DOCUMENT RESUME

ED 469 184 TM 034 489

AUTHOR	Wang, Xiang-Bo; Harris, Vincent; Roussos, Louis
TITLE	Effects of Multidimensionality on IRT Item Characteristics and True Score Estimates: Implications for Computerized Test Assembly. Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-97-06
PUB DATE	2002-07-00
NOTE	23p.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; *College Entrance Examinations; Computer Assisted Testing; Estimation (Mathematics); *Item Response Theory; Law Schools; Test Construction; *Test Items; *True Scores
IDENTIFIERS	Item Characteristic Function; *Law School Admission Test; *Multidimensionality (Tests)

ABSTRACT

Multidimensionality is known to affect the accuracy of item parameter and ability estimations, which subsequently influences the computation of item characteristic curves (ICCs) and true scores. By judiciously combining sections of a Law School Admission Test (LSAT), 11 sections of varying degrees of uni- and multidimensional structures are used to assess the impact of multidimensionality on ICCs and true scores. This combination makes an artificial test for analysis purposes. It was found that multidimensionality decreased item slopes and estimated true score curves. The effect is shown to be bigger in the high ability range than in the low ability range. (Contains 23 figures and 21 references.) (Author/SLD)

TM

■ Effects of Multidimensionality on IRT Item Characteristics and True Score Estimates: Implications for Computerized Test Assembly

Xiang-Bo Wang
Law School Admission Council

Vincent Harris
Law School Admission Council

Louis Roussos
Law School Admission Council

■ Law School Admission Council
Computerized Testing Report 97-06
July 2002

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 200 law schools in the United States and Canada.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

During the past 10 years, item response theory (IRT) models have been increasingly used for a wide variety of purposes, such as score equating, identification of differential item functioning, and, most recently, computer-adaptive testing (CAT). One of the principal assumptions of IRT is unidimensionality—a single latent trait or ability underlies a set of item responses. This assumption is especially relevant for CAT, because the majority of current CAT alogrithms rely heavily on item characteristic curves (ICC) and item information of unidimensional IRT models for test assembly and ability estimation.

However, the assumption of unidimensionality is more often violated than observed in practice because of the multidimensional nature of common test purposes and items. Multidimensionality is known to affect item parameter estimates which, in turn, influence item characteristic curves and true scores. The purpose of this study is to further assess the impact of multidimensionality on ICC and true scores.

Data from one Law School Admission Test (LSAT) were analyzed in this study, due to its well-documented two-dimensional, though highly correlated, structure. The LSAT consists of four sections: Analytical Reasoning (AR), Logical Reasoning A (LR:A), Logical Reasoning B (LR:B) and Reading Comprehension (RC). The first LSAT dimension, composed soley of AR items, can be labeled as "deductive reasoning," whereas the second dimension, loading on LR:A, LR:B, and RC items, can be interpreted as "reading/informal reasoning." By judiciously combining certain sections of the LSAT data, 11 artificial tests are created that have 11 degrees of dimensionality structures. By analyzing these artificially created data sets, the effects of multidimensionality on the estimation of item response functions and true scores are investigated. Note that such effects are controlled with the current paper-and-pencil LSAT by maintaining a constant dimensional structure from form to form.

This study has found that multidimensionality decreases item slopes and the estimated true score curve in the high ability range. The estimated true score reduction is about 1 point at the sectional level of 24 items, and about 4 points at the total test level of 102 items. Multidimensionality seems to have a greater effect in the higher ability range than in the low ability range. It can be inferred that more diverse multidimensional structures may exhibit much bigger effects on ICCs and true scores. These effects may be more significant within a CAT framework, because true scores are estimated from fewer items.

Abstract

Multidimensionality is known to affect the accurary of item parameter and ability estimations, which subsequently influences the computation of item characteristic curves and true scores. By judiciously combining sections of a Law School Admssion Test (LSAT), 11 data sets of varying degrees of uni- and multidimensional structures are used to assess the impact of multidimensionality on ICCs and true scores. It has been found that multidimensionality decreases item slopes and estimated true score curves. The effect is shown to be bigger in the high ability range than in the low ability range.

Introduction

Item Response Theory (IRT) models are used by practitioners and researchers alike for a multitude of purposes, such as score equating (Lord, 1977, 1980; Petersen, Kolen, & Hoover, 1989), the identification of differentially functioning items (Thissen, Steinberg, & Wainer, 1993), and more recently, in the development of computer-adaptive tests (CAT), (Hambleton, Zaal, & Pieters, 1991; Wainer, 1990). One of the important assumptions underlying most IRT models is unidimensionality of latent ability space (Lord & Novick, 1968). That is, common IRT models assume that a single latent trait, or ability underlies a set of item responses. For example, the probability that examinee j correctly answers test item i, denoted by Y_{ij},= 1 based on the three-parameter logistic IRT model is given by

$$P\left(Y_{ij}=1|a_i,b_i,c_i,\theta_j\right)=c_i+\left(1-c_i\right)\frac{e^{Da_i\left(\theta_j-b_i\right)}}{1+e^{Da_i\left(\theta_j-b_i\right)}}, \quad (1)$$

where

θj = the latent trait value, interpreted as examinee j's ability level on the hypothesized construct;
a_i = an item discrimination parameter estimate for item i;
b_i = an item difficulty parameter estimate for item i; and
c_i = a lower asymptote parameter estimate for item i.

In practice, the assumption of unidimensionality of ability is rarely met with actual test data where the response to an item is often dependent upon several factors. In addition, researchers who have investigated the robustness of IRT item and ability estimates with simulated multidimensional data sets have shown that these (true) multidimensional parameters are often poorly recovered by the unidimensional models, especially when several equally important abilities are required to correctly answer a set of test items (Doody, 1985; Drasgow & Parsons, 1983; Hsu & Yu, 1989; Reckase, Carlson, Ackerman, & Spray, 1986).

The assumption of unidimensionality is especially relevant for computer-adaptive testing (CAT) where a large number of forms are administered ("tailored") to different groups of test takers in order to estimate their ability levels with the highest degree of precision (and usually with fewer items than are required with a linear paper-and-pencil test). The majority of current CAT algorithms rely heavily on IRT models for the assembly of forms. Hence, the impact of multidimensionality on item characteristics is potentially even more severe within a CAT framework given the reliance on IRT models for the selection of items to be included in each tailored form. The CAT forms assembled from item characteristics should be comparable across test takers in order to enable valid score-based inferences for all test takers, irrespective of the set of items selected. It would therefore seem imperative to assess the impact of multidimensionality on ICCs and true scores in order to determine how the CAT assembly process might be affected by calibrations conducted under unidimensional versus multidimensional conditions. Also, it would be important to assess the degree to which true score estimates vary when based on item response calibrations that differ with respect to their underlying dimensional structure. Preliminary research suggests that true score estimates are fairly robust to violations of unidimensionality (Camilli, Wang, & Fesq, 1995; Dorans & Kingston, 1985). Nonetheless, more research needs to be undertaken in a larger number of conditions before making any definite conclusions about the effect of multidimensionality on the estimation of true scores using an IRT model.

The purpose of this study is to further assess the effects of multidimensionality on ICCs and true scores. The ICCs and true score estimates of the same sets of items, but from different data dimensional compositions, will be compared. The results are based on artificially constructed compositions of actual item responses.

