The effect of dimensionality on an adaptive test's ability estimation was examined. Two-dimensional data sets, which differed from one another in the interdimensional ability association, the correlation among the difficulty parameters, and whether the item discriminations were or were not confounded with item difficulty, were generated for 1,600 simulated examinees. The generated data were used for Bayesian computerized adaptive testing (CAT) simulations (three-parameter logistic model), and the CAT ability estimates were compared with the simulated examinees' known abilities. The dimensionality of response data shifted the focus for the minimization of measurement errors from known abilities (with unidimensional data) to the average of the latent abilities (with bidimensional data). Three tables and 24 graphs summarize the study results. (Author/SLD)
The Influence of Dimensionality on CAT Ability Estimation

R.J. De Ayala

University of Maryland

Please send correspondence to:

R.J. De Ayala
Measurement, Statistics, and Evaluation
Benjamin Building
University of Maryland
College Park, MD 20742

Paper presented at the NCME Annual Meeting, April 1990, Boston
ABSTRACT

This study examined the effect of dimensionality on an adaptive test's ability estimation. Two-dimensional data sets were generated which differed from one another in the interdimensional ability association, the correlation among the difficulty parameters, and whether the item discriminations were or were not confounded with item difficulty. The generated data were used for Bayesian CAT simulations (three-parameter logistic model) and the CAT ability estimates were compared with the simulees known abilities ($\theta_T$s). Results show that dimensionality of the response data shifts the focus for the minimization of measurement errors from $\theta_T$ (with unidimensional data) to the average of the latent abilities (with bidimensional data).
Computerized adaptive testing (CAT) is concerned with the minimization of measurement errors in the estimation of an examinee's ability. To achieve this goal the examinee is administered items based on his or her current ability estimate. These items are selected such that the examinee is expected to have about a fifty percent chance of correctly answering the items. Some of CAT's benefits include equiprecise measurement throughout the ability continuum and adaptive tests which are shorter than the corresponding paper-and-pencil tests.

CATs typically are based on one of the dichotomous unidimensional IRT models, such as the three-parameter logistic (3PL) or Rasch models (e.g., McBride & Martin, 1983; Kingsbury & Houser, 1988). The development of the CAT item pool requires the identification of the data's dimensionality before fitting the IRT model. That is, although some items may be considered unidimensional, other test items may require more than one ability to obtain a correct response. For instance, correctly answering a mathematical word problem may be considered to be a function of reading and mathematical abilities. Implications of the violation of unidimensionality for CAT item pool development (e.g., equating, scale shrinkage) may be found in Doody-Bogan and Yen (1983) as well as in Yen (1985).

Multidimensional models have been developed in order to address the issue of multiple latent dimensions (e.g., McKinley & Reckase, 1983; Sympson, 1978). These models are classified as either compensatory or noncompensatory. Conceptually, a compensatory model is one in which an examinee's latent traits interact to produce a response to an item. This interaction may take the form of an examinee's facility in one latent trait (θ) compensating for a deficiency in another θ. In contrast, in a noncompensatory model the examinee's θs do not interact to yield a response. Although these models have been used in some research they have yet to obtain widespread acceptance or use in applications.
Given that the dimensionality assumption of unidimensional IRT subsumes the principle of local independence (Lord, 1980), violation of this assumption should affect the likelihood function used for parameter estimation. A number of studies (e.g., Ackerman, 1989; Way, Ansley, & Forsyth, 1988; Ansley & Forsyth, 1985; Reckase, 1979) have examined the effect of multidimensional response data on unidimensional IRT parameter estimates. These studies have been primarily concerned with the effects of dimensionality on the calibration of a multidimensional data set by either LOGIST (Wingerskey, Barton, & Lord, 1982) or BILOG (Mislevy & Bock, 1982). Although the models used for data generation differed, the results of these studies have found that dimensionality affects parameter estimation. In general, when a compensatory multidimensional IRT model was used for data generation $\hat{b}$ was found to be an estimate of the average of the true $b$s (Way et al., 1988), $\hat{a}$ was an estimate of the sum of $a_1$ and $a_2$ (Way et al., 1988), and ability estimates $\hat{\theta}$ to be an estimate of the average true $\theta$s (Ackerman, 1989; Way et al., 1988). In contrast, data generation using a noncompensatory model showed that $\hat{b}$ was an overestimate of or correlated more highly with $b_1$ than with $b_2$ (Ackerman, 1989; Way et al., 1988; Ansley & Forsyth, 1985), $\hat{a}$ was an estimate of the average of the true $a$s (Way et al., 1988; Ansley & Forsyth, 1985), and $\hat{\theta}$ to be an estimate of the average true $\theta$s (Ackerman, 1989; Way et al., 1988; Ansley & Forsyth, 1985). In general, these conclusions come from correlational analyses of the parameters with their estimates and an assessment of the accuracy of parameter estimation by the calculation of the mean absolute difference (a.k.a., MAD or AAD) across whichever was pertinent, examinees or items.

