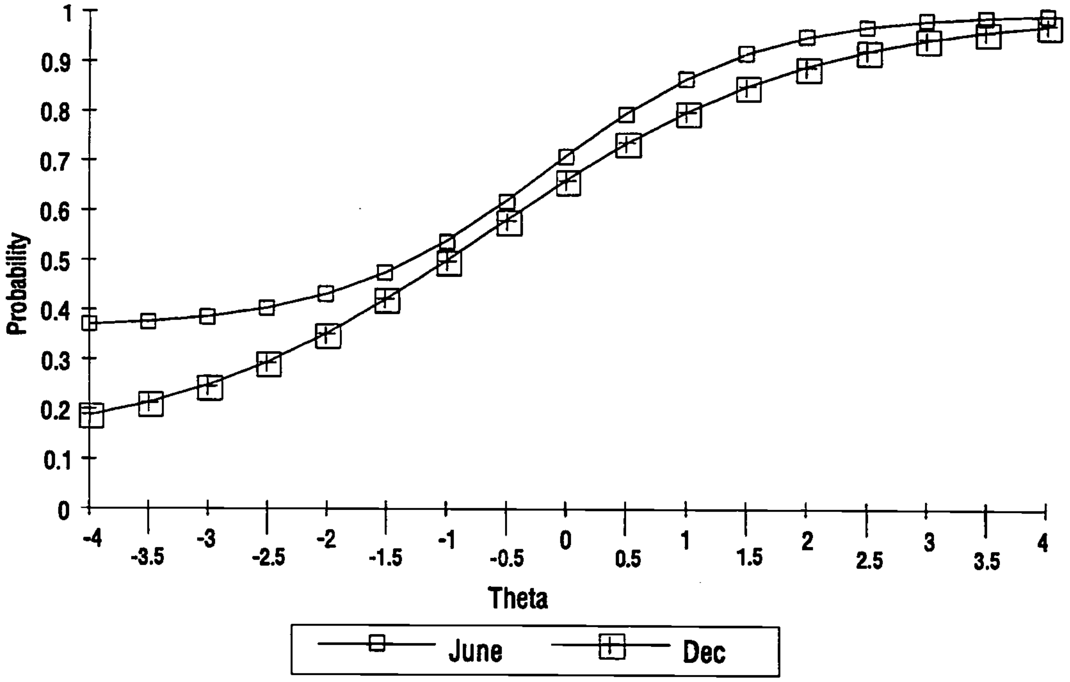


Item D6

Figure 9 (continued)

