Running head: ITEM CALIBRATION IN COMPUTERIZED ADAPTIVE TESTING

Application of Direct Optimization for Item Calibration in Computerized Adaptive Testing

Iosif A. Krass

Defense Manpower Data Center

The author would like to thank Daniel Segall, Robert Holmes, and Bruce Bloxom for their supporting comments and helpful suggestions.

Requests for reprints should be sent to Iosif A. Krass, Defense Manpower Data Center, 400 Gigling Road, Seaside, CA 93955-6771. E-mail: krassia@pentagon.mil

Abstract

In the process of item calibration for a CAT test, many well-established calibrating methods can show significant increase in biases or variability in the estimation of item parameters. This paper introduces a calibration algorithm based on the convexity of likelihood functions, which can be successfully applied to data resulting from a CAT test. This package consists of: (a) an algorithm that estimates examinee ability, and (b) an algorithm that estimates the parameters for a new item that is seeded into the CAT test. The precision (in the sense of biases and variability) of the new package in estimation of item parameters is comparable with BilogMG, and in some cases exceeds it.

Key Words: computerized adaptive testing, CAT, item calibration, item parameters, maximization of likelihood, log-likelihood function, precision, BilogMG, DMAP, ICCs, multi-dimensional test, convexity.

Application of Direct Optimization for Item Calibration in Computerized Adaptive Testing

1. Introduction

The problem of item calibration--estimation item parameters when the model of responses is fixed--is very old and has been well discussed in the psychometric literature (e.g., Bock & Aitkin, 1981; Thissen & Steinberg, 1984; Samejima, 1969; Levine, 1984). There are a few packages available which are designed to do the job of calibration, for example, BilogMG (Zimowski et. al. 1996), Multilog, (Thissen, 1988), ASCAL (Assessment System Corporation, 1988), Logist (Wingersky et. al. 1982). However, nearly all available packages and algorithms are designed to use results of tests given in the paper-and-pencil mode.

Beginning from 1994, in Armed Services Vocational Aptitude Battery (ASVAB) computerized adaptive testing mode (CAT) was developed with seeded-item design scheme (Segall, Moreno, Bloxom, & Hetter, 1997) which allows relatively fast and cheap to get test results for calibrating of new items from unbiased examinee population (see more explanations below). Unfortunately our attempts to apply mentioned above calibrating packages to test results from CAT ASVAB seeded item algorithm lead to large biases and standard errors in item-parameter estimates which is probably due to specifics of CAT testing data and sometimes due to special features of CAT ASVAB testing data.

In CAT testing data, the matrix of examinee-by-item responses is rather sparse, in comparison with the paper-and-pencil test. Though the package like BilogMG can handle some amount of missing

data (which is making matrix of responses not fully compete comparing with case of classical paper-and-pencil tests) quantity of missing data in the case of CAT exams looks like excessive. The CAT tests are rather short (at most 15 items in the case of CAT ASVAB) because of computerized-adaptation to each examinee. Different items have different chance to be present to any examinee. For example, very hard or very easy items will be hardly ever be exposed, so the missing data for those items is much bigger than for other items. Also in the case of ASVAB the tested examinee population is sometimes considerably different from a standard normal population (i. e. population with ability distribution normal with mean zero and standard error one) due to season and geographical changing. If not those specific features, adjusted BilogMG (described below) is rather good, precise and practical method for calibrating of CAT seeded items. Therefore, the main reason which force us to develop another practical calibrating algorithm was failure of BilogMG (due to increase in biases or variability of parameter estimations) in the case of not standard examinee population. BilogMG also lost some robustness due to increase variability of estimated parameters in the case of non one-dimensional items (see Appendix 4). Because this kind of features can be found not only in CAT ASVAB but another CAT tests we think the paper can be interested for other researchers in CAT area.

Due to all that, we have developed an algorithm based on likelihood optimization, which is a version of Joint Likelihood Optimization and is not marginal; or by Baker classification (Baker, 1992) it belongs to the class of Direct Maximization Aposteriori algorithms (DMAP), which name will be used for the algorithm. In this paper we will describe the new algorithm and compare it with adjusted BilogMG, the most widely used parametric calibration package, which is marginal likelihood

optimization algorithm rather good described in the literature (for references see Zimowski et. al. 1996) and as we will demonstrate it is rather good calibrating tool under the "standard" conditions.