In general, studies which have investigated the operating characteristics of CAT have involved the simulation of unidimensional data and item pools (e.g., Weiss, 1982; McBride, 1977; Jensema, 1974). However, given "...that no actual psychological measurement instrument is likely to be exactly unidimensional..." the issue becomes one of
whether the "...instrument is sufficiently unidimensional to allow application of IRT" (Hulin, Drasgow, & Parsons, 1983, p. 40). In live testings, where the possibility of less than ideal unidimensional data may exist, the primary concern has been with the estimation of the reliability and validity of CAT (e.g., McBride & Martin, 1983; Weiss & Kingsbury, 1984). Further, because in these studies the examinee's true ability is unknown the influence of dimensionality on the accuracy of ability parameter estimation cannot be investigated.

This study investigated the effect of varying degreee of dimensionality on CAT ability estimation. That is, an adaptive test based on unidimensional item parameter was administered to an simulee who used more than one ability to respond. Two-dimensional data sets were generated which differed from one another in the interdimensional ability association, the correlation among the difficulty parameters, and whether the item discriminations were or were not confounded with item difficulty. This latter factor is included because of Reckase, Carlson, Ackerman, and Spray's (1986) finding that upper deciles of a unidimensional ability differ mainly on $\theta_2$ while at lower deciles the ability differed primarily on $\theta_1$ (cited in Ackerman, 1989). Simulees with known abilities were administered unidimensional tests and their abilities estimated on the basis of their multidimensional responses. In contrast to the studies mentioned above (i.e., Ackerman, 1989; Way et al., 1985; Ansley & Forsyth, 1985), the accuracy and bias of the $\hat{\theta}$s at various points along the ability continuum was assessed.

**METHOD**

*Data*: The data were generated according to a multidimensional 3PL (M3PL) model (Doody-Bogan & Yen, 1983). This model requires a set of multidimensional $\theta$s as well as a set of (multidimensional) item parameters. The multidimensional $\theta$s were generated such that the examinee's ability on dimension 1 ($\theta_1$) was evenly distributed between -3.0 and 3.0 using 0.4 logit interval between successive $\theta$ levels (i.e., for 100 examinees $\theta_1 = -3.0$, for
100 examinees \( \theta_1 = -2.6 \), etc.). The examinee's ability on the second dimension (\( \theta_2 \)) was derived from \( \theta_1 \) by using Hoffman's (1959) technique for generating correlated data. For each of the 160 simulees \( \theta_2 \) was obtained by randomly sampling a normal deviate (Z) from a unit normal curve and calculating:

\[
\theta_2 = \theta_1 + \left( \frac{k}{r} \right)Z
\]

\( k = \sqrt{1-r^2} \), and \( r \) is the desired intercorrelation between \( \theta_1 \) and \( \theta_2 \). Four interdimensional \( \theta \) correlations (\( r_{\theta_1 \theta_2} \)) were investigated from extreme bidimensionality to almost unidimensionality; values for \( r_{\theta_1 \theta_2} \) were 0.03, 0.30, 0.60, 0.90.

In the following an item parameter's subscript refers to a dimension. The difficulty parameters (\( b_1 \) and \( b_2 \)) were generated in a fashion analogous to the generation of \( \theta_1 \) and \( \theta_2 \). That is, the \( b_1 \) for sets of four items was fixed at every 0.1 logit between -3.5 and 3.5 (e.g., for 4 items \( b_1 = -3.5 \), for 4 items \( b_1 = -3.4 \), etc.). The \( b_2 \) for each of the 284 items was derived from the item's \( b_1 \) using the correlated generation method mentioned above. Three \( b_1 b_2 \) correlations (\( r_{b_1 b_2} \)) were used in the study, 0.03, 0.60, and 0.90.

