ED 395 976	TM 025 114
AUTHOR	Wightman, Lawrence E.; De Cham,lain, Andre F.
TITLE	A Comparison of the Propcrties of IRT Parameter Estimates Using Two Different Calibration Designs.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-94-19
PUB DATE	Apr 94
NOTE	33p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; Comparative Analysis; *Estimation (Mathematics); Item Banks; *Item Response Theory; Pretests Posttests; *Test Items
IDENTIFIERS	*Calibration; Graduate Management Admission Test; LOGIST Computer Program; *Three Parameter Model

ABSTRACT

Two different methods of obtaining three parameter logistic item response theory (IRT) pretest item parameter estimated for the Graduate Management Admissions Testing Program. The first method consisted of calibrating pretest and operational items simultaneously in a LOGIST run, that is a concurrent calibration design. The second approach entailed analyzing the pretest items separately from the operational items holding examinee ability scores constant from a previous operational items run, that is, using a two-stage calibration design. Results show that the means of the item difficulty (b-parameter) estimates were very similar, regardless of the method employed. However, the higher b-parameter values using the two-stage calibration run method (i.e., holding ability fixed excluding the studied items from the criterion) were slightly overestimated and the lower b-parameter values were slightly underestimated. The a-parameters were consistently underestimated using the two-stage estimation procedure. Finally, the slopes of the item-ability regressions using a concurrent calibration (including the studied items) are steeper for nearly all of the items. These preliminary results are consistent with those reported in past studies and suggest that non-operational (pretest) items should be calibrated concurrently with operational items for item banking purposes. (Contains 4 tables, 8 figures, and 16 references.) (Author/SLD)

RESEARCH REPORT

A COMPARISON OF THE PROPERTIES OF IRT PARAMETER ESTIMATES USING TWO DIFFERENT CALIBRATION DESIGNS

Lawrence E. Wightman
Andre F. De Champlain

Educational Testing Service
Princeton, New Jersey
April 1994

A Comparison of the Properties of IRT Parameter Estimates Using Two Different Calibration Designs

Lawrence E. Wightman

and

Andre F. De Champlain

October 1993

Abstract

The purpose of this study was to compare two different methods of obtaining 3PL IRT pretest item parameter estimates for the Graduate Management Admissions Testing Program. The first method consisted of calibrating pretest and operational items simultaneously in a LOGIST run, that is, a concurrent calibration design. The second approach entailed analyzing the pretest items separately from the operational items holding examinee ability scores constant from a previous operational items run, that is, using a two-stage calibration design. Results show that the means of the item difficulty (**b**-parameter) estimates were very similar, regardless of the method employed. However, the higher **b**-parameter values using the two-stage calibration run method (i.e. holding ability fixed excluding the studied items from the criterion) were slightly overestimated and the lower **b**-parameter values were slightly underestimated. The **a**-parameters were consistently underestimated using the two-stage estimation procedure. Finally, the slopes of the item-ability regressions using a concurrent calibration (including the studied items) are steeper for nearly all of the items. These preliminary results are consistent with those reported in past studies (Stocking & Eignor, 1986) and suggest that non-operational (pretest) items should be calibrated concurrently with operational items for item banking purposes.

A Comparison of the Properties of IRT Parameter Estimates Using Two Different Calibration Designs[1]

Lawrence E. Wightman
Andre De Champlain
Educational Testing Service

INTRODUCTION

The numerous studies dealing with Item Response Theory (IRT) that have dominated the measurement literature in the past decade attest to its importance in the development and analysis of tests and items. Indeed, IRT models are currently being utilized by large test publishers (Kingston & Stocking, 1986) as well as departments of education (Pandey & Carlson, 1983) for a variety of purposes such as norm- and criterion-referenced test development, test equating (Cook & Eignor, 1991) and the detection of differentially functioning items (Thissen, Steinberg, & Wainer, 1993). Warm (1978) summarizes the importance of IRT as follows:

> "Item Response Theory (IRT) is the most significant development in psychometrics in years. It is, perhaps, to psychometrics what Einstein's relativity is to physics. I do not doubt that during the next decade it will sweep the field of psychometrics." (p.11).

Several IRT models have been proposed (Hambleton & Swaminathan, 1985). One popular model is the three-parameter logistic function which can be written as follows.

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}, \tag{1}$$

where,

$P_i(\theta_j)$ = The probability of correctly answering item *i* given ability level θ;
a_i = The item discrimination parameter value for item *i*;
b_i = The item difficulty parameter value for item *i*;
c_i = The lower asymptote parameter value for item *i*;
D = A scaling factor, approximately equal to 1.7, to approximate a normal ogive model.

[1] The authors express their appreciation to Daniel Eignor and Linda Wightman for their comments on an earlier draft, to Ruth Mroczka for the production of the item/ability plots and, especially, to Martha Stocking for her encouragement and technical assistance along the way, as well as for her substantive comments on an earlier draft.

Indeed, IRT models possess several advantages over traditional (classical) test theory models, two of which are that they can provide sample-free item parameter estimates as well as test-free ability estimates. In addition, they can yield information pertaining to a wide range of examinee abilities. When the IRT model of interest fits the set of item responses that are to be analyzed, item parameter estimates are independent of the sample of examinees drawn from the population towards which the test is targeted and examinee ability estimates are independent of the particular set of items selected from the pool (Hambleton & Swaminathan, 1985).

These properties have important ramifications for several areas within educational measurement, including score equating and adaptive testing. The item invariance property enables the equating of a test before the actual administration assuming that item parameters are known *a priori*. Within the context of adaptive testing, the ability invariance property makes it possible to compare examinee abilities even if they did not respond to the same set of items.

The extent to which these properties hold for any given testing situation is indicated by the degree of fit between the IRT model of interest and the set of item responses. Though the advantages of IRT are numerous, the strong assumptions underlying the majority of these models generally must be met or the advantages of an IRT model do not hold.