Data Collection

Data from one administration of the Law School Admission Test (LSAT) were analyzed in this study. The LSAT has a stable dimensional structure that has been well studied and documented by past research. The LSAT consists of four sections: Analytical Reasoning (AR), Logical Reasoning A (LR:A), Logical Reasoning B (LR:B) and Reading Comprehension (RC). The LR:A and LR:B sections are virtually identical in terms of item content and specifications. A typical LSAT form is composed of 101 to 102 items, with each section containing between 24 and 27 items. The specific number of items and test takers used in this study are summarized in Table 1.

TABLE 1
Summary of item numbers

Number of Items Per Section				Total Test	Approximate Number of Examinees
AR	LR:A	LR:B	RC		
24	25	26	27	102	46,000

Past research has shown that a two-dimensional structure underlies the item responses to the LSAT (Ackerman, 1994; Camilli, Wang & Fesq, 1995; De Champlain, 1995, 1996; Roussos & Stout, 1994). The first dimension, composed solely of AR items, can be labeled as "deductive reasoning" whereas the second factor, loading on LR:A, LR:B, and RC items, can be interpreted as "reading/informal reasoning." In spite of the fact that the AR section is dimensionally distinct, it still correlates substantially with the LR:A, LR:B, and RC sections. Using NOHARM, a nonlinear factor-analysis program (Fraser & McDonald, 1988), Roussos (1996) obtained a correlation of 0.62 between AR and LR, and a correlation of 0.72 between AR and RC for the December 1991 LSAT. The correlation between the LR and RC sections was 0.89. In terms of number-right scores, AR had correlations of 0.59 and 0.49 with LR and RC, respectively (Douglas, Kim, Roussos, Stout, & Zhang, 1999).

By judiciously analyzing certain combinations of sections from the LSAT data, we can create artificial tests having a variety of different dimensionality structures. Then by analyzing these artificially created data sets, we investigate the effects of these different dimensionality structures on the estimation of item response functions and true scores.

Analysis Methods

The analyses in this paper are divided into two components—a set of item calibrations and a series of comparisons of ICCs and true scores that resulted from the calibrations. The following details the two components.

Analysis Component 1: Item Calibrations

In order to reveal the effects of multidimensionality on item parameter estimates, this study undertook 11 calibrations as outlined in Table 2 on the same 102 items and approximately 46,000 test takers using BILOG (Mislevy & Bock, 1990). In order to maximize item response information in the data, every test taker is used for each of the 11 calibrations, instead of using the BILOG default option of randomly sampling 1,000 test takers. As shown in Table 2, the 11 calibrations are divided into four groups, and the data set of each of the 11 calibrations reflects a certain dimensional structure as will be described below.

TABLE 2
Item calibration schemes

Group	Calibration No.	ItemType Section Combination	Dimensionality
1	Calibration 1:	AR + LR:A + LR:B + RC	Essentially two-dimensional
2	Calibration 2:	AR + LR:A	Two-dimensional
	Calibration 3:	AR + LR:B	Two-dimensional
	Calibration 4:	AR + RC	Two-dimensional
3	Calibration 5:	LR:A + LR:B	Unidimensional
	Calibration 6:	LR:A + RC	Essentially unidimensional
	Calibration 7:	LR:B + RC	Essentially unidimensional
4	Calibration 8:	AR only	Unidimensional
	Calibration 9:	LR:A only	Unidimensional
	Calibration 10:	LR:B only	Unidimensional
	Calibration 11:	RC only	Unidimensional

Group 1 has only one calibration that calibrates all 102 items concurrently. It is obvious that Calibration 1 is dominated by items representing the reading/informal reasoning dimension, since there are only 24 AR items versus 77 LR:A, LR:B, and RC items combined. Thus, the unidimensional IRT estimates from BILOG can be viewed as heavily weighted by the reading/informal reasoning dimension. All the results related to this calibration are called "essentially two-dimensional."

Group 2 consists of three calibrations, Calibrations 2–4, with AR items combined with LR:A, LR:B, and RC items, respectively. Since the data set for each of the three calibrations are still two-dimensional, the item parameter estimates obtained in this group are constrained to be unidimensional despite their original two-dimensional structure. However, the main difference between Group 1 and Group 2 calibrations is that Group 2 calibrations have approximately equal numbers of items representing each of the two dimensions of the original data. All the results related to this group of calibrations are referred to as "two-dimensional."

Group 3 also has three calibrations, Calibrations 5–7, with two distinct features. First, Calibration 5, LR:A and LR:B items, is strictly unidimensional, since it consists of all logical reasoning items. Calibrations 6-7, LR:A and LR:B plus RC items, contain essentially unidimensional items, reflecting the reading/informal reasoning dimension, because RC is not exactly the same as the LR items but quite similar in construct. Thus, all the results related to Calibration 5 are considered unidimensional, while all the results associated with Calibrations 6–7 are deemed "essentially unidimensional." In addition, all three calibrations contain approximately the same number of items.

Group 4 is composed of four independent calibrations, one for each of the four sections. Two characteristics are clear. First, each of the four calibrations is unidimensional, since all items are from one section only. The second characteristic is that the number of items in each of the four calibrations is fairly small, from 24 to 27 items. These represent only half the items in Groups 2 and 3 and nearly one quarter of the items in Group 1 calibrations.

In order to compare the item parameter estimates obtained in the 11 calibration runs, the items must be placed on the same scale. This was accomplished by constraining the ability distribution of each calibration to a standard normal distribution since the same test takers are used in each of the 11 calibrations. This will force differences due to multidimensionality to be exhibited in the ICCs only.

Analysis Component 2: Comparisons of ICCs and True Scores

Once the items were calibrated and scaled, the ICCs and true scores based on each calibration were compared in order to assess the effect of multidimensionality at the item, section, and total test levels.

According to the 11 calibrations listed in Table 2, any single item may have up to five ICCs. For example, an AR item has five ICCs, four from Calibrations 1, 2, 3, and 4, and one from Calibration 8. Based on the two-dimensional structure known to be present in the complete data set, it is hypothesized that:

- the ICCs across Calibrations 1– 4 will be similar because of the comparable two-dimensional structure underlying the item responses;
- the ICCs across Calibrations 5–11 will be similar because of the comparable unidimensional structure underlying the item responses;
- ICCs from Calibrations 1–4 will be flatter than their counterparts from calibrations 5–11 since multidimensionality is known to reduce item discrimination. Such differences can potentially be attributed to differences in dimensional structure.

True scores are estimated using the following formula,

$$\tau j = \sum_{i=1}^{n} P(Y_{ij} = 1 | a_i, b_i, c_i, \theta_j) \qquad (2)$$

where

τ_j = estimated true score for test taker j; i = the number of items the true score is based on; and the probability P is defined in Equation 1.

True scores are computed at both the sectional and total test levels. It is hypothesized that the true scores estimated using the two-dimensional data sets will differ noticeably from those derived with the unidimensional datasets.

As for the total test level, there are four true score lines representing varying degrees of multidimensionality. It is hypothesized that the more unidimensional true score lines will be steeper than the more multidimensional true score lines.

Results

The results of this study will be summarized in three sections: (1) ICC comparisons; (2) sectional true score comparisons; and (3) total test true score comparisons.