The discrimination parameters (\( a_1 \) and \( a_2 \)) were created by randomly sampling from a uniform distribution with a minimum value of 0.20 and a maximum value of 1.8. This set of \( a_1 \) was combined with the three sets of \( b_1 \) to form three item pools where all item pools had the same set of \( a_1 \); this combination of the randomly ordered \( a_1 \) with the \( b_1 \) form the nonconfounding condition. The confounding between \( a_1 \) and \( b_1 \) was obtained by sorting \( a_1 \) into ascending order and sorting \( a_2 \) into descending order (cf., Ackerman, 1989). The pseudo-guessing parameter, \( c \), was set to 0.20.

The interdimensional correlations of 0.30, 0.60, and 0.90 were obtained from the literature (Ackerman, 1989; Way et al., 1988; Ansley and Forsyth, 1985); the \( r_{\theta_1 \theta_2} = 0.03 \) was used as an approximation to \( r_{\theta_1 \theta_2} = 0.0 \) because this latter value could not be used with the Hoffman's technique. The \( r_{b_1 b_2} = 0.03 \) was obtained from Yen (1985), whereas
the $rb_2 = 0.60$ ($rb_2 = 0.36$) and $rb_2 = 0.90$ ($rb_2 = 0.81$) were used to simulate moderate and high linear relationships. The minimum and maximum $as$ are the same as those in Ackerman (1984). The constant used for $c$ came from Way et al. (1988).

To summarize, the data generation was based on 6 different combinations of item parameters (3 levels of $rb_2$ by 2 levels of confounding) and four levels of interdimensional ability association. The crossing of these three factors produced 24 response data sets. For each data set the true $\theta_T$s plus the relevant 284 true item parameters were used to generate binary response strings with a random error component for each simulated examinee. Generation of the binary response strings was accomplished by calculating for a given $\theta_T$ pair and a given item the probability of a correct response according to the M3PL model. To create the random error component for a response, a random number was selected from a uniform distribution $[0,1]$ and compared to the calculated probability. If the random number was less than or equal to the calculated probability, then a response of 1 was produced (a correct answer), otherwise a 0 was generated (an incorrect response).

**Program**: A computer program was written that simulated a CAT based on the 3PL model and which used Bayesian ability estimation with Owens Bayes updating (i.e., Jensema's (1974) alpha technique) for item selection. The adaptive testing simulation was terminated when either of two criteria were met: a maximum of thirty items was reached or when a standard error of estimate (SEE) of 0.05 or less was obtained.

A unidimensional item pool was created for use with the Bayesian CAT. Discrimination, difficulty, and pseudo-guessing parameters were generated for 284 items. The discrimination ($a$) and pseudo-guessing ($c$) parameters were generated by random sampling from a uniform distribution with the following restrictions: ($a$) $a$ were restricted to the inclusive range of 0.80..2.00; and ($b$) $c$ were allowed to vary between 0.00 and 0.20. The difficulty parameters ($b$) were uniformly distributed between -3.5 and 3.5 (inclusive) with four items at each 0.10 of an interval (i.e., there were four items with $b=$
Dimensionality and CAT estimation

-3.5, four items with $b = -3.4$, etc.). The use of multiple items at each 0.1 interval was
done to ensure that items of appropriate difficulty would always be available for the
Bayesian CAT's ability estimation. These item parameters values are consistent with
desirable item pool characteristics (Patience and Reckase, 1980; Urry, 1977). Therefore,
to each of the 1600 examinees in each of the 24- multidimensional response data sets a
Bayesian CAT was administered.

**Analyses**: Analysis of the CAT simulations involved using root mean square error (RMSE),
bias, and correlations (Pearson product-moment, Spearman rank-order) between the $\theta$ and
$\theta_1, \theta_2$, and between $\theta$ and the average of $\theta_1$ and $\theta_2 (\bar{\theta})$. Descriptive statistics were
calculated on the number of items administered, the $\hat{\theta}$s as well as on various item pool
characteristics.

**RESULTS**

For the 0.03, 0.30, 0.60, and 0.90 interdimensional ability conditions the observed
correlations were -0.028, 0.303, 0.590, and 0.964, respectively. Table 1 shows the item
parameters' interdimensional correlations for the confounded and nonconfounded condi-
tions. As can be seen, for the desired $r_{b1b2}$ of 0.03, 0.60, and 0.90 the observed correla-
tions were 0.095, 0.678, and 0.946. In addition, for the confounded conditions the correla-
tion between $a_1$ and $b_1$ approached -1.00 and between $a_2$ and $b_1$ approximated 1.00. The
unidimensional item pool used for the CAT simulations had an average $a$ of 1.410 (median
of 1.421) and a mean $c$ of 0.102 ($\text{median}=0.101$). The Pearson product-moment correlation
between $a$ and $b$ for the unidimensional item pool was 0.077 (Spearman rank-order was
0.076).