Specific IRT models have particular assumptions of their own. For example, the Rasch model makes the very strong assumption that item discriminations are equal for all items in the test of interest and that the lower asymptote parameter is equal to zero (Hambleton & Swaminathan, 1985).

In addition, the assumptions of local independence and unidimensionality of the latent space are central to most IRT models. Briefly stated, local independence entails that for subsets of examinees of equal ability, the (conditional) distributions of item scores are statistically independent (Lazarsfeld, 1950; Lord & Novick, 1968). Local independence generally implies that:

$$P(u_1=1,u_2,\ldots,u_n=1\,|\,\theta)=\prod_{n}^{i=1} P(u_i=1\,|\,\theta) \qquad (2)$$

Therefore, at any given ability level, the probability of correctly answering all items on a test is equal to the product of the separate probabilities of correctly answering each item. Unidimensionality entails that the probability of a correct response on a given item is a function of a single latent trait (Lord & Novick, 1968).

Once the IRT model has been selected, several procedures are available for the estimation of the item and ability parameters. Joint maximum likelihood (JML) (Wingersky, Patrick, & Lord, 1991), marginal maximum likelihood (MML) (Bock & Lieberman, 1970) and Bayesian estimation procedures (Birnbaum, 1969; Owen, 1975) have been implemented in several IRT computer programs. LOGIST (Wingersky, Patrick, & Lord, 1991), the IRT computer program that is used operationally at ETS, uses a joint maximum likelihood estimation procedure.

Stocking and Eignor (1986), in a study dealing with the impact of different ability distributions on IRT preequating, examined the effect of certain calibration procedures on IRT parameter estimates with actual achievement test data. Specifically, in a supplement to the study presented in an appendix, item-ability regressions were plotted for two actual achievement test distributions: a first plot outlining observed proportions of correct responses for 'operational' and 'pretest' items for abilities estimated on operational items only and a second regression plot showing observed proportions of correct responses for 'operational' and 'pretest' items for abilities based solely on pretest items. The item-ability regression plot slopes were clearly steeper for the ability points that *included* the item in the ability estimate being considered.

Wingersky and Lord (1984) showed that sampling errors for item parameter estimates using JML are correlated. That is, lower *a*'s will produce underestimated low *b*'s and overestimated high *b*'s. Although Wingersky and Lord used simulated data, the same types of effects were noted with the *real* data sets in Stocking and Eignor (1986). The slope of the item-ability regression plots (either operational or pretest) were consistently steeper when the item of interest was included in the calibration. Calibrations made on the 'other' item set tended to produce depressed **a** parameters. Stocking and Eignor suggest that this depressed slope value in the item-ability regression plot could be attributable to a similar type of phenomenon noted with traditional item discrimination statistics, specifically item biserial correlation coefficients. Indeed, the biserial correlation coefficient value tends to be higher when the item of interest is included in the criterion. Also, Lord (1980) has shown that item biserial correlations are related to IRT item discrimination parameters in the following fashion:

$$\rho^{!}_{ix} = \frac{a_i}{\sqrt{1 + a_i^2}} \qquad (3)$$

where $\rho^{!}_{ix}$ = the item biserial correlation, and
a_i = the IRT item discrimination parameter

Based on this relationship, Stocking and Eignor (1986) suggest that a similar type of effect may be occurring with IRT item discrimination parameters. It is important to point out, however, that the relationship between the biserial correlation coefficient and the IRT discrimination parameter outlined in equation three holds only if there is no "pseudo-guessing", that is, a zero lower asymptote parameter value, and if the latent trait is normally distributed. These conditions were violated to a certain extent in the Stocking and Eignor study (1986) and hence their interpretations should be interpreted with some degree of caution. Clearly, more research needs to be undertaken before any definite conclusions are made regarding the estimation of IRT item and ability parameters under various calibration designs.

The purpose of the present study was to compare IRT item and ability parameter estimates obtained using two calibration procedures with item responses obtained from two forms of the Graduate Management Admissions Test: a first design that entailed calibrating non-operational (pretest) and operational items concurrently and a second design that involved analyzing the pretest items separately, holding constant examinee ability scores obtained from a previous operational items run. The item and ability parameters obtained from each procedure were compared to see if the effect noted by Stocking and Eignor (1986) was replicated in the present study. The comparison of the two procedures was replicated using a second form.

The study was motivated by a concern about the effect of pretest item type composition on item parameter estimates. More specifically, the proportion of items for each item type is not necessarily identical for the operational and pretest items. For example, the number of items of different types that appear in the pretests is determined by the needs of the GMAT item bank. The items for pretesting at a particular administration may be predominantly Sentence Correction items, for example. Since the proportions of verbal item types found in the pretest items may vary from the proportions of item types found in the operational sections, one question of interest is whether IRT item and ability parameter estimates should be obtained from the total set of items (i.e., operational, preequating and pretest items) or for the pretest items separately. However, the issue of parameter estimation bias resulting from different calibration procedures needs to be resolved before the proportionality question can be pursued.

METHODOLOGY

The testing program

The Graduate Management Admissions Test (GMAT) is comprised of seven sections: six scored sections, hereafter referred to as operational sections, as well as one additional section, hereafter referred to as the non-operational section. The non-operational section is not included in examinees' final scores. Although the non-operational section occupies a different position in the test from administration to administration, it is usually placed in sections one, two or three. Also, the content of the non-operational section differs from candidate to candidate. That is, the subforms are 'spiralled' so that an essentially random 1/*n* of the candidates receive each subform, where *n* is the number of subforms. The content of the non-operational section normally is in one of four categories:

1) a final form *verbal item type* section for the preequating[2] of a new form,

2) a final form *quantitative item type* section for the preequating of a new form,

3) a section of new items of a *verbal item type* for pretesting, or,

4) a section of new items of a *quantitative item type* for pretesting.