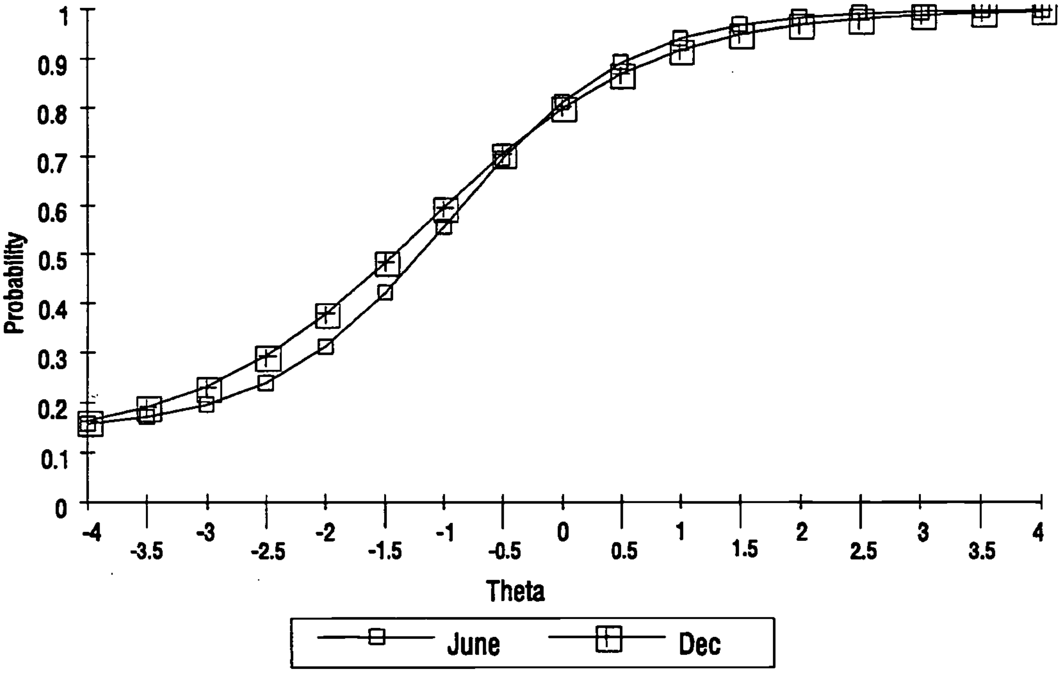


Item D8

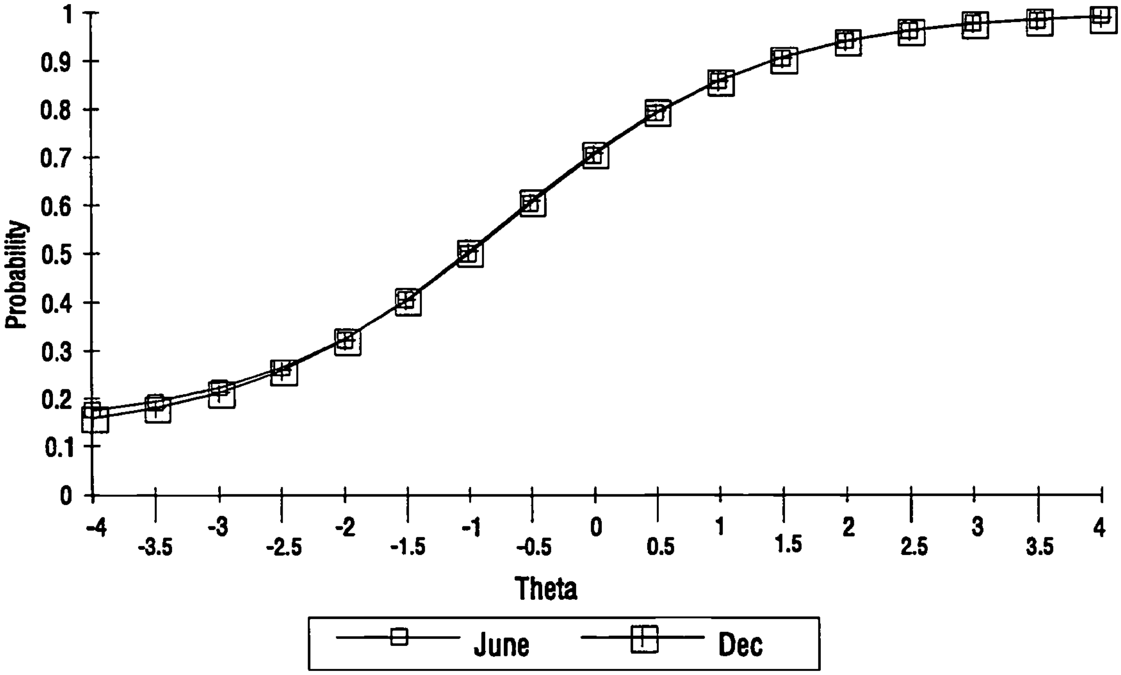


Item D9

Figure 10
C Estimated for Neither Administration