Insert Table 1 about here

Table 2 shows the correlational analyses between $\theta$ and $\theta_1, \theta_2$, and $\bar{\theta}$ for the
nonconfounded conditions. As can be seen, for each level of the $r_{b1b2}$ factor the
association between CAT $\theta$ and $\theta_1$ and with $\theta_2$ became increasingly stronger as $r_{b1b2}$
increased. The intercorrelation between bs appeared to have a slight effect on the correlation between \( \theta \) and \( \theta_1 \) and between \( \theta \) and \( \theta_2 \). Further, there was a slight decrease in the average number of items administered with increasing intercorrelation between the bs. Although for the \( rb_{1b2} = 0.90 \) and \( r_{\theta_1\theta_2} = 0.90 \) conditions there were minimal differences between \( \hat{\theta} \), \( \hat{\theta}_1 \), and \( \hat{\theta}_2 \), for all combinations of the \( rb_{1b2} \) and \( r_{\theta_1\theta_2} \) factors the linear association between \( \hat{\theta} \) and \( \hat{\theta} \) was greater than for either \( \hat{\theta}_1 \) or \( \hat{\theta}_2 \).

Figure 1 shows the RMSE analysis for the three levels of \( rb_{1b2} \) and the \( r_{\theta_1\theta_2} = 0.03 \) and \( r_{\theta_1\theta_2} = 0.90 \) conditions; the differences in the plotted \( \theta \) values reflect the differences in the \( r_{\theta_1\theta_2} \) conditions. As can be seen, the RMSE with respect to \( \hat{\theta} \) was less than that of the RMSE of either \( \theta_1 \) or \( \theta_2 \) for all nonconfounded conditions. In fact, the RMSE with respect to \( \hat{\theta} \) for the \( r_{\theta_1\theta_2} = 0.90 \) condition is comparable to RMSE for when \( r_{\theta_1\theta_2} = 0.03 \), \( r_{\theta_1\theta_2} = 0.30 \), and \( r_{\theta_1\theta_2} = 0.60 \); the RMSE plots for these latter two conditions are the intermediate steps in the progression from \( r_{\theta_1\theta_2} = 0.03 \) RMSE plots to those of \( r_{\theta_1\theta_2} = 0.90 \). The RMSE with respect to \( \hat{\theta} \) decreased slightly as \( rb_{1b2} \) increased. As \( r_{\theta_1\theta_2} \) increased, the RMSE of \( \theta_1 \) or \( \theta_2 \) approached that of \( \hat{\theta} \).

As would be expected from a Bayesian CAT, the CAT overestimated low ability on \( \theta_1 \) and \( \theta_2 \) (i.e., \( \theta_T < -2.0 \)) and underestimated high ability on \( \theta_1 \) and \( \theta_2 \) (i.e., \( \theta_T > 2.0 \)). Figure 2 shows the bias plots for the nonconfounded conditions presented in Figure 1. As \( r_{\theta_1\theta_2} \) increased the bias with respect to \( \theta_1 \) and \( \theta_2 \) decreased. For all combinations of the \( r_{\theta_1\theta_2} \) and \( rb_{1b2} \) factors, minimal bias was obtained when \( \hat{\theta} \) was considered an estimate of \( \hat{\theta} \). The \( rb_{1b2} \) factor does not appear to have a meaningful effect on bias for \( \theta_1 \), \( \theta_2 \), and \( \hat{\theta} \).
Table 3 presents the results from the confounded conditions. As was the case with the nonconfounded condition, \( \hat{\theta} \) is more highly related to \( \theta \) than to either \( \theta_1 \) or \( \theta_2 \). In general, \( r_{\theta_1} \) tends to be larger than \( r_{\theta_2} \) for \( r_{\theta_1 \theta_2} \) values of 0.03 and 0.30, whereas for the \( r_{\theta_1 \theta_2} = 0.60 \) and \( r_{\theta_1 \theta_2} = 0.90 \) conditions the opposite is true. Unlike the nonconfounded condition, when \( r_{\theta_1 \theta_2} = 0.90 \) the \( r_{\theta_2} \) and \( r_{\hat{\theta}} \) correlations are more similar to one another and higher in magnitude than \( r_{\theta_1} \). Further, for all combinations of the \( r_{b_1 b_2} \) and \( r_{\theta_1 \theta_2} \) factors the average test length in the confounded condition was slightly less than the corresponding nonconfounded condition test length. The pattern of decreasing test length with increasing \( r_{b_1 b_2} \) association was not as evident with the confounded condition as it was under the nonconfounded condition.