The sections of the test that are counted in candidates' scores are called operational sections, categories 1) and 2) are called preoperational sections and categories 3) and 4) are called pretest sections.

Typically, there are two new forms being preequated at a given administration. Thus, there are twelve subforms dedicated to preequating, that is, three verbal sections and three quantitative sections per new form, plus as many additional subforms as are needed to pretest available new items. There are usually 25 subforms in total. The layout of the operational, pre-operational and pretest sections in the forms examined and the number of items in each section is illustrated in Table 1.

Insert Table 1 about here

Analyses

The analyses conducted for this study were limited to the verbal items. As part of a series of analyses carried out on the data sets, three-parameter IRT models were fit to examinee responses using the joint maximum likelihood estimation procedure implemented in LOGIST VI (Wingersky, Patrick, & Lord, 1991). Specifically, verbal item data sets were analyzed using the *missing data* option in LOGIST given that only samples of candidates received any given subform. The

[2] Pre-equating for this testing program consists in administering all the sections of a new intact test form in the non-operational section of an existing form for the purpose of putting a new form on the established scale. Pre-equating sections are also called pre-operational (PreOP) sections.

operational items are administered to all candidates. There are typically 1,200 to 1,800 (or more) observations for each preequating and pretest item, with as many as 35,000 observations for the verbal (or quantitative) operational items used in a calibration run.

Two LOGIST runs were made for each of two administrations (January and October 1990) using the item responses from the verbal sections. The first run was based on the total set of items, that is, a concurrent calibration was used. The latter run involved only the pretest items holding examinee thetas constant from a previous operational items calibration. The results in each case were then put through a characteristic curve transformation run (Transformation of B's using Least squares Technique; Stocking and Lord, 1983) to put all the parameters and thetas on the original GMAT IRT scale.

RESULTS

Item difficulty parameter estimates

Table 2 presents summary statistics for the verbal item difficulty (***b***) parameters for the January and October 1990 administrations by calibration method for the operational (OP), preoperational (PO) and pretest (PT) items.

Insert Table 2 about here

Inspection of Table 2 shows that a difference in the mean ***b*** parameter values using the two calibration procedures is apparent for the January 1990 data only. Specifically, there was a difference of .12 in the mean item difficulty value obtained for pretest items when using the concurrent run and the two-stage calibration runs. Regarding the October 1990 data, the difference between the mean item difficulty was very small (.04) and hence would suggest that either method would yield very similar estimates.

In addition to examining the mean differences between the estimates obtained using the two calibration procedures, it is of some import to consider the differences in the parameters of the individual items. Figures 1 and 2 plot the October and January 1990 item difficulty estimates derived from the concurrent (single, combined LOGIST) run and the two-stage calibration (separate LOGIST runs/Fixed thetas) procedure. The forty-five degree line indicates the location of identical parameter estimates yielded by the two calibration procedures.

Insert Figures 1 and 2 about here

Item difficulty parameter estimates tend to be very similar irrespective of the calibration procedure. However, Figures 1 and 2 do show that lower item difficulty values tend to be estimated slightly below the equal estimation line when the two-stage calibration design was used, whereas higher item difficulties tend to be generally above the equal estimation line for the two-stage calibration procedure. In fact, this pattern occurred in almost identical fashion for the January and October data.

Item discrimination parameter estimates

Statistics for the verbal item discrimination parameters (*a*) for the January and October 1990 administrations by calibration method for the operational (OP), preoperational (PO) and pretest (PT) items are presented in Table 3.

Insert Table 3 about here

There were substantial differences in the mean discrimination parameter values depending on the calibration procedure that was used. More precisely, there was a difference of over .14 in the mean item discrimination value obtained for pretest items when using the concurrent calibration run versus the two-stage calibration runs for the January as well as October 1990 administrations. It would therefore appear as though the calibration procedure employed has a direct bearing on the item discrimination estima*es.

Figures 3 and 4 plot the October and January 1990 item discriminations for both calibration procedures. That is, the concurrent calibration (single, combined LOGIST run) and the two-stage calibration procedure (Separate LOGIST runs/Fixed thetas) parameter values are plotted against each other.

Insert Figures 3 and 4 about here

For both the October and January administration data, these plots clearly show that using the two-stage calibration design yielded substantially lower item discrimination estinates (compared to the concurrent calibration), regardless of item discrimination value.

Lower asymptote parameter estimates

The lower asymptote parameter estimate means for each calibration procedure are presented in Table 4.

Insert Table 4 about here

Table 4 shows that the lower asymptote parameter estimate means, for the January as well as October administrations, are very similar for the pretest items regardless of which calibration procedure is utilized. The largest discrepancy between any two means is .01 (for the October 1990 pretest items).

Figures 5 and 6 plot the October and January 1990 lower asymptote parameter estimates using the concurrent calibration procedure (single LOGIST run) and the two-stage calibration method (Separate LOGIST runs/Fixed thetas).

Insert Figures 5 and 6 about here

The results are consistent for both the October and the January data. These plots show that the ıse of one calibration procedure rather than another will have very little effect on the lower parameter estimates of the pretest items. Note that more items are estimated at COM C (common C) for the two-stage/fixed theta estimation procedure than for the combined procedure. (See the horizontal lines of observations in the area of .1 to .16 on the Y-axes, the separate runs axes, of Figures 1 and 2.)

Ability estimates

In order to examine item performance for given ability levels, item-ability regressions were plotted. The X-axis of a given plot corresponds to the ability (theta) estimates whereas the Y-axis corresponds to the proportion of examinees responding correctly at each score point. Illustrative examples of these plots are presented in Figures 7 and 8. An '*X*' *symbol* corresponds to the proportion of candidates correctly answering an item at a given ability level using the concurrent calibration (based on operational and pretest items) while a '*hexagon*' corresponds to the proportion of candidates correctly answering an item at that same ability level based on the two-stage calibration procedure. The solid line is the theoretical curve resulting from the IRT parameter estimation procedure based on the single LOGIST run; the dotted line is the theoretical curve resulting from the IRT parameter estimation procedure based on the two-stage LOGIST runs. If the theoretical fit is good, the curves will closely follow the sets of X's or hexagons.