For most AR items, four of the five ICCs were found to be very similar and one ICC separate from the rest to varying degrees. Which calibration does the deviant AR ICC usually represent? A closer look at the amplified plate for AR Item 5 in Figure 1 shows that it is the ICC for AR-Calibrated-by-Itself that deviates significantly from the rest. In terms of the ICC slopes, AR-Calibrated-by-Itself is the steepest. The ICCs of the AR-Calibrated-with-LR:A, -LR:B, and -RC are virtually identical. The ICC for the AR-Calibrated-with-the-Total-Test is the flattest. This order of slope steepness is true of all the AR items, even for AR Item 21 in Figure 2, which does not possess as dramatic an ICC departure for AR-Calibrated-by-Itself from the rest as shown in the previous figure.

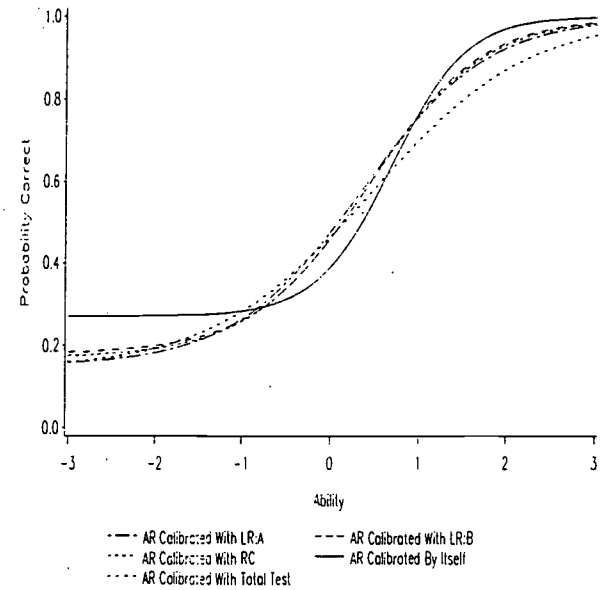

FIGURE 1. *ICC comparisons, item AR 5*

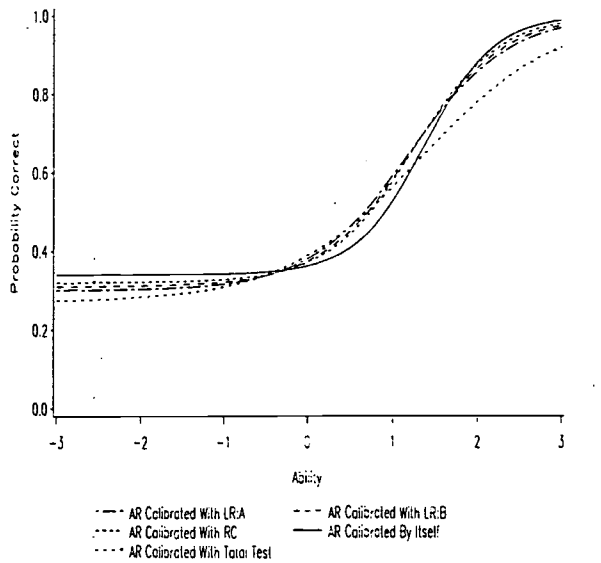

FIGURE 2. *ICC comparisons, item AR 21*

One thing that should be pointed out is that the AR-Calibrated-by-Itself ICCs are significantly different from the others, especially toward the lower-ability range, like AR Item 13 in Figure 3. Such a dramatic departure might have been caused by the increased standard error of measurement due to the small number of items in the calibration. Caution should be exercised in interpreting such ICC departures.

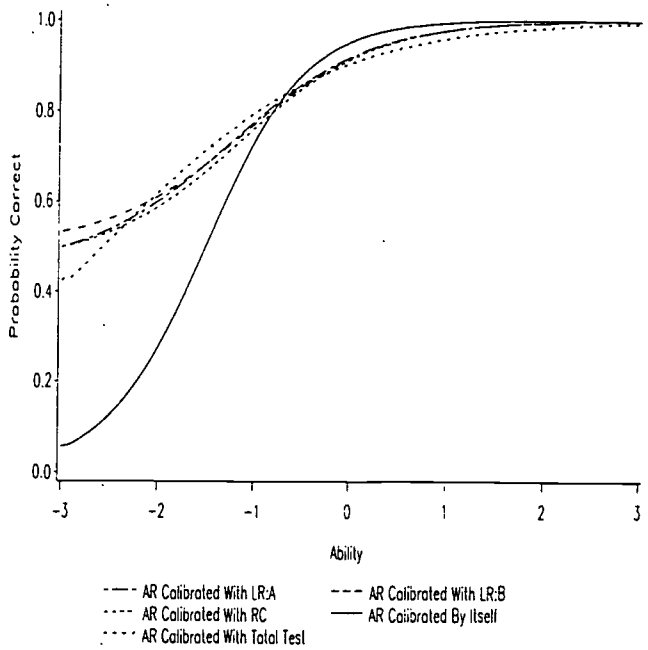

FIGURE 3. *ICC comparisons, item AR 13*

Being of the same item type, LR:A and LR:B items were not found to present significant departures among the five ICCs, except for some separation with very few items. Figure 4 shows the five ICCs for LR:A Item 23 which has a relatively large separation as compared with the rest of the LR items. The deviant ICC line stands for LR:A-Calibrated-With-AR which is the flattest line. The rest of the ICCs are all bundled together with the steepest ICC standing for LR:A-Calibrated-by-Itself. These same patterns are true of LR:B Item 16 in Figure 5 which also contains one of the biggest ICC separations among LR items.