Insert Table 3 about here

Inspection of the confounded conditions' RMSE plots showed the same relationship between \( \hat{\theta}, \theta, \theta_1, \) and \( \theta_2 \); Figure 3 contains the confounded condition sample RMSE plots for the same conditions presented in Figure 1. For the \( r_{b_1 b_2} = 0.03 \) and \( r_{b_1 b_2} = 0.60 \) conditions and for the approximate range \(-2.0 \leq \theta \leq 2.0\), the RMSE for the confounded conditions are lower than those for the nonconfounded conditions, regardless of the \( r_{\theta_1 \theta_2} \) condition; as \( r_{b_1 b_2} \) increases the difference in RMSEs diminishes. As was the case for the nonconfounded condition, the RMSE of \( \theta_2 \) was less than that of \( \theta_1 \) for high ability examinees for the \( r_{b_1 b_2} = 0.03 \) condition. In contrast to the nonconfounded condition, the RMSE with respect to \( \theta_1 \) was less for lower ability examinees than that of the RMSE \( \theta_2 \). For all combinations of interdimensional ability and difficulty association the RMSE of \( \theta \) was less than that of \( \theta_1 \) and \( \theta_2 \).

Insert Figure 3 about here

Figure 4 presents the corresponding bias plots to those in Figure 2's, but for the confounded condition. As can be seen, compared to the nonconfounded condition there was
less bias for $\theta_1$ at low abilities, but no meaningful difference at upper abilities. Although at the $r_{b_1b_2} = 0.33$ and $r_{b_1b_2} = 0.60$ conditions there is no difference between the confounded and nonconfounded conditions in bias with respect to $\theta_2$, for the $r_{b_1b_2} = 0.90$ there was an increase in bias for $\theta < -1.0$. For all interdimensional difficulty levels there was an increase in bias for $\theta_2$ in the $\theta$ range 1.0 to 3.0. In general, as $r_{\theta_1\theta_2}$ increased, this pattern was evident, although with decreasing levels of bias in the estimation of $\theta_1$ and $\theta_2$. As was the case with the nonconfounded condition, the bias in $\bar{\theta}$ with respect to $\bar{\theta}$ was less than that of estimating either $\theta_1$ or $\theta_2$, except when $r_{\theta_1\theta_2} = 0.90$. In this latter condition, the differences in bias with respect to $\theta_1$, $\theta_2$ and $\bar{\theta}$, may not be considered meaningful by some; for this $r_{\theta_1\theta_2}$ condition there does not appear to be any difference in bias between the confounded and nonconfounded conditions. For $r_{\theta_1\theta_2} = 0.90$ and regardless of $r_{b_1b_2}$ level, the CAT overestimated low ability more than it underestimated high ability.

As stated above, for the nonconfounded condition there was a slight decrease in the average number of items administered with increasing $r_{b_1b_2}$, although this pattern was not as evident with the confounded condition. Calculation of the average number of items administered at each of the 16 levels of $\theta$ showed that, in general, shorter tests were administered for $\theta < 0.0$ (e.g., average test lengths of 15-16 items depending on the condition) to longer tests for $\theta > 2.0$ (e.g., mean test lengths of 17-20 items depending on particular data set; the $r_{b_1b_2} = 0.03$, $r_{\theta_1\theta_2} = 0.90$ condition had an atypical mean test length of 22 items for $\theta = 3.0$). With increasing $r_{b_1b_2}$ and $r_{\theta_1\theta_2}$ the mean test lengths became less variable across $\theta$. Of the 38,400 adaptive tests simulated the absolute maximum and minimum test lengths were 28 and 11 items, respectively.

Conclusion and Discussion

In general, increasing interdimensional difficulty association produced a slight decrease on test length and an increase in the accuracy of ability estimation as assessed
Dimensionality and CAT estimation

by RMSE. The associations between \( \theta \) and \( \xi, \theta_1, \) and \( \theta_2 \) increased as the correlation between interdimensional difficulties and interdimensional ability increased. The largest associations were between \( \theta \) and \( \xi \); 0.957 and 0.961 for the nonconfounded and confounded conditions, respectively. For comparative purposes, a Bayesian CAT (maximum test length of 20 items and termination SEE of 0.05) using a unidimensional data set (generated according to the 3PL model and using this item pool) had a \( r_{\theta\theta} \) of 0.988 (for both Pearson and Spearman coefficients) and an average test length of 15.613.