Insert Figures 7 and 8 about here

As expected from the data presented in Figure 3 and 4, the item-ability regression curves tend to be steepest when using the concurrent calibration procedure. The higher item discrimination parameter estimates presented previously would indeed yield steeper curves due to larger slope values. The examples that are presented in Figures 7 (January 1990 administration) and 8 (October 1990 administration) vary in terms of item difficulty, discrimination, and the lower asymptote parameter value. However, regardless of the characteristics of the items, this effect (i.e., steeper slope for the empirical curve based on estimated abilities that *include* the item of interest) is noted for all items.

The maximum amount of information based on a three-parameter logistic IRT function is found at the following theta value:

$$\theta_{\max} = b_i + \frac{1}{Da_i}(\ln \frac{1+(1+8c_i)^{1/2}}{2}) \qquad \textbf{(4)}$$

where,
$\theta_{\max}$ = The theta value where the item provides the most information;
a_i = The item discrimination parameter value for item *i*;
b_i = The item difficulty parameter value for item *i*;
c_i = The lower asymptote parameter value for item *i*;
D = A scaling factor, approximately equal to 1.7, to approximate a normal ogive model.

This theta maximum value is primarily a function of item difficulty and discrimination. Hence, not surprisingly, the "underestimation effect" observed throughout the majority of the non-operational (pretest) items in the January and October 1990 GMAT forms tends to manifest itself more strongly with items that show the most discrepancy with regards to their difficulty and discrimination estimates (based on concurrent or two-stage calibration procedures): for example, item 407 in Figure 7 and item 246 in Figure 8. In addition, this effect is more pronounced in the range of the ability scale where the most information is provided by the item. In the case of the items shown in Figures 7 and 8, it occurs in the upper range.

Conversely, for items with lower discrimination parameter values, that is, items that provide information along the entire ability scale rather than concentrated information in a narrow band, the underestimation effect tends to be less severe but evident throughout the whole range of thetas (c.f. items 323 and 417 from Figure 7 and items 244 and 357 from Figure 8).

DISCUSSION

The purpose of this research was to compare IRT item and ability parameter estimates using two distinct calibration procedures. Items analyzed were verbal operational, preoperational and pretest item responses taken from the January and October 1990 GMAT administrations. The first calibration procedure entailed estimating item and ability parameters using a concurrent calibration, that is, analyzing all items simultaneously (non-operational and operational). The second procedure involved calibrating items and abilities utilizing a two-stage calibration, that is, analyzing pretest items separately from the operational items holding examinee ability scores constant from a previous operational items run. Previous findings in the literature have indicated that the item as well as ability parameter estimates obtained using these two methods differ appreciably (Stocking & Eignor, 1986).

The results obtained in this study parallel very closely those presented by Stocking and Eignor (1986). In particular, pretest item discrimination parameter estimates were substantially lower when using a two stage calibration procedure (that is pretest items were not included in the ability estimate) than when using a concurrent procedure (pretest items did contribute to the ability estimates). The item difficulty and lower asymptote parameter estimates, however, were less affected by the calibration procedure utilized. When item-ability regression plots were examined, the underestimation effect noted by Stocking and Eignor (1986) was also omnipresent for the items examined in the present investigation.

It is somewhat more difficult to interpret the findings obtained in this study than those presented by Stocking and Eignor. Stocking and Eignor analyzed simulated date partially, while this study used actual test data where *true* item parameter as well ability parameter values are unknown.

However, the results obtained in this study so strongly confirm those that would be predicted by Stocking and Eignor's results that they strongly suggest the same underlying mechanism is accounting for them. That is, when an item is not included in the criterion score, the ***a*** parameter is estimated lower and since errors between the ***a*** and ***b*** parameter estimates are correlated, the ***b*** parameter estimates are also affected.

The findings obtained in this study support the following general recommendation: Item and ability parameter estimates are directly affected by the calibration procedure used. Given that test assemblers make use of item information provided from calibrations developed at pretest time, it is important to realize that the target information function curves devised from lower discrimination parameter values developed with a two-stage procedure will more than likely *not* reflect the amount of information provided by these items at the administration of a final form assembled using these pretest parameters. Accordingly a concurrent calibration IRT analysis design of pretest items should be specified (i.e. the items should be included in the criterion (theta estimates).

Further research is needed to investigate effects on parameter estimation when the proportions of item types for the items being pretested are different from the proportions of those item types in the operational sections.

REFERENCES

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. Journal of Mathematical Psychology, 6, 258-276.

Bock, R.D. & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. Psychometrika, 35, 179-197.

Cook, L.L, & Eignor, D.R. (1991). IRT equating methods. Educational Measurement: Issues and Practice, 10, 37-45.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications, Boston, MA: Kluwer-Nyjhoff.

Kingston, N.M, & Stocking (1986, August). Psychometric issues in IRT-based test construction. Paper presented at the meeting of the American Psychological Association, Washington, DC.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer et al. (Eds.), Measurement and prediction. Princeton, NJ: Princeton University Press.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Pandey, T.N., & Carlson, D. (1983). Application of item response models to reporting assessment data. In R.K. Hambleton (Ed.), Applications of item response theory, Vancouver, BC: Educational Research Institute of British Columbia.

Stocking, M.L., & Eignor, D.R. (1986). The impact of different ability distributions on IRT preequating (Research Report No. 86-49). Princeton, NJ: Educational Testing Service.

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Warm, T.A. (1978). A primer of item response theory (Tech. Rep. No. OG-941278), Oklahoma City, OK: U.S. Coast Guard Institute.

Wingersky, M.S., & Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Wingersky, M.S., Patrick, R., & Lord, F.M. (1991). LOGIST VI users guide. Princeton, NJ: Educational Testing Service.