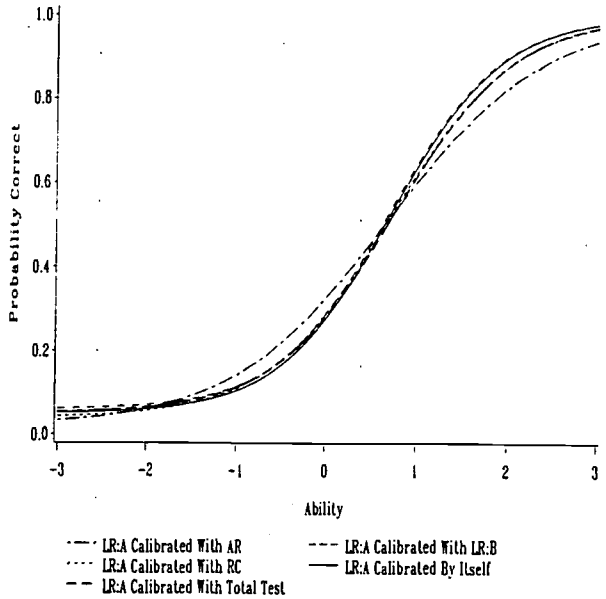

FIGURE 4. *ICC comparisons, item LR:A 23*

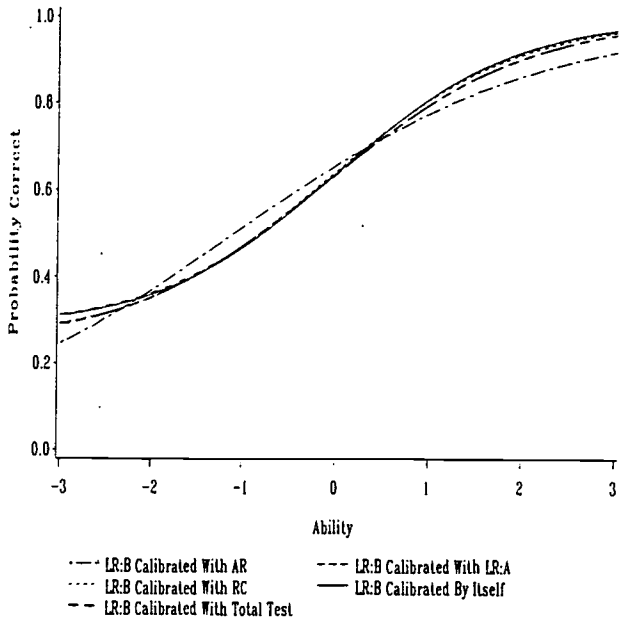

FIGURE 5. *ICC comparisions, item LR:B 16*

An overview of all the RC items revealed that four ICCs clump closely with one ICC branching out mostly toward the lower-ability range. According to Figures 6 and 7 for RC items 10 and 21, respectively, the ICC that stands out corresponds to RC-Calibrated-by-Itself and has the steepest slope. In retrospect, the ICC patterns of the RC items resemble those of AR items whose dissimilar ICCs were also the ones that were calibrated by themselves. Again, the smaller number of items included in the calibration may be at least partially responsible for these departures.

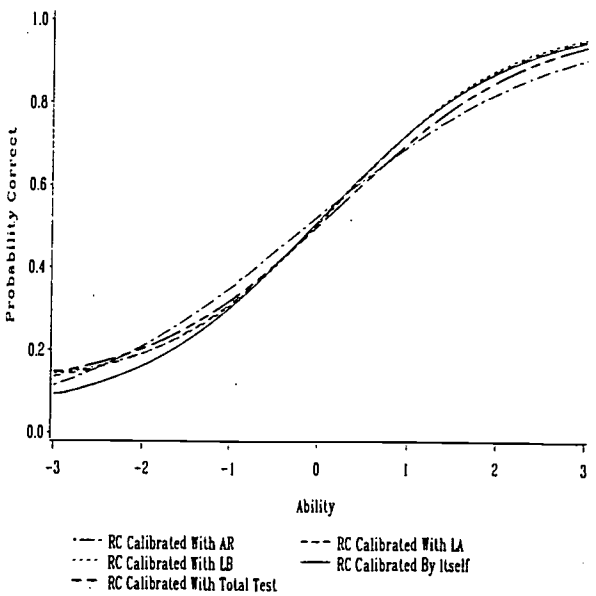

FIGURE 6. *ICC comparisons, RC item 10*

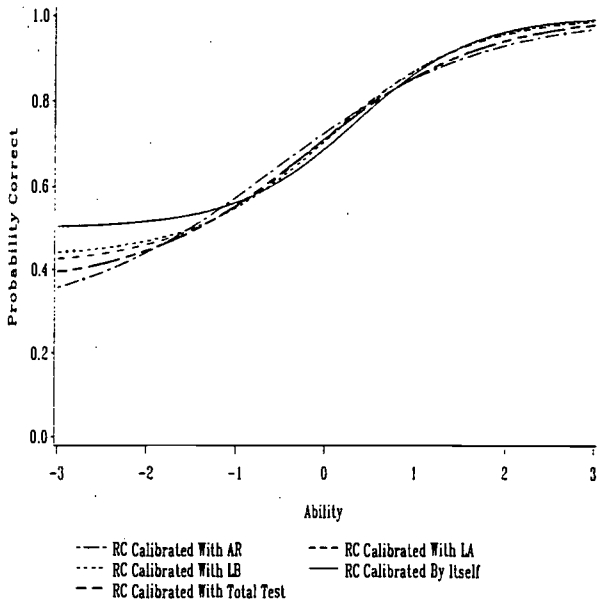

FIGURE 7. *ICC comparisons, item RC 21*

Four conclusions can be drawn from the above analyses. First, the two sections of the LR items seem to dominate when AR, LR:A, LR:B, and RC items are calibrated together. That is, both AR and RC items seem to have been compressed toward the LR dimension through the unidimensional calibration of BILOG. This conclusion is not only supported by the fact that LR constitutes half of the items, but also by the findings of the above analyses that both AR and RC items "stand out" with their increased slopes when they are calibrated alone. On the contrary, the slopes of the LR:A and LR:B items remain virtually the same when they are calibrated alone or in combination with all other items.

Under the condition of a similar number of items, both LR:A and LR:B items seem to be equally affected when calibrated with AR items. The extent that LR:A and LR:B items are affected when calibrated with RC items is minimal, which is similar to the extent that they are affected when calibrated with the entire test. This second finding is consistent with the first conclusion.

Third, the slopes of AR items are the flattest when calibrated with the total test. Under the condition of similar item numbers, AR items are affected in a similar way when calibrated with LR and RC items, respectively. Compared with LR and RC items, the slopes of AR items when calibrated alone increase the most sharply, as exemplified by Figure 2.

Fourth, when calibrated alone, the slopes of RC items increase, but not as much as those of AR items. The ICCs of RC items when calibrated with the total test are closer to those when calibrated with LR:A and LR:B items than it is to the ICC when calibrated with AR items.

In view of the above four conclusions, the approximate order of dimensional dominance of all items can be summarized as LR, the most dominant dimension; RC, the secondary dimension very similar to the LR section; and AR, the tertiary dimension. AR seems to be significantly different from LR and RC, as supported by previous studies. Whether or not LR and RC are significantly different from each other is an empirical question. There is some evidence from the study of Douglas, Kim, Roussos, Stout, and Zhang (1996) that LR and RC can be significantly different from each other under some comparison conditions.

Note that the small numbers of items (24 to 27) of Calibrations 8–11 do not seem to have as significantly affected the LR-by-Itself and RC-by-Itself calibrations as the AR-by-Itself calibration. The ICCs for LR:A and LR:B items when calibrated alone are virtually identical to their ICCs when calibrated with the total test.

Comparing True Scores at the Sectional Level

To what extent do the effects of multidimensionality observed at the item level show up in true scores at the sectional level? This part will compare the true scores based on each of the four item-type sections according to the varying degrees of dimensionality in their calibration data sets.

Figure 8 compares five true score lines for the AR section when it is calibrated with LA, LB, RC, all items and by itself. In comparison to the ICC patterns in Figures 1–3, Figure 8 shows that the true score line for the AR-Calibrated-with-the-Total-Test is apart from the other four closely bundled lines, especially for the relatively high-ability range. The AR true score of a relatively high-ability test taker on the AR section is lower when it is calibrated along with the total test. How much lower is it? Plotting the true score differences between the AR-Calibrated-by-Itself and the AR-Calibrated-with-the-Total-Test, Figure 9 shows that test takers of relatively high AR ability can differ as much as 1.6 points on their predicted AR true score. For test takers below the middle-ability scale, the difference ranges from positive 0.24 to negative 0.61 points.