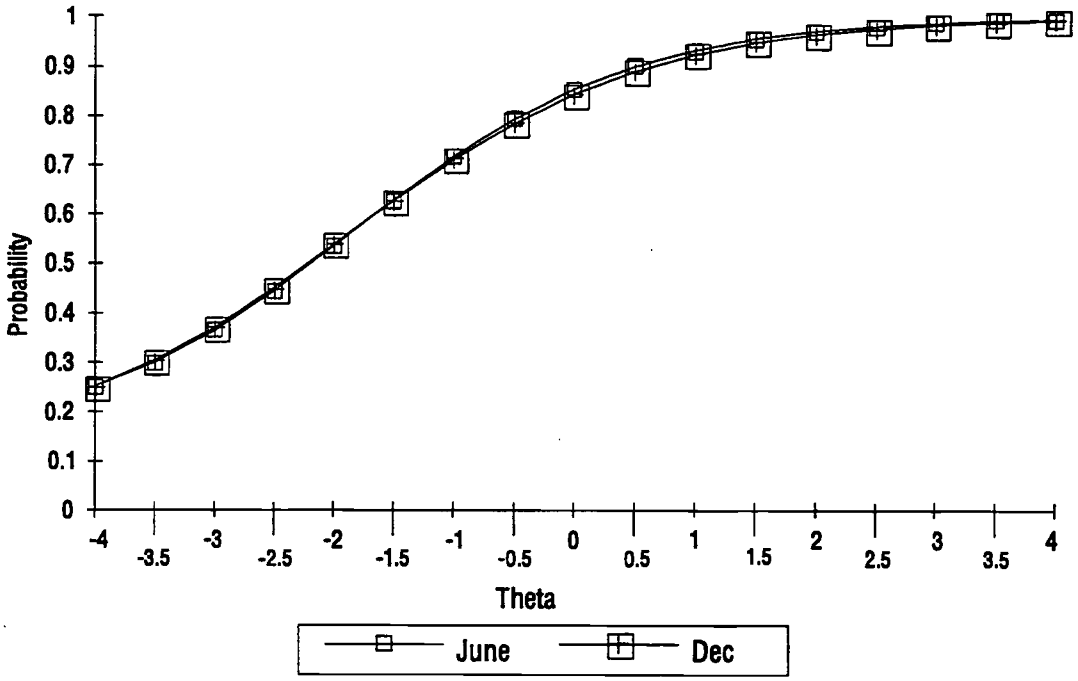


Item A1

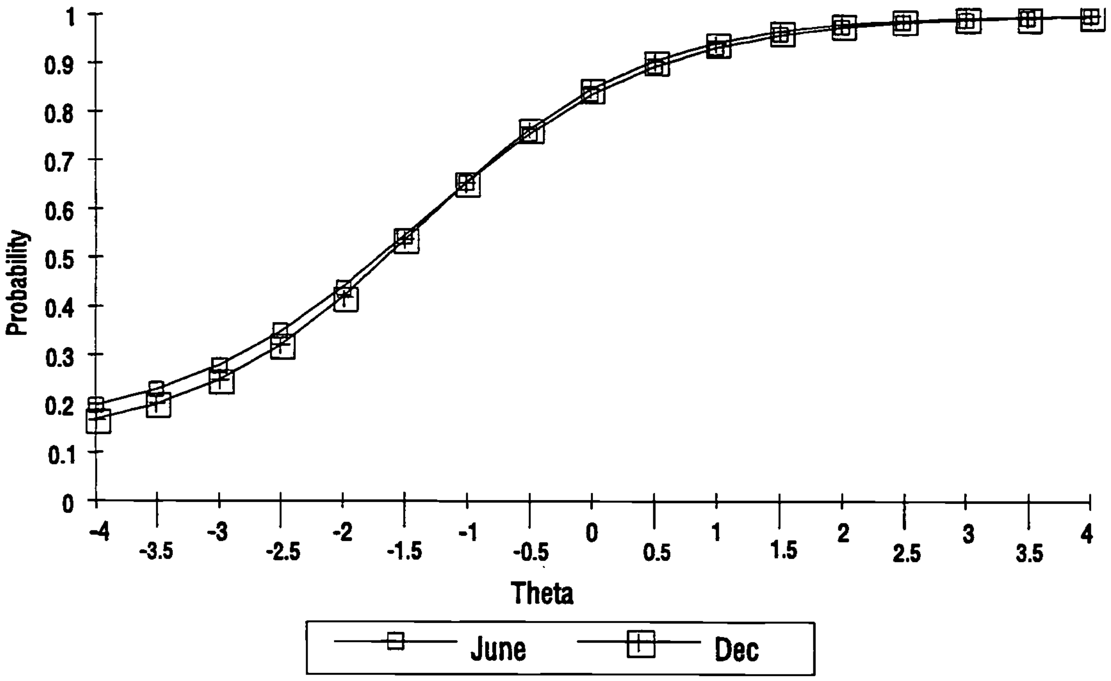


Item A7

Figure 10 (continued)