Table 1

Typical layout of GMAT Verbal sections

SAMPLE	Operational			PreOp 1			PreOp 2			Pretesting					
	1-23	24-50	51-70	71-93	94-120	121-140	141-163	164-190	191-210	211-233	234-256	257-281	282-306	307-331	332-351
1	X[3]	X	X	X											
2	X	X	X		X										
3	X	X	X			X									
.	X	X	X				.	.	.	.	.	.	.		
.	X	X	X				.	.	.	.	.	.	.		
.	X	X	X				.	.	.	.	.	.	.		
24	X	X	X											X	
25	X	X	X												X

[3] X indicates that data are observed; all other cells are 'missing' (i.e. coded as '3' for LOGIST runs).

Table 2

Verbal item difficulty parameter estimates by IRT calibration procedure

IRT Calibration procedure		January '90				October '90			
		OP	PO 1	PO 2	PT's	OP	PO 3	PO 4	PT's
Concurrent calibration	Run n	475				435			
	n	70	70	70	265	70	70	70	225
	Mean	-.2404	-.1680	-.0074	-.2350	-.0893	-.3175	-.1229	-.1451
	SD	1.2214	1.2184	1.2115	2.1196	1.3658	1.2427	1.2210	2.0003
Two-stage calibration (fixed Thetas)	Run n	335				295			
	n	70			265	70			225
	Mean	-.2192			-.3552	-.0881			.1839
	SD	1.2177			1.7046				2.6400

Table 3

Verbal item discrimination parameter estimates by IRT calibration procedure

IRT calibration procedure		January '90				October '90			
		OP	PO 1	PO 1	PT's	OP	PO 3	PO 4	PT's
Concurrent calibration	Run n	475				435			
	n	70	70	70	265	70	70	70	225
	Mean	.6759	.7606	.7190	.7257	.7145	.6932	.7499	.6851
	SD	.2003	.2167	.2513	.2580	.2499	.2170	.2587	.2632
Two-stage calibration (fixed Thetas)	Run n	335				295			
	n	70			265	70			225
	Mean	.6786			.5813	.7198			.5426
	SD	.1977			.2079	.2503			.2018

Table 4

Verbal lower asymptote parameter estimates by IRT calibration procedure

IRT calibration procedure		January '90				October '90			
		OP	PO 1	PO 1	PT's	OP	PO 3	PO 4	PT's
Concurrent calibration	Run n	475				435			
	n	70	70	70	265	70	70	70	225
	Mean	.1657	.1995	.2196	.1856	.1700	.1865	.1825	.1766
	SD	.0751	.1416	.1554	.1050	.1016	.1212	.1069	.1132
Two-stage calibration (fixed Thetas)	Run n	335				295			
	n	70			265	70			225
	Mean	.1725			.1811	.1719			.1665
	SD	.0709			.0967	.0991			.1146