When discrimination was confounded with difficulty, the ability estimates showed a differential association with one of the two latent traits, however, the correlation between \( \theta \) and \( \xi \) was always greater than that of \( r_{\theta\theta_1} \) and \( r_{\theta\theta_2} \). For all combinations of the \( rb \) and \( \theta_1 \theta_2 \) factors the correlation between \( \theta \) and \( \xi \) for the confounded condition was always greater than for the correlation for the corresponding nonconfounded condition.

From the results of the studies on the effects of dimensionality on the calibration of compensatory multidimensional data it may be hypothesized that the finding that \( \hat{\theta} \) was an estimate of the average true \( \theta_s \) was, in part, a result of the fact that \( \hat{\xi} \) was an estimate of the average of the true \( \xi_s \). That is, because \( \theta \) and \( \xi \) are on the same scale, when the separate dimensions are collapsed in the estimation of \( \theta \), the subsequent stage of estimation will also reflect the collapsed difficulty scale; both BILOG and LOGIST obtain \( b \)'s prior to estimating \( \theta \). However, given that in CAT the item parameters are assumed true when the collapsing of the two difficulty scales does not account for \( \hat{\theta} \) being an estimate of the average \( \theta \)'s.

Conceptually, the item pool may be considered to have come from the calibration of a unidimensional data set. However, the results should be generalizable to those situations where item parameters are obtained from data which are not truly unidimensional (i.e., the situations investigated by Ackerman, 1989; Way, et al., 1988; Ansley & Forsyth, 1985). For item selection it is the distribution of \( b \) and the magnitude of \( a \) and \( c \).
which are important; CAT makes no distinction with respect to whether \( \hat{\theta} = b \) or \( \hat{\theta} = (b_1 + b_2)/2 \) and \( \hat{\alpha} = a \) or \( \hat{\alpha} = a_1 + a_2 \).

As stated above, CAT is concerned with minimizing the measurement errors associated with the estimation of an examinee's ability. It was shown that the dimensionality of the response data shifts the focus for the minimization of measurement errors from \( \theta_T \) (with unidimensional data) to the average of the latent abilities (with bidimensional data). Although the results may be considered problematic by some, there may be situations where one is only interested in ordering examinees on their ability to perform or solve certain types of problems and not in ordering them on the separate latent abilities which may be required to solve the problems. For example, on a statistics exam the instructor may only be interested in a student's understanding of the appropriateness and use of t-tests. The problems may be stated as word problems and require stating the appropriate statistical hypotheses, identification of and calculating the relevant t-statistic, arriving at conclusions concerning the truth or falsity of hypotheses, etc. Most likely the instructor is not interested in the student's standing on the separate abilities required to answer the problem (e.g., his or her reading ability, math ability, etc), but in the student's understanding of t-tests. Reckase, Ackerman, and Carlson (1988) have concluded that IRT's unidimensionality assumption does not necessarily require test items to measure a single ability, but rather the unidimensionality assumption requires the test items to measure the same composite of abilities. For this study, this composite of abilities was the average of \( \theta_1 \) and \( \theta_2 \).
REFERENCES


Table 1. Item parameters interdimensional correlations$^a$.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Item</th>
<th>Parameter</th>
<th>$a_1$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{b1b2} = 0.03$</td>
<td>$a_1$</td>
<td>-0.990</td>
<td>-0.995</td>
<td>-0.088</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.032)</td>
<td>(-0.022)</td>
<td>(0.055)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$a_2$</td>
<td>0.997</td>
<td>0.112</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.014)</td>
<td>(-0.034)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1$</td>
<td>0.095</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.095)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{b1b2} = 0.60$</td>
<td>$a_1$</td>
<td>-0.990</td>
<td>-0.997</td>
<td>-0.677</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.032)</td>
<td>(-0.022)</td>
<td>(0.027)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$a_2$</td>
<td>0.995</td>
<td>0.671</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.014)</td>
<td>(-0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1$</td>
<td>0.678</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.678)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r_{b1b2} = 0.90$</td>
<td>$a_1$</td>
<td>-0.990</td>
<td>-0.997</td>
<td>-0.944</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.032)</td>
<td>(-0.022)</td>
<td>(-0.002)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$a_2$</td>
<td>0.995</td>
<td>0.940</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.014)</td>
<td>(0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1$</td>
<td>0.946</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.946)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$Pearson product-moment correlations for confounded and (nonconfounded) conditions.
### Table 2. Intercorrelations$^a$ between $\hat{\theta}$ and $\theta_1, \theta_2, \theta_0$ and the average number of items administered (Mean NIA) for the nonconfounded conditions.