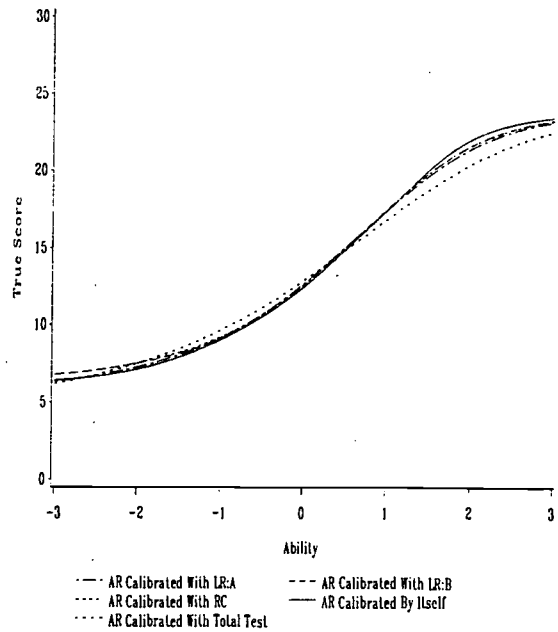

FIGURE 8. *True score comparison on the AR section*

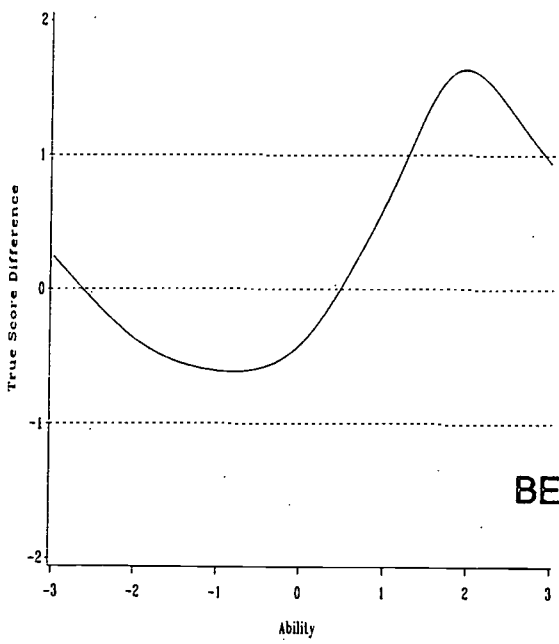

FIGURE 9. *AR true score differences between AR-calibrated-by-itself minus AR-calibrated-with-the-total test*

Figures 10 and 12 show that the effects of multidimensionality on LR:A and LR:B are virtually identical, which is expected because they are of the same item type and specifications. Both LR:A and LR:B true scores are similarly affected when calibrated with the AR items. How much effect is there? As shown in Figures 11 and 13, multidimensionality would increase the true scores of the below-middle-ability range by as much as half a point, but decrease the true scores of the above-middle-ability test takers by a maximum of 1.5 points.

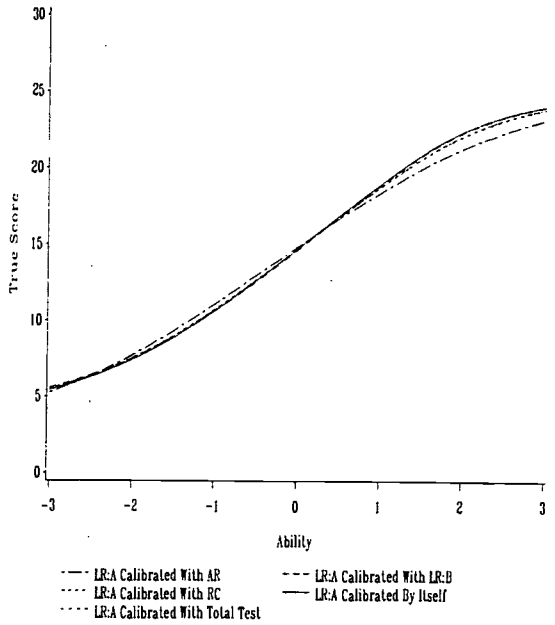

FIGURE 10. *True score comparison on the LR:A section*

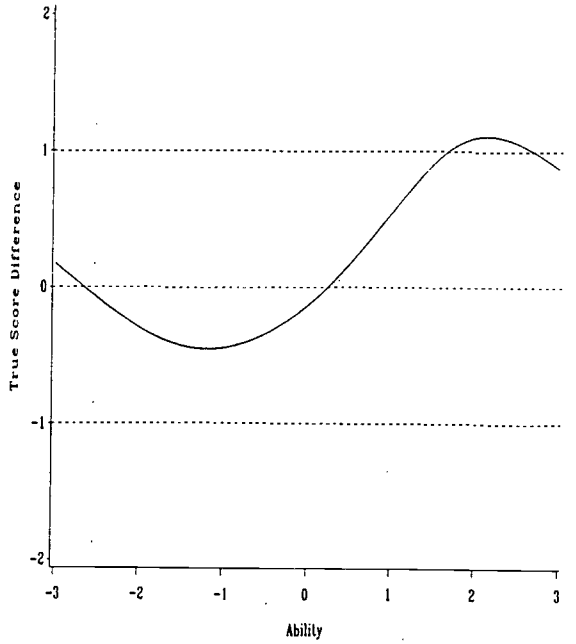

FIGURE 11. *LR:A true score differences between LR:A-calibrated-by-itself minus LR:A-calibrated-with-AR*

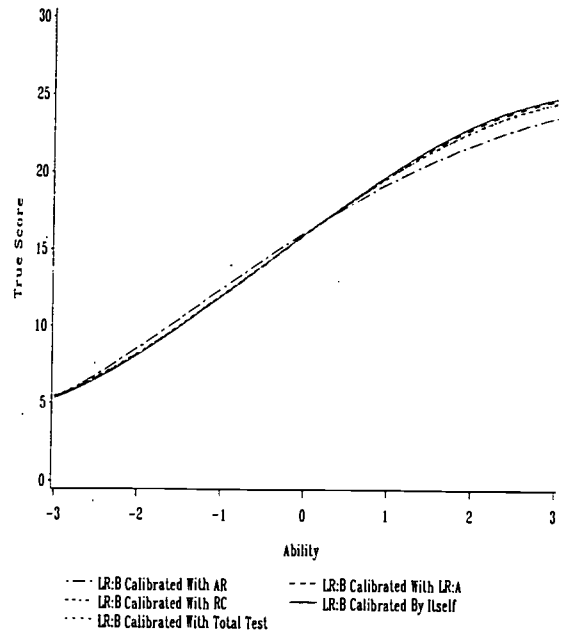

FIGURE 12. *True score comparisons on the LR:B section*

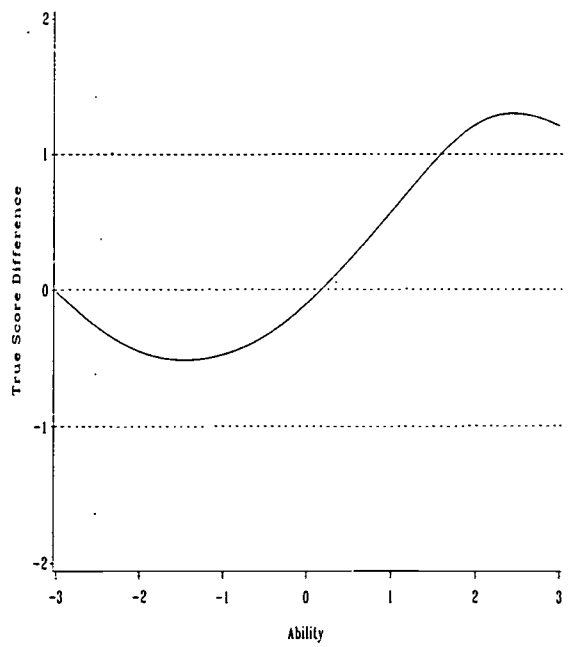

FIGURE 13. *True score differences between LR:B-calibrated-by-itself minus LR:B-calibrated-with-AR*

As for the RC section, the five true score lines in Figure 14 are somewhat separate. As expected, the true score line for the RC-Calibrated-by-Itself differs the most from the RC-Calibrated-with-AR, the second from RC-Calibrated-by-the-Total-Test, and finally from the RC-Calibrated-with-LR:A and LR:B. Figure 15 shows the true score differences between the RC-Calibrated-by-Itself and the others. It can be seen that the effects on true scores range from 0.77 points for the extremely low-ability test takers, to 0.51 points for relatively low-ability test takers to as high as 1.58 points for above-middle-ability test takers, depending on whether or not the RC section is calibrated by itself or with the AR section.