Figure 1

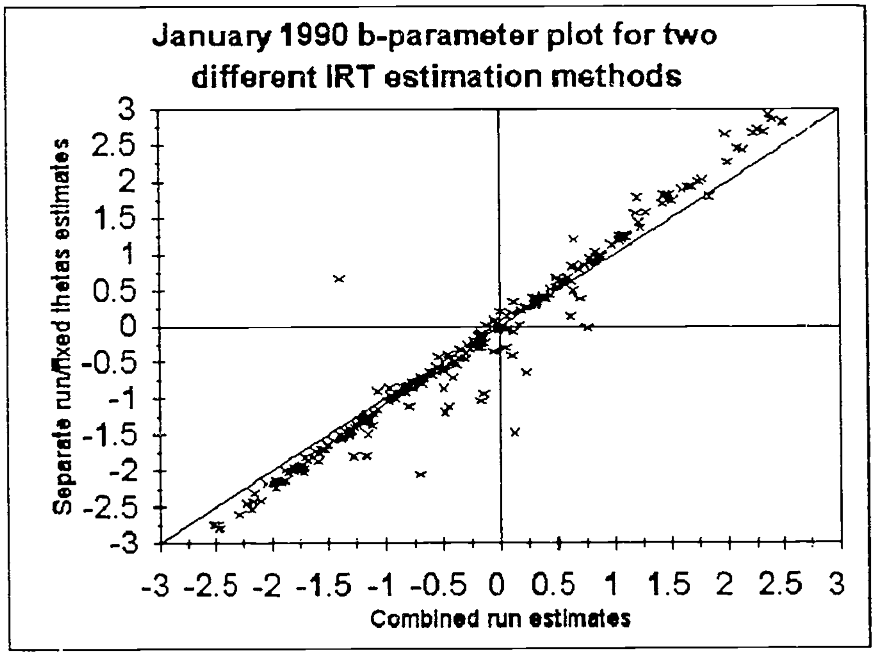


Figure 2

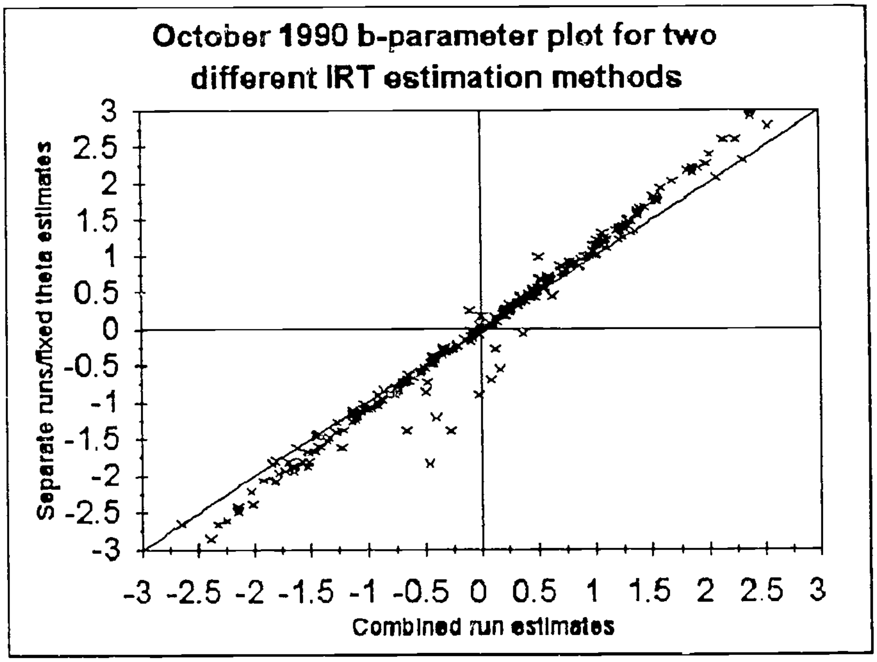


Figure 3

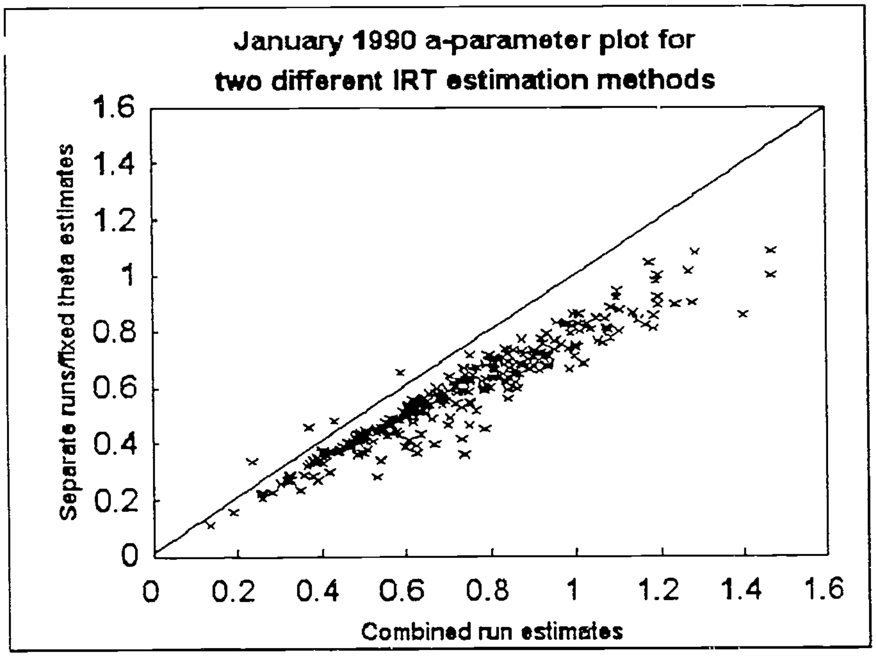


Figure 4

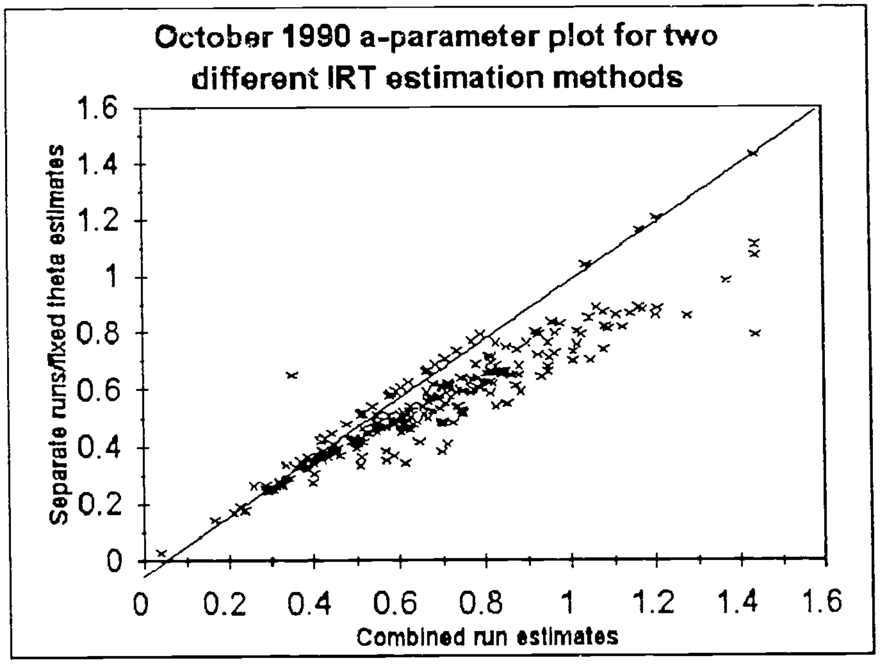


Figure 5

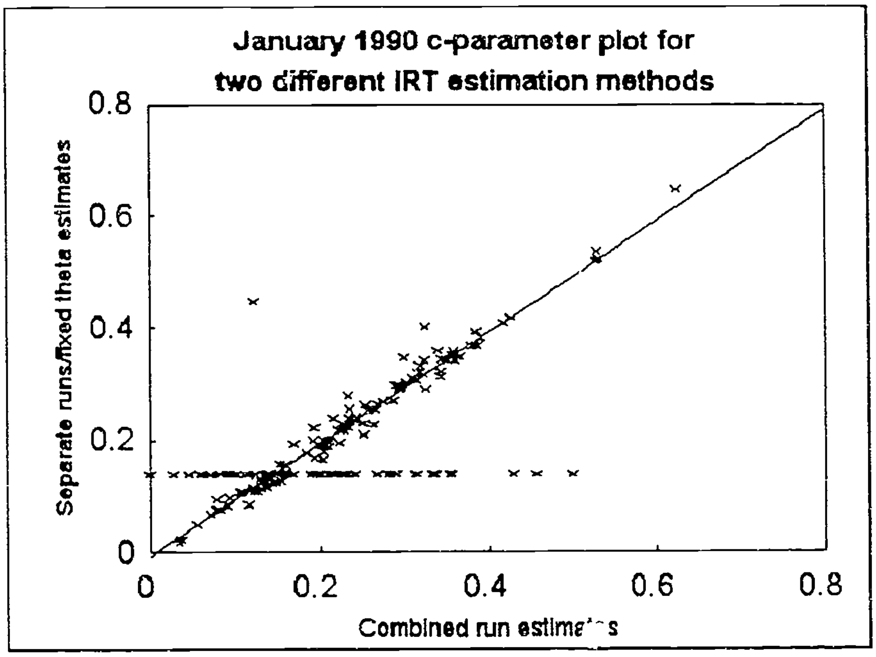


Figure 6

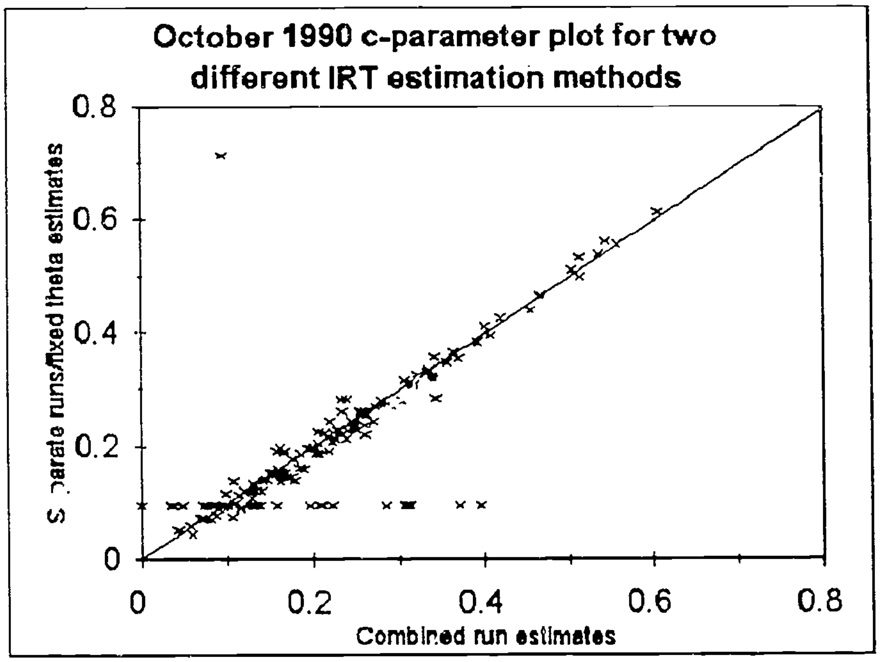


Figure 7

Item-ability regression plots: January 1990 GMAT form