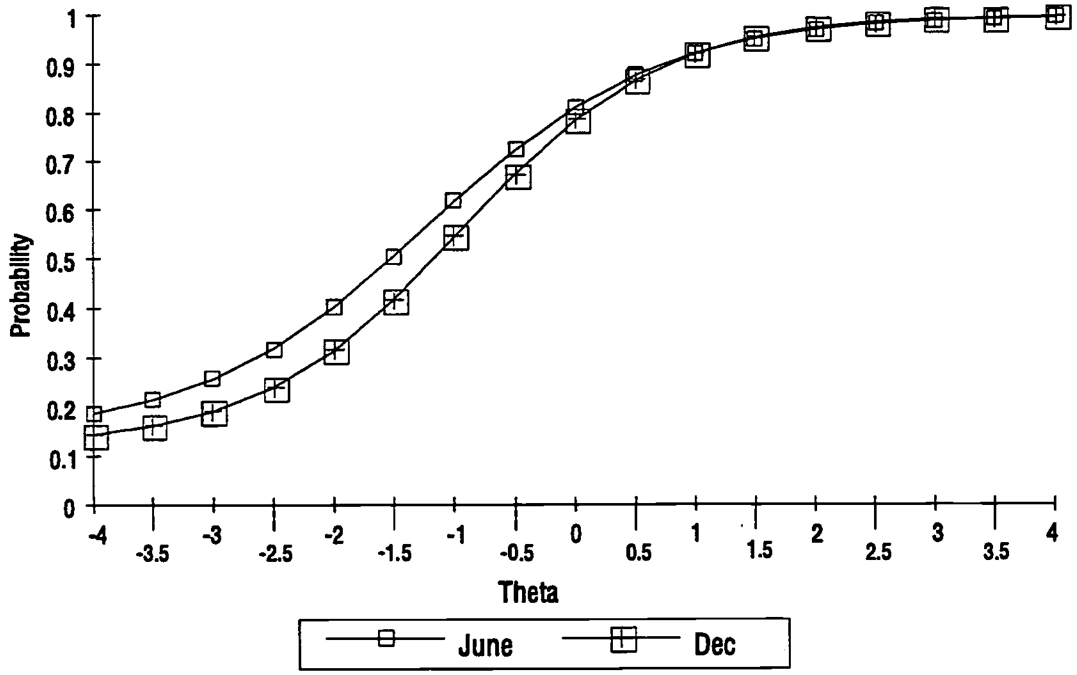


Item A9

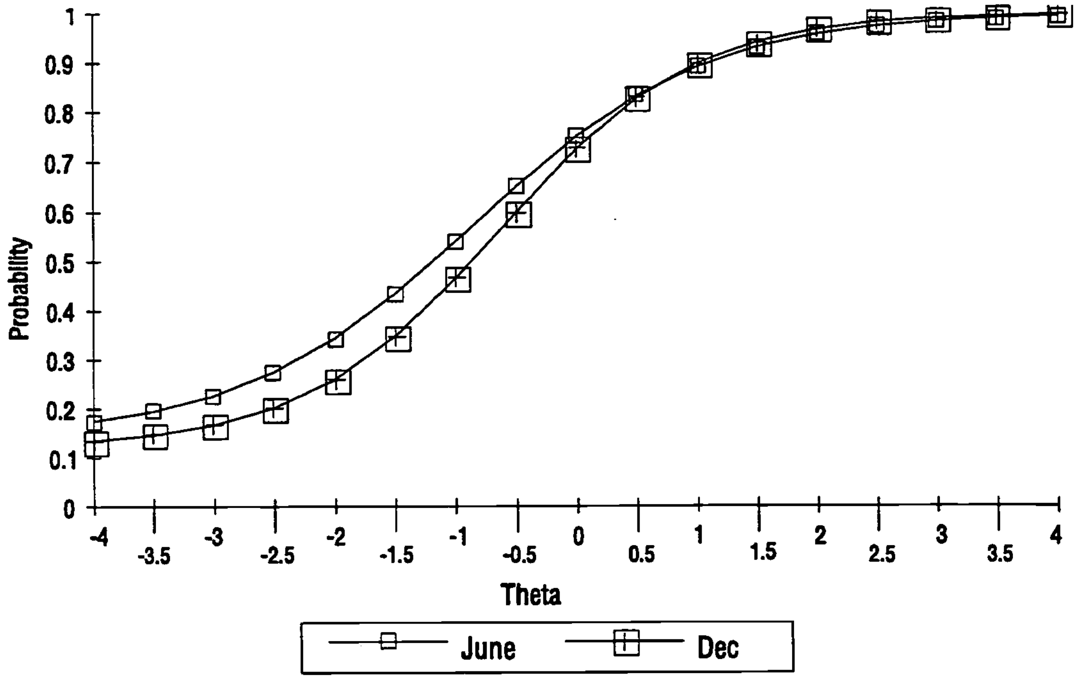


Item B2

Figure 10 (continued)