The DMAP algorithm begins by estimating examinee aposteriori distribution ability based on the test results (Krass, 1997), which considered to be "nuisance" parameters in EM language (here we are using terminology of McLachlan & Krishnan, 1997). DMAP is getting those estimate by likelihood maximization without any assumption on the estimated 3PL parameters of the seeded item which is analog of EM unobserved data. Getting estimation of examinee abilities by maximizing likelihood, DMAP estimates 3PL parameters of seeded item. Then it re-estimates examinee ability using estimated 3PL parameters of seeded item and continues this process up to convergence inside of the required tolerance bounds. This approach is similar to approach of splitting calibrating estimation problem into two sub-problems which is used in Logist algorithm (Wingersky et. al. 1982) and, more generally, described in the theory of EM algorithms. (McLachlan & Krishnan, 1997). In this paper we will describe estimating examinee ability by DMAP and then its estimating seeded item parameters, and we will present some simulation results to compare the performances of the algorithm and BilogMG.

DMAP algorithm as well as adjusted BilogMG (adjustment is presented below) consists part of the on-line calibrating algorithmic package for seeded item in CAT ASVAB which installed in 1998 and successfully implemented (Krass, 1998).

2. Estimation of Examinee Ability

Let in our test the item pool consists of I items, with Item Characteristic Curve (ICC) $P_i(q)$, $i = 1,\cdots, I$ being 3PL ICC, i.e.,

$$P_i(q) = c_i + \frac{1 - c_i}{1 + \exp(l_i(q))}, \quad (1)$$

Where $l_i(q) = -D \cdot a_i \cdot (q - b_i)$, and a_i, b_i, c_i; $i = 1,\cdots, I$ the item discriminating, difficulty, and guessing indexes, correspondingly; q is the latent ability of an examinee and $D = 1.7$ is a scaling constant (Lord, 1980). We assume that examinee ability $q \in [q_{\min}, q_{\max}]$, which means that the optimization, described below, should be done as a constrained optimization. In CAT-ASVAB we have assumption$q_{\min} = -3.0$ and $q_{\max} = +3.0$. This type of optimization feature cannot be done with an "internal" algorithm type such as Newton-Raphson, which design to find zeros of second derivative of maximized function. If the optimized function reach its maximum on the edge of the ability segment, it second derivative generally speaking is not equal to zero. Let our examinee get a sequence $\{i_1, i_2, \cdots, i_k\}$ of items generated by CAT, where $k \le K$, and K is the length of the CAT-ASVAB test (usually $10 \le K \le 15$). CAT ASVAB test algorithm is totally driven by an information table based on an item pool with a rather large exposure control factor (at least 0.7 in CAT3-CAT4 which means that next item selected by information table can be blocked for the examinee with 30% probability, forcing CAT algorithm to present for examinee another item) (Hetter & Sympson, 1997). Due to this properties item-response matrix for CAT test data is not only sparse but also somewhat chaotically filled, because

ideally every examinee should have in CAT exam his/her own "unique" sequence of items. However, design of DMAP algorithm is made to use the fact that the test data is result of a CAT exam.

CAT-ASVAB items are multiple-choice items, so the examinee produces a dichotomous answer sequence $\bar{u}_k = \{u_1, u_2, \cdots, u_K\}$. Then, his/her likelihood function, under independence of answers assumption after the first k items of the test is:

$$L(\bar{u}_k, q) = g(q) \cdot \prod_{i=1}^{k} P_i(q)^{u_i} \cdot Q_i(q)^{(1-u_i)} \tag{2}$$

where $Q_i(q) = 1 - P_i(q)$ and $g(q)$ is the density of prior ability distribution in the population of examinees. The value $\bar{q}_k$ which maximizes likelihood

$$L(\bar{u}_k, \bar{q}_k) = \max_{q \in [q_{\min}, q_{\max}]} L(\bar{u}_k, q) \tag{3}$$

is considered to be the best estimator of the examinee's ability after the first k items of the test. As usual, we assume that prior ability distribution is normal $N(m, s)$, i.e.,

$g(q) = \frac{1}{s\sqrt{2p}} \exp(-\frac{(q-m)^2}{2 \cdot s^2})$, where m and s are the mean and SD of prior distribution.