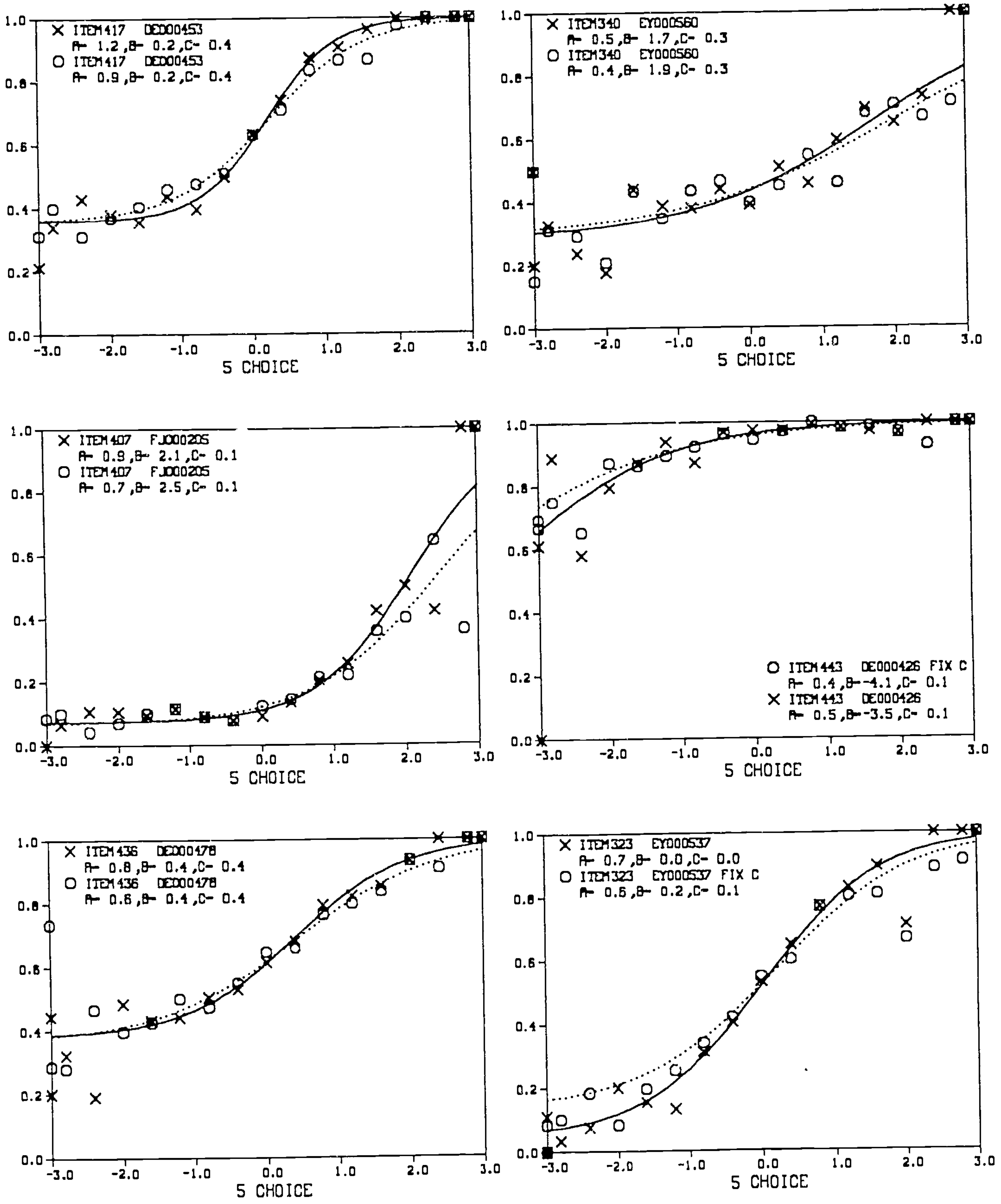
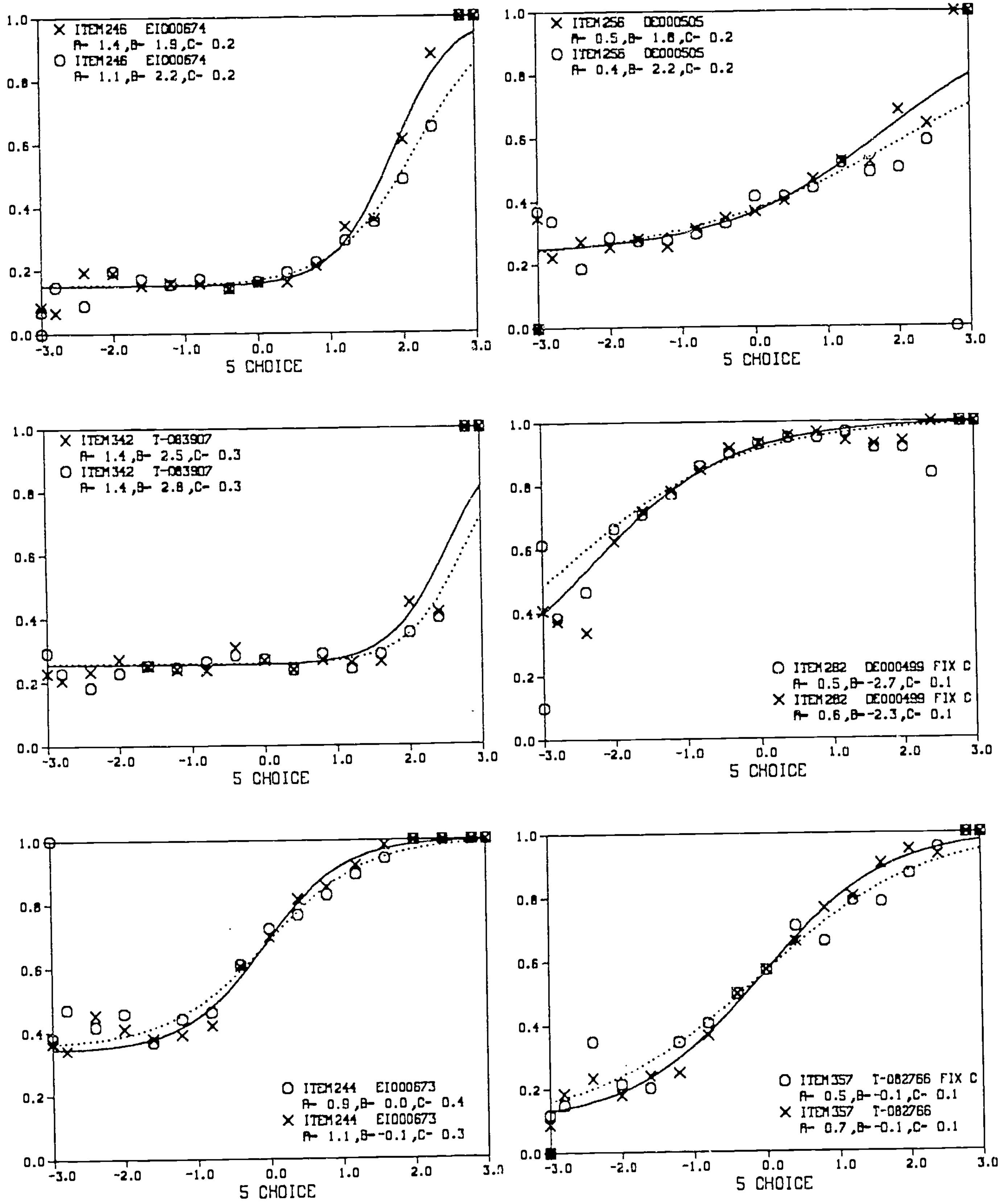


Figure 8

Item-ability regression plots: October 1990 GMAT form


× ITEM246 EI000674
A- 1.4 ,B- 1.9 ,C- 0.2
○ ITEM246 EI000674
A- 1.1 ,B- 2.2 ,C- 0.2
× ITEM256 DE000505
A- 0.5 ,B- 1.8 ,C- 0.2
○ ITEM256 DE000505
A- 0.4 ,B- 2.2 ,C- 0.2
× ITEM342 T-083907
A- 1.4 ,B- 2.5 ,C- 0.3
○ ITEM342 T-083907
A- 1.4 ,B- 2.8 ,C- 0.3
○ ITEM282 DE000499 FIX C
A- 0.5 ,B- -2.7 ,C- 0.1
× ITEM282 DE000499 FIX C
A- 0.6 ,B- -2.3 ,C- 0.1
○ ITEM244 EI000673
A- 0.9 ,B- 0.0 ,C- 0.4
× ITEM244 EI000673
A- 1.1 ,B- -0.1 ,C- 0.3
○ ITEM357 T-082766 FIX C
A- 0.5 ,B- -0.1 ,C- 0.1
× ITEM357 T-082766
A- 0.7 ,B- -0.1 ,C- 0.1
5 CHOICE