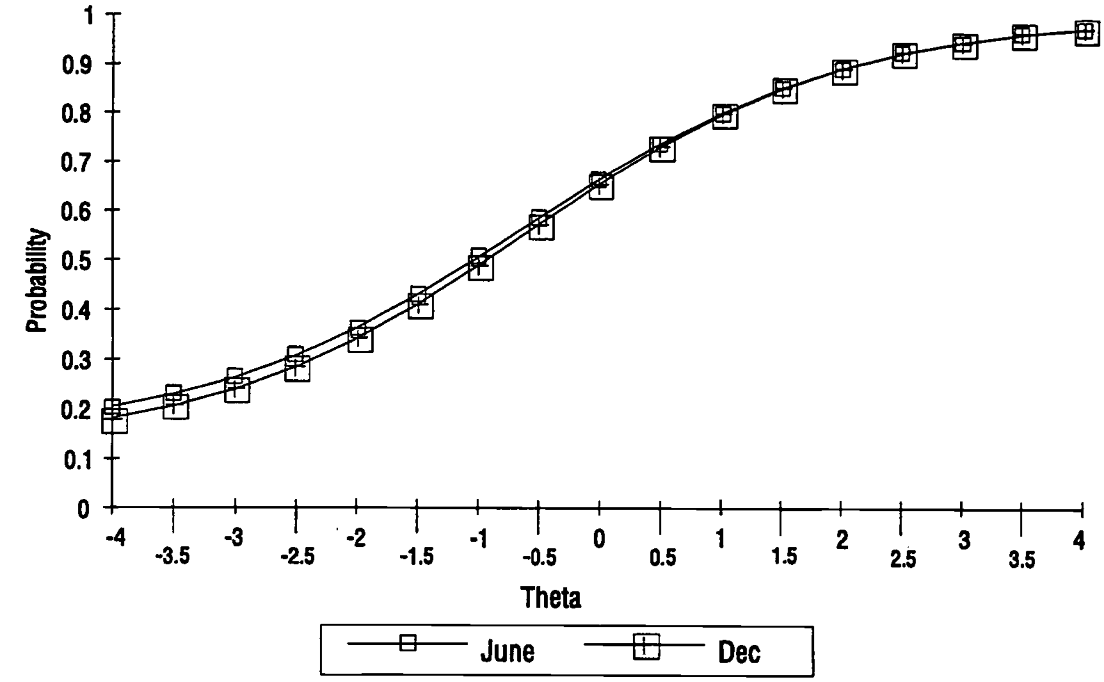


Item B10

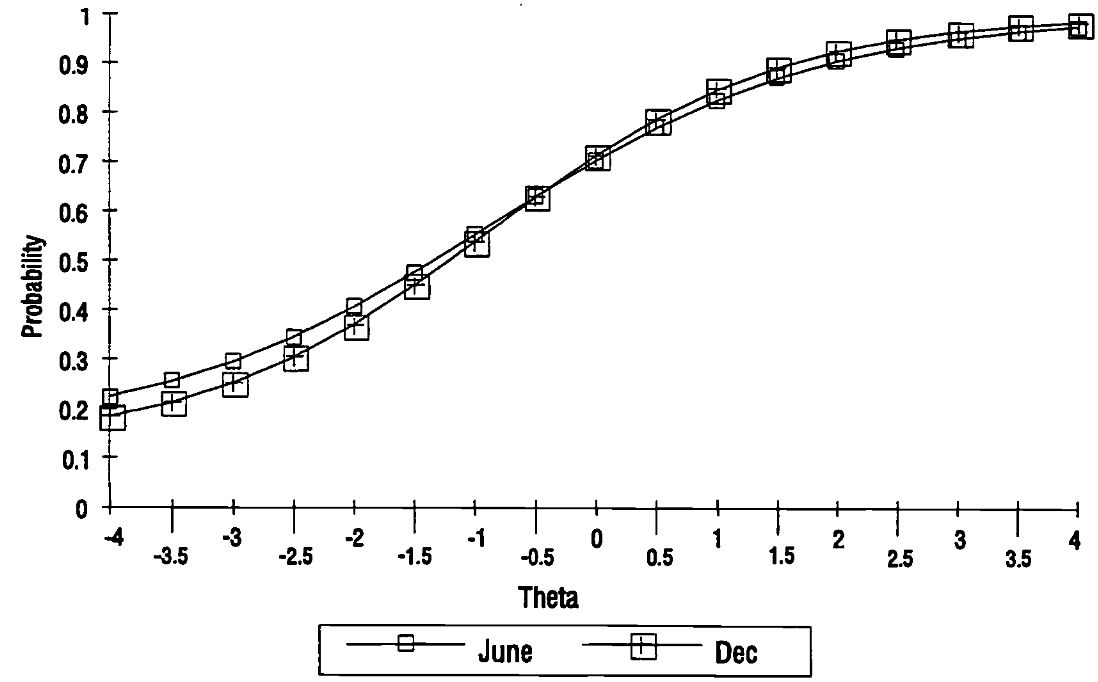


Item C1

Figure 10 (continued)