The first part of DMAP algorithm design to find value of maximizing abilities $\bar{q}_k, k = 1, \ldots, K$, which approximate solution of (3) and giving estimates of "nuisance" parameters – examinee abilities. Technical details and description of the algorithm is done into Appendix 1, but now we will stop on results of this part of DMAP algorithm application.

In the current CAT-ASVAB, the Owen-Bayesian algorithm (Owen, 1975) is applied to estimate ability of the examinee "on-the-fly," and the Bayesian-Modal (Segall, et al., 1997) algorithm is

applied to the total test sequence to make the final tuning in ability examinee estimation. The above-described DMAP algorithm requires a little bit more computer time (about 1.5 more), in estimation of particular examinee ability, but it gives more precision in the estimation in the densest part of the ability distribution. However, the effect of computer time increasing can be felt only in the scale of big simulation with several thousands simulees. In the case of one examinee we still in the area of parts of seconds[1] with modern PC (described in the Appendix 1, which now is standard for CAT ASVAB).

The results of a simulation for 3,000 examinees for Arithmetic Reasoning in CAT-ASVAB Form 1, where the size of the item pool is equal to $I = 94$, is shown in Figures 1 and 2. In this simulation experiment, we took 3,000 examinees with standard normal ability distribution and "recovered" their known "true" ability by standard Bayesian (i. e. Owen-Bayesian plus Bayesian-Modal algorithms) methods (Figure 1) and by the DMAP algorithm (Figure 2). To get those graphs we took grid of equal distant abilities $q_i = 0.1 \cdot i - 2.5;\ i = 0, \ldots, 49$ beginning from –2.5 to +2.5. For every interval $[q_i, q_{i+1})$ we estimated Maximum deviation between "real" ability of a simulee q and its estimated ability $\hat{q}$ (so called theta-hat) "recovered" ability by correspondent to the graph method. This Maximum Deviation is denoted as Max. Dev. on graphs. In the same mode we compute Minimum Deviation (Min. Dev. on graphs), Standard Deviation and Deviation Mean (Std Error and Mean on graphs correspondingly).

(Figures 1 and 2 about here.)

[1] Ability estimating part of DMAP can be used as "on-fly" estimator of examinee ability in CAT test.

As we can see, DMAP has about the same precision (in the sense of SD or maximum–minimum deviation) as a standard Bayesian algorithm for $q \le -1.85$ but does better than the standard from $q \ge -1.05$. In the area of ability $q < -2.00$, where guessing is a decisive factor for examinees, DMAP typically loses to the standard Bayesian methods, but there is not a large population in that ability area. We make the same kind of simulation experiments for other available CAT ASVAB tests (CAT1 through CAT4) results of those simulations was quite close to the test AR CAT1 described above, so we did not present those results here to save space. These results are shown to us that application of DMAP for re-estimation of examinees ability increase precision of estimation, because, generally speaking, we can use estimation of examinees ability which is done by standard Owen-Bayesian method implemented by CAT ASVAB to proceed with DMAP algorithm. The increasing of precision of examinee ability is essential in the second part of DMAP, which is estimation of 3PL parameters of seeded item.

The presented part of DMAP algorithm is heavily using adpativeness of individual examinee item sequence in CAT testing data. This property of item sequence is used to define from what side of ability interval dichotomy process should began (see Appendix 1) corresponding to the current item and response in CAT adaptive sequence. It helps to contract area of search of maximizing likelihood (3) ability $\overline{q}_k$ estimation, increasing speed of algorithm convergence.

3. Estimation of ICC parameters

In this section we will demonstrate the implementation of the second part of DMAP algorithm which design to get 3PL ICC parameters on unknown (seeded) items, assuming that the ability of participating examinees has already been estimated. Let set $P(a_i, b_i, c_i)(q)$; $i = 1, \cdots, I$ of 3PL functions (1) present CAT pool (so called set of adaptive items), where; a_i, b_i, c_i are parameters of the item $i = 1, \cdots, I$. There is a new $(I+1)$-st item with unknown parameters which is called a CAT seeded item; it is usually given to an examinee in the second, third, or fourth (random) position of his or her exam. If the CAT test of the length K is given to M examinees with abilities q_m and $g(q_m)$ his/her estimation of prior distribution, $m = 1, \ldots, M$, then the joint likelihood of the response vectors can be written as

$$L = \prod_{m=1}^{M} g(q_m) \cdot \prod_{j=1}^{K+1} (P(a_{i_j^m}, b_{i_j^m}, c_{i_j^m})(q_m))^{u(i_j^m)} \cdot (Q(a_{i_j^m}, b_{i_j^m}, c_{i_j^m})(q_m))^{(1-u(i_j^m))}, \quad (4)$$