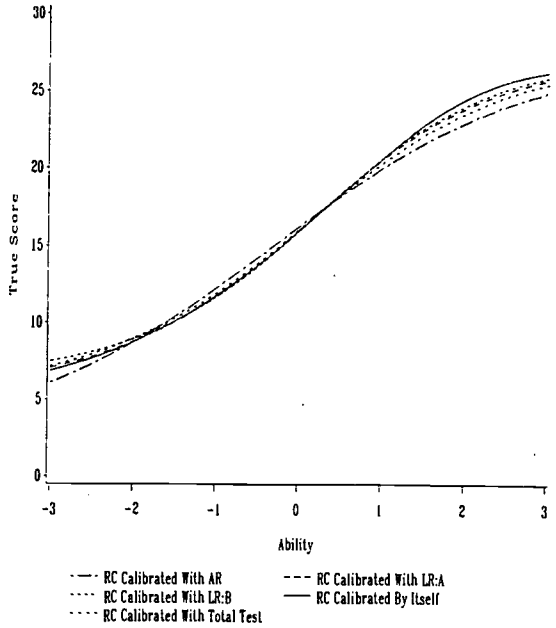

FIGURE 14. *True score comparisons on the RC section*

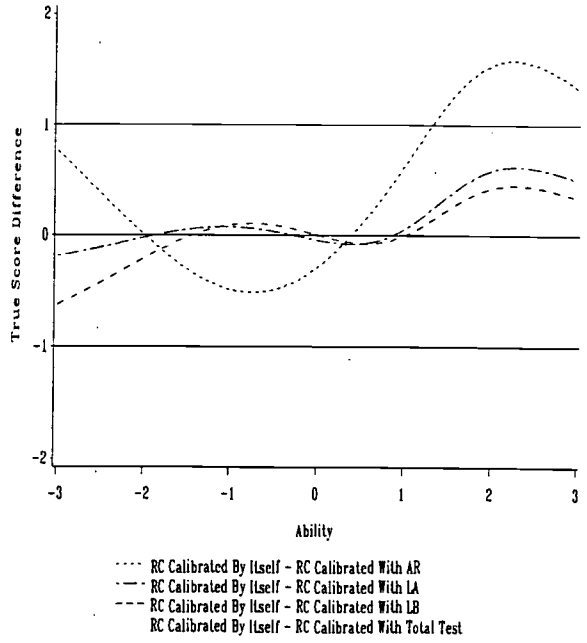

FIGURE 15. *True score differences on the RC section*

Figure 15 also shows the effects of multidimensionality on the RC section when it is calibrated with the LR:A and LR:B sections, respectively, and with the total test. The maximum effects on the RC section when calibrated with the LR:A and LR:B are about half a point for relatively higher-ability test takers. The effects on the RC section when calibrated with the total test vary between half a point and 1 point.

In view of the above findings, two conclusions can be offered. First, as expected, the effects of multidimensionality at the sectional level are the biggest when one section representing one dimension is calibrated with one section representing the second dimension in terms of numbers, such as the AR section calibrated with the LR:A, LR:B, or RC sections, respectively. Calibrating with the total test has the second

biggest effects for the AR and RC sections, but minimal effect for the LR:A and LR:B sections because LR:A and LR:B dominate the other item types, as pointed out earlier.

The second conclusion is that multidimensionality seems to affect the estimation of the true score in the high-ability range more than in the low-ability range. More specifically, multidimensionality lowers the true score curve in the high-ability range more than it raises the curve in the lower-ability range.

Comparing True Scores Based on Two Sections

Having shown the aggregate effect of multidimensionality at the level of individual sections, one might wonder how much bigger the effect of multidimensionality may be when two or more sections are combined. Figure 16 displays three true score lines for the combination of the LR:A and LR:B sections after being calibrated two-dimensionally, essentially unidimensionally and unidimensionally. It can be seen that the unidimensional and the essentially unidimensional true score lines are very close to each other, but apart from and steeper than their two-dimensional true score line. According to Figure 17, the unidimensional true score line differs from the two-dimensional true score line by about 1 point for the relatively low-ability range, and by slightly more than 2 points for the relatively high-ability range. The maximum difference between the unidimensional and the essentially unidimensional true score lines is negligible, about 0.4 points for the relatively high-ability range.

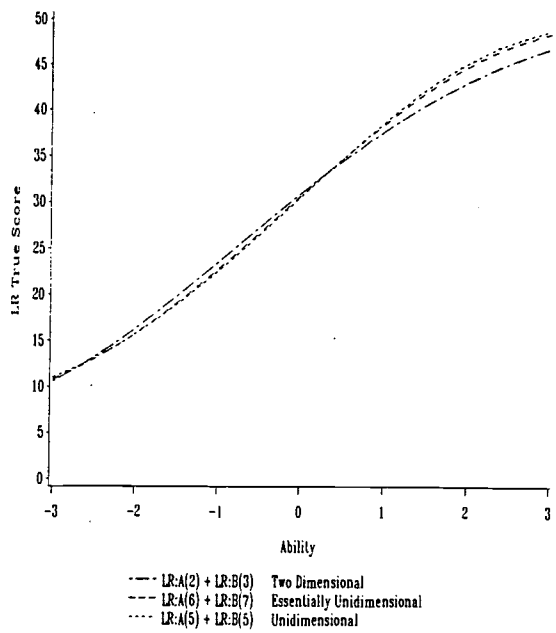

FIGURE 16. *True score comparisons on two sections: LR:A+LR:B*

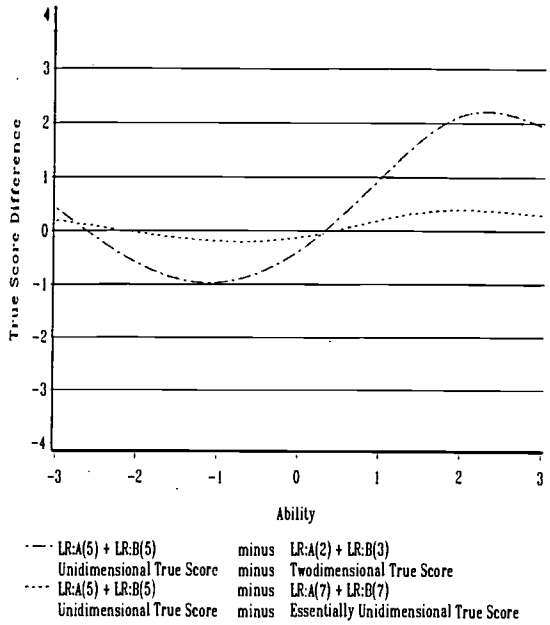

FIGURE 17. *True score differences on two sections: LR:A+LR:B*

Two important features about Figures 16–17 need to be emphasized. First, LR:A and LR:B are of the same item type and specifications, and naturally, one dimensional. Second, although the LR:A and LR:B sections come from five different calibrations, each of the five calibrations has a similar number of items. Because of these two features, it can be concluded that the 1 to 2 point differences observed in Figure 17 can be attributed mostly to the effect of multidimensionality.