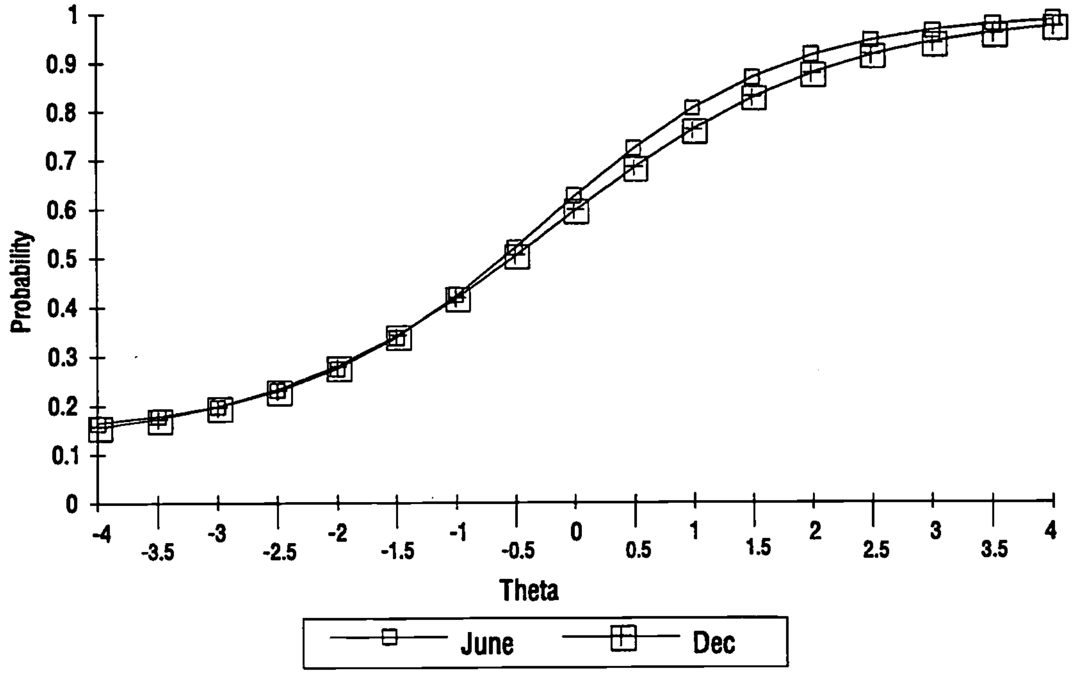


Item C2



Item D2

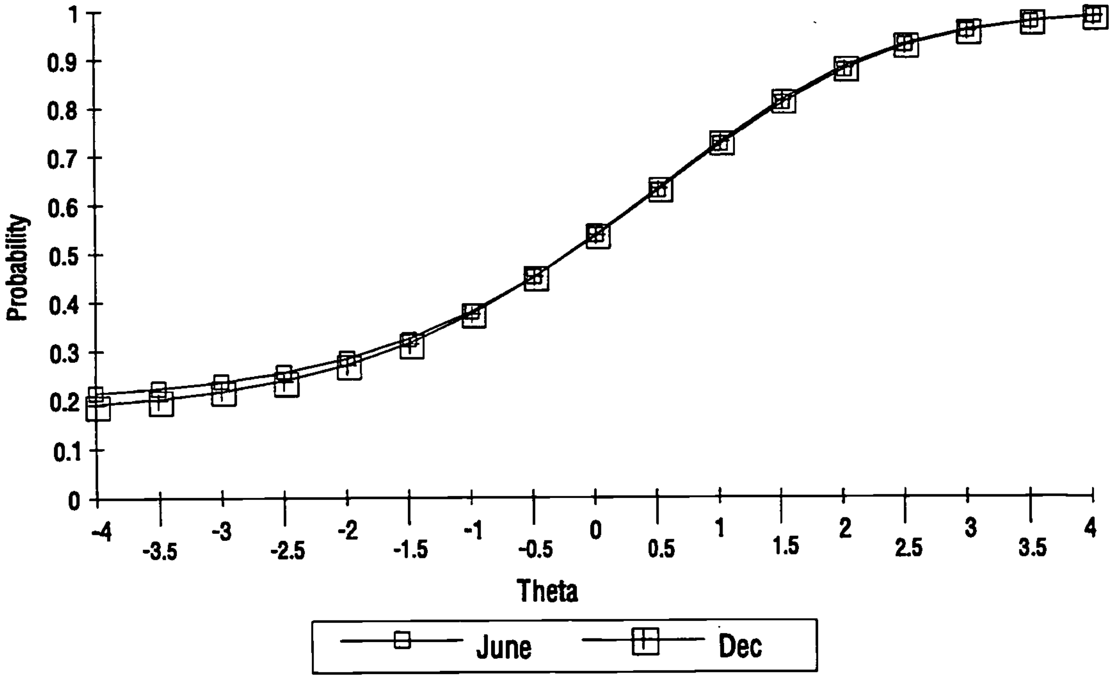
Figure 10 (continued)