Here i_j^m is index of item which is given to examinee number $m = 1, \ldots, M$ on $j = 1, \ldots, K+1$ step of CAT test, and $u(i_j^m)$ is the binary response of the examinee m on the test item i_j^m which he/she got in the test in the j selection of CAT algorithm. In expression (4) we took into account that the length of the test is increased to $(K+1)$ due to administration of the seeded item. Thus if the seeded item is given to examinee m in the step $j_m = 2,3,4$, then $i_{j_m}^m = I+1$, because by our agreement seeded item is $(I+1)$-th item of the CAT pool. As we told before, the seeded item is given to every examinee participating in the test, due to this relation (4) can be rewritten in the form:

$$L = L_0 \cdot \prod (P(\bar{a}, \bar{b}, \bar{c})(q_m))^{u_m} \cdot (Q(\bar{a}, \bar{b}, \bar{c})(q_m))^{(1-u_m)}, \quad (5)$$

where $(\bar{a}, \bar{b}, \bar{c}) = (a_{I+1}, b_{I+1}, c_{I+1}) = (a_{i_{j_m}^m}, b_{i_{j_m}^m}, c_{i_{j_m}^m})$ are the item parameters of the seeded item, and $u_m = u(i_{j_m}^m)$ is the response of m-th examinee on the seeded item in the test. Here L_0 is the joint likelihood of the test without the seeded item. Let us note that expression (4) implied that presentation of items in CAT test are independent, which is arguable (see Levine & Williams, 1988). But due to random mechanism of administration of seeded item in CAT test independence of answer on seeded item with respect to answer on other CAT items is more plausible.

From the point of view of item calibration the seeded item is presented to the person who is trying to get through the exam as a common examinee. Application of a seeded item does not influence examinee ability (Krass, 1998), so we get test from the "operating" level of exams, which is different from calibrating data in P&P modes. In this mode examinees very often know or suspect that the whole exam or part of it with seeded item in is not influence the final exam results. Thus test data using in this item is not, generally speaking, from examinee population with "original" aptitude. This we meant when we mention in introduction about unbiasness of test data produced by seeded item design.

To estimate item parameters for the seeded item, we must solve the problem of maximization of log-likelihood of (5), i.e., find a solution to the problem:

$$\ln L = \ln L_0 + \sum (u_m \cdot \ln P(\bar{a}, \bar{b}, \bar{c})(q_m) + (1-u_m) \cdot \ln(1-P(\bar{a}, \bar{b}, \bar{c})(q_m)) \Rightarrow \max, \quad (6)$$

where $(\overline{a}, \overline{b}, \overline{c}) \in [a_{\min}, a_{\max}] * [b_{\min}, b_{\max}] * [c_{\min}, c_{\max}]$. Upper and low boundaries such as $a_{\min}, a_{\max}, \ldots$ for different parameters are user-defined for a test, as the boundaries in the case of ability estimation.

In the practical application of DMAP algorithm for On-line calibration parameters search area $[a_{\min}, a_{\max}] * [b_{\min}, b_{\max}] * [c_{\min}, c_{\max}]$ defines through the checking parameters of all item pools in implemented CAT ASVAB-s for the particular test to get upper and lower limit for every item parameter. Typical example of the border: $a_{\min} = 0.1$, $a_{\max} = 4.0$, $b_{\min} = -4.0$, $b_{\max} = 4.0$, $c_{\min} = 0.0$, $c_{\max} = 0.5$. Technical details of design and implementation of this part of DMAP algorithm are presented in the Appendix 2.

As we mentioned in Introduction the both parts of described DMAP algorithm united like in usual EM type algorithm through successive iterative process. Estimation of examinee abilities, then estimation of parameter of seeded item. Re-estimation of examinee abilities, using newly estimated seeded item as addition to CAT test data, and so on up to reaching needed bound of tolerance.