<table>
<thead>
<tr>
<th>Item Pool Characteristics</th>
<th>$r_{\hat{\theta}1}$</th>
<th>$r_{\theta2}$</th>
<th>$r_{\hat{\theta0}}$</th>
<th>Mean NIA (SD NIA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>0.500</td>
<td>0.645</td>
<td>0.821</td>
<td>17.206 (2.577)</td>
</tr>
<tr>
<td></td>
<td>(0.518)</td>
<td>(0.520)</td>
<td>(0.785)</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.611</td>
<td>0.752</td>
<td>0.844</td>
<td>17.187 (2.649)</td>
</tr>
<tr>
<td></td>
<td>(0.630)</td>
<td>(0.660)</td>
<td>(0.825)</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.741</td>
<td>0.816</td>
<td>0.875</td>
<td>17.146 (2.587)</td>
</tr>
<tr>
<td></td>
<td>(0.751)</td>
<td>(0.825)</td>
<td>(0.854)</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.890</td>
<td>0.893</td>
<td>0.900</td>
<td>17.408 (2.773)</td>
</tr>
<tr>
<td></td>
<td>(0.890)</td>
<td>(0.891)</td>
<td>(0.891)</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.513</td>
<td>0.694</td>
<td>0.866</td>
<td>16.896 (2.390)</td>
</tr>
<tr>
<td></td>
<td>(0.539)</td>
<td>(0.564)</td>
<td>(0.840)</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.678</td>
<td>0.779</td>
<td>0.903</td>
<td>16.878 (2.440)</td>
</tr>
<tr>
<td></td>
<td>(0.697)</td>
<td>(0.686)</td>
<td>(0.888)</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.799</td>
<td>0.849</td>
<td>0.924</td>
<td>16.703 (2.442)</td>
</tr>
<tr>
<td></td>
<td>(0.801)</td>
<td>(0.863)</td>
<td>(0.902)</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.922</td>
<td>0.929</td>
<td>0.934</td>
<td>16.926 (2.543)</td>
</tr>
<tr>
<td></td>
<td>(0.921)</td>
<td>(0.915)</td>
<td>(0.924)</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.552</td>
<td>0.723</td>
<td>0.914</td>
<td>16.407 (2.249)</td>
</tr>
<tr>
<td></td>
<td>(0.562)</td>
<td>(0.601)</td>
<td>(0.889)</td>
<td></td>
</tr>
<tr>
<td>0.30</td>
<td>0.727</td>
<td>0.794</td>
<td>0.942</td>
<td>16.323 (2.243)</td>
</tr>
<tr>
<td></td>
<td>(0.734)</td>
<td>(0.707)</td>
<td>(0.925)</td>
<td></td>
</tr>
<tr>
<td>0.60</td>
<td>0.829</td>
<td>0.874</td>
<td>0.955</td>
<td>16.265 (2.257)</td>
</tr>
<tr>
<td></td>
<td>(0.823)</td>
<td>(0.886)</td>
<td>(0.928)</td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>0.945</td>
<td>0.951</td>
<td>0.957</td>
<td>16.428 (2.408)</td>
</tr>
<tr>
<td></td>
<td>(0.940)</td>
<td>(0.935)</td>
<td>(0.942)</td>
<td></td>
</tr>
</tbody>
</table>

$^a$Pearson product-moment correlation coefficient (Spearman rank-order correlation coefficient)
### Dimensionality and CAT estimation

Table 3. Intercorrelations\(^a\) between  \( \theta \) and  \( \theta_1, \theta_2, \theta \) and the average number the of items administered (Mean N.A) for the confounded conditions.