Comparing True Scores Based on Three Sections

The three true score lines in Figure 18 are formed by adding the RC section to what consists of the previous Figure 16, as the RC section is calibrated two-dimensionally, essentially unidimensionally and unidimensionally. The pattern of the three score lines in Figure 18 presents a similar pattern as that of Figure 16—the slopes of the unidimensional and essentially unidimensional true score lines are steeper than that of the two-dimensional true score line. The differences between the unidimensional and two-dimensional lines vary from about 1.4 points to about 3.7, as shown in Figure 19. The differences between the unidimensional and the essentially unidimensional lines remain the same for the lower half of the ability range, but increase to 1 point for the relatively high-ability range.

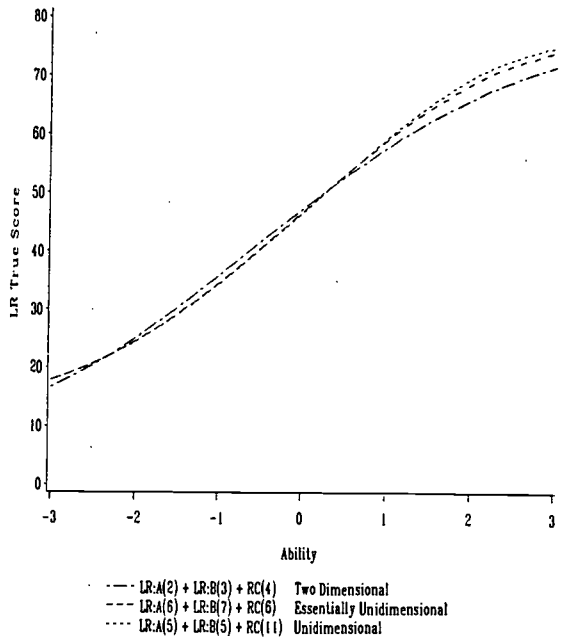

FIGURE 18. *True score comparisons on three sections: LR:A+LR:B+RC*

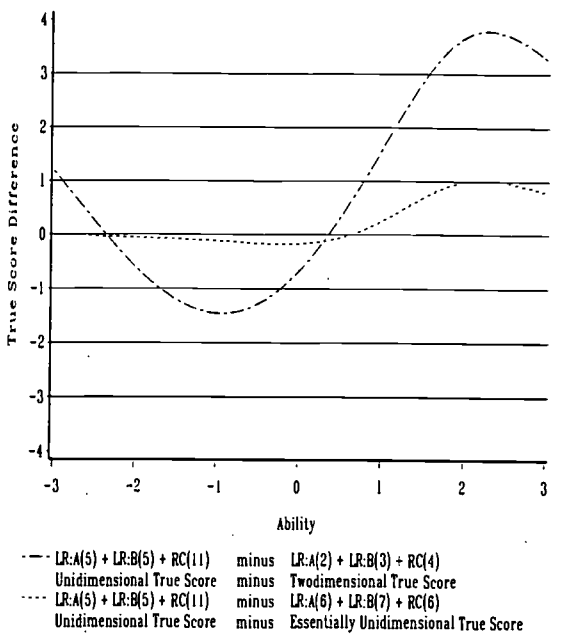

FIGURE 19. *True score differences on three sections: LR:A+LR:B+RC*

Comparing True Scores at Total Test Level

Adding the AR section to what constitutes Figure 18 above forms Figure 20. Figure 20 represents three true score lines for the four sections combined, corresponding to the three cases where each section was calibrated two dimensionally, essentially unidimensionally and unidimensionally, respectively. What is the total effect of these types of multidimensional calibrations on the four sections combined? Figure 21 shows that the maximum differences between the unidimensional vs. the two dimensional calibrations can be as much as 4.3 points for the relatively high-ability range and 1.6 points for the relatively low-ability range.

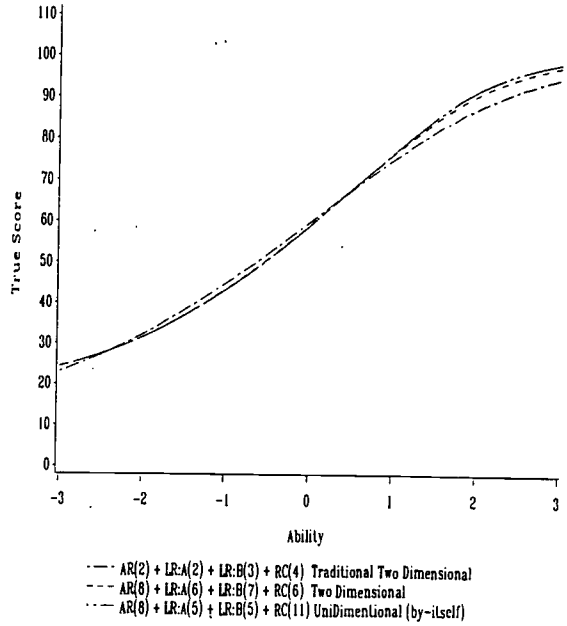

FIGURE 20. *True score comparisons on total test lengths*

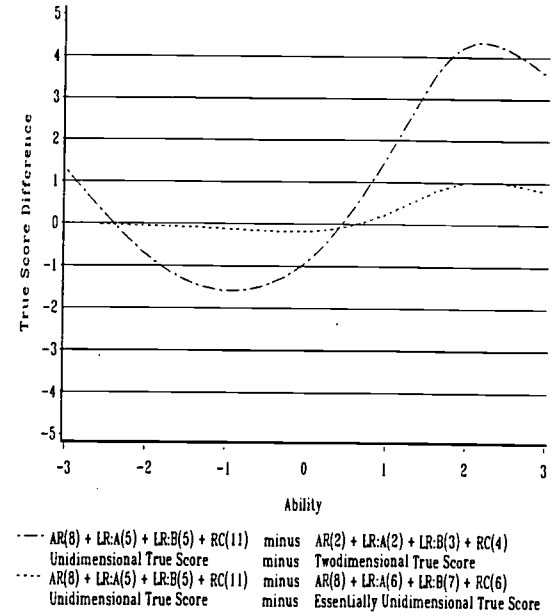

FIGURE 21. *True score differences on the total LSAT*

Figure 22 further compares the two-dimensional and the unidimensional true score lines of Figure 20 with another two true score lines. The first one represents the true score line when the four sections are calibrated as a single unit (the essentially two-dimensional case). The second true score line stands for the unidimensional case where each of its four sections is calibrated separately.