Item D10

Figure 11

Test Characteristic Curve



Conclusions

In comparing the item characteristic curves for the different patterns of c- estimation, it seems that the most discrepant item characteristic curves are produced when the c-parameter is fixed for only one of the administrations. The items for which the c-parameter was estimated for both administrations produced more stable item parameter estimates than the case where the c-parameter was fixed for one administration. However, the items for which the c-parameter was fixed for both administrations seem to produce the most stable item characteristic curves. These differences at the item level do not present themselves at the test level.

For applications which rely on item level statistics, the results observed here are troublesome. For example, in computer adaptive testing algorithms which employ a maximum information item selection technique, the item which contributes the highest level of information at the current estimated ability level is chosen for administration to the test taker. Since the item parameters are utilized in calculating the information function, the discrepancies in item parameters observed here could result in less efficient and less precise measurement.

In test assembly, the instability in the item parameter estimates resulting from c-estimation problems may be still more problematic. Both the values of the IRT parameters and the correlations among the parameters have a significant impact on the psychometric properties of a test form. The precision of expectations based on pretest statistics may be significantly reduced if pretest statistics misrepresent the functional characteristics of the items on a test form.

The fact that the differences observed at the item level diminish at the test level is encouraging in some sense. This implies that for test level applications such as test equating, the issue of whether or not the c-parameter is set to the COMC value may not be a problem. An important issue to explore is the extent to which these differences impact on scaling results. In "mean- sigma" scaling, only the b-parameters are used, and this is thought to produce more stable results since the b-parameters are generally more stable than the a- and c-parameters. This was not the case, however, in this study. Our analyses indicated that the b-parameters were not as stable from one administration to the other as would normally be expected. In the characteristic curve transformation method of scaling (Stocking & Lord, 1983), which is the method used for the LSAT, the entire item characteristic curve is utilized in the scaling procedure. Intuitively, it would seem that this should produce more stable scaling results since differences in the c-parameter could be balanced out by the a- and b-parameters. However, our results imply that this may not be the case. When the c-parameter was set to COMC for only one administration, the item characteristic curves produced for an item were often radically different. It would seem that these differences could have a substantial effect upon the results obtained in the characteristic curve transformation method of scaling. Perhaps further research should be carried out in order to explore this issue.

A limitation of this study is that for the items having the c-parameter estimated for only one administration, we do not know how similar the a- and b-parameter estimates would have been if the c-parameter had been estimated for both administrations. The c-parameter was set to COMC for these items because there was not sufficient data to derive a stable parameter estimate. Therefore, it is possible that the a- and b-parameters for these items would have been even more disparate had the c-parameter been estimated rather than set to COMC. This possibility should be investigated. The important issue here, however is not whether or not COMC is an improvement over estimating the c-parameter, but whether or not the use of COMC is an acceptable procedure for dealing with the instability of the c-parameter for these items. We have learned from this study that the use of COMC is not a suitable solution, and a means of deriving a more reasonable estimate of the c-parameter for these items should be identified.

References

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wingersky, M.S., Patrick, R., & Lord, F.M. (1987). LOGIST User's Guide. Princeton, NJ: Educational Testing Service.