<table>
<thead>
<tr>
<th>Item Pool Characteristics</th>
<th>( r_{\theta_1} )</th>
<th>( r_{\theta_2} )</th>
<th>( r_{\theta} )</th>
<th>Mean NIA (SD NIA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>0.628 (0.674)</td>
<td>0.623 (0.437)</td>
<td>0.897 (0.851)</td>
<td>16.433</td>
</tr>
<tr>
<td>0.30</td>
<td>0.692 (0.747)</td>
<td>0.768 (0.617)</td>
<td>0.905 (0.886)</td>
<td>16.689</td>
</tr>
<tr>
<td>0.60</td>
<td>0.765 (0.801)</td>
<td>0.850 (0.836)</td>
<td>0.905 (0.883)</td>
<td>16.703</td>
</tr>
<tr>
<td>0.90</td>
<td>0.907 (0.907)</td>
<td>0.921 (0.910)</td>
<td>0.922 (0.909)</td>
<td>17.394</td>
</tr>
<tr>
<td>0.60 0.03</td>
<td>0.663 (0.720)</td>
<td>0.620 (0.422)</td>
<td>0.920 (0.877)</td>
<td>16.337</td>
</tr>
<tr>
<td>0.30</td>
<td>0.760 (0.797)</td>
<td>0.750 (0.614)</td>
<td>0.926 (0.923)</td>
<td>16.293</td>
</tr>
<tr>
<td>0.60</td>
<td>0.834 (0.850)</td>
<td>0.856 (0.851)</td>
<td>0.948 (0.916)</td>
<td>16.276</td>
</tr>
<tr>
<td>0.90</td>
<td>0.928 (0.927)</td>
<td>0.942 (0.933)</td>
<td>0.943 (0.930)</td>
<td>16.730</td>
</tr>
<tr>
<td>0.90 0.03</td>
<td>0.704 (0.749)</td>
<td>0.608 (0.418)</td>
<td>0.941 (0.907)</td>
<td>15.917</td>
</tr>
<tr>
<td>0.30</td>
<td>0.787 (0.802)</td>
<td>0.756 (0.635)</td>
<td>0.956 (0.942)</td>
<td>16.066</td>
</tr>
<tr>
<td>0.60</td>
<td>0.840 (0.845)</td>
<td>0.874 (0.868)</td>
<td>0.961 (0.923)</td>
<td>16.031</td>
</tr>
<tr>
<td>0.90</td>
<td>0.944 (0.939)</td>
<td>0.955 (0.945)</td>
<td>0.958 (0.941)</td>
<td>16.230</td>
</tr>
</tbody>
</table>

\(^a\)Pearson product-moment correlation coefficient (Spearman rank-order correlation coefficient)
Figure Captions

**Figure 1.** RMSE analysis for the nonconfounded conditions for $r_{b_1b_2} = 0.03$, $r_{b_1b_2} = 0.60$, $r_{b_1b_2} = 0.90$ and $r_{q_1q_2} = 0.03$, $r_{q_1q_2} = 0.90$. 
RMSE
Ability r=0.03; Difficulty r=0.03; Nonconfounded

RMSE
Ability r=0.90; Difficulty r=0.03; Nonconfounded

RMSE
Ability r=0.03; Difficulty r=0.60; Nonconfounded

RMSE
Ability r=0.90; Difficulty r=0.60; Nonconfounded

RMSE
Ability r=0.03; Difficulty r=0.90; Nonconfounded

RMSE
Ability r=0.90; Difficulty r=0.90; Nonconfounded
Figure Captions

Figure 2. Bias analysis for the nonconfounded conditions for $rb_1b^2 = 0.03$, $rb_1b^2 = 0.60$, $rb_1b^2 = 0.90$ and $\theta_1\theta_2 = 0.03$, $\theta_1\theta_2 = 0.90$. 
Figure Captions

**Figure 3.** RMSE analysis for the confounded conditions for $r_{bb_2} = 0.03$, $r_{bb_2} = 0.60$, $r_{bb_2} = 0.90$ and $r_{\theta_1\theta_2} = 0.03$, $r_{\theta_1\theta_2} = 0.90$. 
Dimensionality and CAT estimation

Figure Captions

**Figure 4.** Bias analyses for the confounded conditions for \( r_{b1b2} = 0.05, r_{b1b2} = 0.60, \)
\( r_{b1b2} = 0.90 \) and \( r_{\theta1\theta2} = 0.03, r_{\theta1\theta2} = 0.90. \)
Bias
Ability r=0.03; Difficulty r=0.03; Confounded

Bias
Ability r=0.90; Difficulty r=0.03; Confounded

Bias
Ability r=0.03; Difficulty r=0.60; Confounded

Bias
Ability r=0.90; Difficulty r=0.60; Confounded

Bias
Ability r=0.03; Difficulty r=0.90; Confounded

Bias
Ability r=0.90; Difficulty r=0.90; Confounded