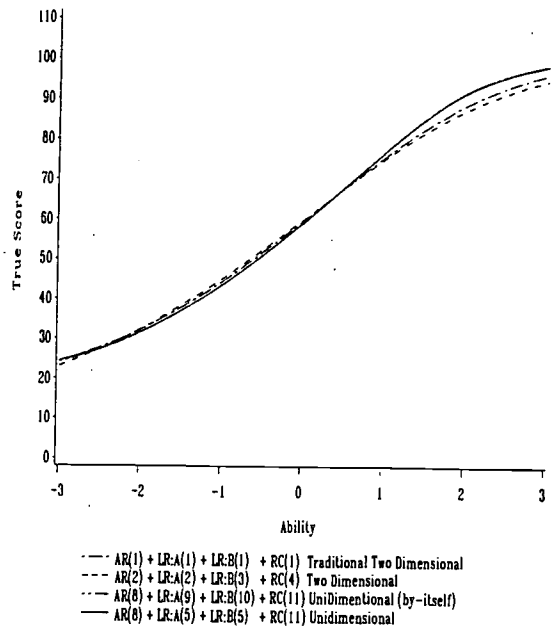

FIGURE 22. *Other true score comparisons on the total LSAT*

As expected, the unidimensional and the unidimensional-by-itself true score lines are virtually identical. Shown in Figure 23, the essentially two dimensional true score is only 1.5 points higher than the two-dimensional true score line for the high-ability range. Two conclusions can be drawn from these findings. The first finding serves as another confirmation that the small number of items used in the by-itself calibrations have not affected the quality of calibrations, since the unidimensional and unidimensional by-itself true score lines differ only in the number of items used in their calibrations. Second, the reason that the essentially two-dimensional true score line has a steeper slope than the other two-dimensional true score lines is that the data set for the former is mainly dominated by the reading/informal reasoning dimension, while the data sets for the latter are more balanced between the two dimensions.

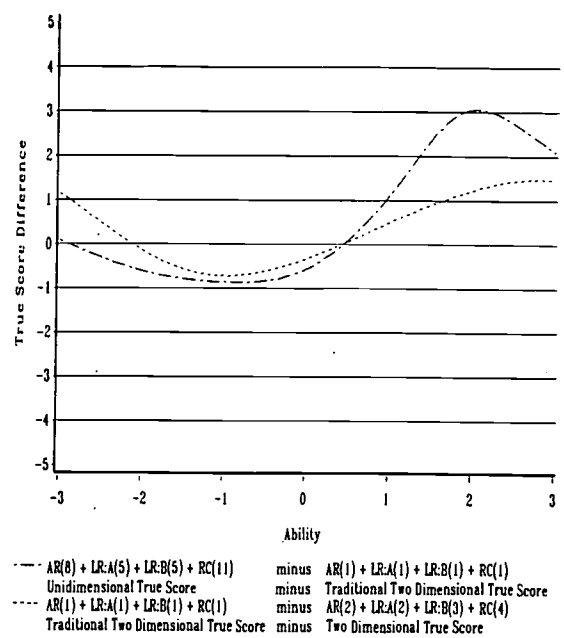

FIGURE 23. *Other true score differences on the total LSAT*

Conclusion and Discussion

Forms of the current paper-and-pencil LSAT are constructed to be as parallel as possible. As such, the effects of multidimensionality on item calibration is kept as constant as possible across administrations. In a CAT environment, however, this approach may not be possible. Because of the tailored nature of a CAT, some items might be calibrated under essentially unidimensional conditions and then subsequently be administered within a multidimensional setting.

This study has systematically investigated the effects of multidimensionality on estimating both ICCs and true scores. It has been shown by systematic manipulation of the two dimensions that such dimensions can exert some noticeable effects on ICC slopes and true scores in spite of the high correlations between the two dimensions. Specifically, to some extent multidimensionality reduces item slopes and decreases the estimated true score curve in the high-ability range. The estimated true score reduction is about 1 point at the sectional level of 24 items and about 4 points at the total test level of 102 items. It can be inferred that more diverse multidimensional structures may exhibit much bigger effects on ICCs and true scores. An important aspect of the results of this study is that multidimensionality seems to have a greater effect at the high-ability range than in the low-ability range. For these items, these differences are within two standard error of measurement for a 102-item test, and as such might be considered negligible. However, within a CAT framework in which true scores are estimated from fewer items, these effects might be more significant.

The significance of this study lies in assessing the extent to which ICCs and true score estimates vary for the same items when they are calibrated with item responses that differ with respect to dimensional structure. This is an important issue associated with CAT. Tailored to individual test takers' ability, a given CAT administration may be composed of a number of items that originate from a number of test forms. How multidimensional the test forms are and how these forms are originally calibrated have been shown in this paper to have noticeable effects on ICCs and true scores, and consequently will affect assessment accuracy. This situation is even more noteworthy if a test possesses more divergent dimensions than were studied here.

This paper has at least two shortcomings. First, it has investigated the impact of multidimensionality only with regard to fixed-length tests, but not on tests of variable lengths. Second, this study stops short of assessing the impact of multidimensionality on the accuracy of ability estimates for individual test takers. The authors intend to continue research in these two directions using both simulated and real data.

The advent of CAT has opened new horizons with respect to testing practices, as well as psychometrics in general. The issues of the effects of multidimensionality on item and test characteristics as well as assessment accuracy will become more salient and attract more attention from both psychometricians and practitioners alike. It is hoped that the results from this study will provide valuable information regarding how and to what extent multidimensionality would impact CAT form assembly, and these results will foster future research in the area.

References

Ackerman, T. (1994, April). *Graphical representation of multidimensional IRT analysis*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Camilli, G., Wang, M. M., & Fesq (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 1*, 79-96.

De Champlain, A. (1995). *Assessing the effect of multidimensionality on LSAT equating for subgroups of test takers* (Statistical Report 95-01). Newtown, PA: Law School Admission Council.

De Champlain, A. (1996). Assessing the effect of multidimensionality on IRT true score equating for subgroups of examinees. *Journal of Eucational Measurement*.

Doody, E.N. (1985). *Examining the effects of multidimensional data on ability and item parameter estimation using the three-parameter logistic model* (Report No. TM 850 360). Monterey, CA: CTB/McGraw-Hill. (ERIC Document Reproduction Service No. ED 258 992).

Dorans, N.J., & Kingston, N.M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 4*, 249-262.

Douglas, J., Kim H., Roussos, L., Stout, W. & Zhang, J., (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992 administrations* (Statistical Report 95-05). Newtown, PA: Law School Admission Council.

Drasgow, F., & Parsons, C.K. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Fraser, C., & McDonald, R. P. (1988). *NOHARM: a computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. New South Wales, Australia: Center for Behavioral Studies, the University of New England.

Hambleton, R.K., Zaal, J.N., & Pieters, J.P.M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R.K. Hambleton and J.N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 341-366). Boston: Kluwer Academic Publishers.

Hsu, T.C., & Yu, L. (1989). Using computers to analyze item response data. *Educational Measurement: Issues and Practice, 8*, 21-28.

Lord, F.M. (1977). Practical applications of characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

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.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software, Inc.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming and equating. In R.L. Linn (Ed.), *Educational Measurement* (pp. 221-262). New York: American Council on Education and Macmillan Publishing Company.

Reckase, M.D., Carlson, J.E., Ackerman, T.A., & Spray, J.A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data*. Paper presented at the annual meeting of the Psychometric Society, Toronto, Ontario, Canada.

Roussos, L. (1996). Personal communication.

Roussos, L., & Stout, W.F. (1994, April). *Analysis and assessment of test structure from the multidimensional perspective*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum Associates.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").