4. Comparing precision DMAP and BilogMG

Comparing the performance of the DMAP algorithm with the BilogMG algorithm is done through a set of simulations arrange to recover values of known parameters of seeded items. But first the BilogMG package must be adjusted to get a reasonable performance. As we have explained, the matrix of responses for a CAT test is rather sparse. Furthermore, items with low information are used very rarely, and items with high information are used too often this together with randomness providing

by CAT exposure control mechanism resulting in very non-uniform filling of the response matrix. Due to that application of BilogMG very often leads to a non-convergence runs which finish by exhausting number of permitted iterations cycle or reasonable limit for computer time. In this case BilogMG provides "recovered" parameters too far from reality. To avoid this inconsistency, we run BilogMG in two stages. First, we simulate a paper-and-pencil test for our set of M examinees on items that belong to the CAT-ASVAB item pool. In this part of simulation we are using ability estimation done by CAT ASVAB and simulate response on a particular item with help of 3PL-model (1). Then we run BilogMG and save the result of item pool estimation with help of the BilogMG "SAVE" statement. After that, we run BilogMG for the data obtained from the simulated CAT-ASVAB test, with the seeded item included, using the preliminary estimation through the "IFNAME" subcommand in the "GLOBAL" statement in BilogMG. With this approach, BilogMG always converges and provides a rather reasonable and stable estimation for the population of examinees with normal distributed abilities and with one-dimensional item, where item dimensionality was estimated by preliminary factor analysis for the correspondent test using "TESTFACT" package (Wilson, Wood, Gibbons, 1991). After much experimentation, we are decided to use 30 quadrature points in the marginal estimations for BilogMG. The first stage in BilogMG running process provides estimation of item parameters of adaptive item set as well as of ability distribution of participating examinees. Moreover, other attempt to input in BilogMG information of item parameters of adaptive items (CAT item pool), for example, using prior mechanism, does not improve it performance[2].

[2] Performance of BilogMG in the "normal" cases was so good that we keep it in our calibrating package.

To compare performances of algorithms in the "normal" situation, i.e. when simulees ability satisfy standard normal distribution and all tested items is one dimensional (see the remark above), we use three typically representative items from the item pool for AR: an "easy" item $(a,b,c) = (1.17,-1.63,0.13)$, a "normal" or "average" item $(a,b,c) = (1.3,0.12,0.15)$, and a "hard" item $(a,b,c) = (1.23,1.63,0.07)$. (The above classification is done based on item difficulties). All three items are rather informative in their areas of difficulty. Then, for each item we run the CAT-ASVAB test simulation twenty times for the set of M examinees, changing random seeds each time to generate different response matrixes. In every run we use DMAP and adjusted BilogMG to re-estimate item parameters for the above-described items. We found that both packages in those experiments show no biases in parameter estimations, the major differences are in the size of variability of those estimations.

First of all, we run our simulation for a different number of examinees with standard normal distribution of their abilities, changing examinee number as: $M \in \{300, 500, 750, 1000, 1500, 2000\}$. In this experiment, we try to identify the number of examinees needed to provide estimation of parameters with satisfactory variability. In Table 1 we show estimation of SD for three parameters in our experiment.

(Table 1 about here.)

These results are graphically shown in the Figure 3.

(Figure 3 about here.)

Let us note that even seeded item design for collecting data for calibration process is relatively cheap, it requires some time to get enough data. For this reason number of examinees (sample size) more than 2000 per one calibrating item looks like unreasonable big. Usually we have to calibrate few

hundred items per test for next generation of CAT ASVAB, and it take at least three - four month to collect 1500 answers per item. In simulation studies we goes beyond sample size 2000, and found that variability of estimation of item parameter do not significantly change comparing with sizes 1500, 2000. As we can see, DMAP requires at least 750 examinees per test to get variances in a and b parameters comparable with BilogMG, and BilogMG is always better in the estimating of parameter c. However, the last advantage (more precise estimation of parameter c) disappears if we measure weighted average distances between "true" ICCs of studied items and ICCs built with estimated 3-PL parameters. Here, under weighted distance between two ICCs curves, we mean

$$D = \sqrt{\sum_{j=1}^{T} w_j \cdot (P(a,b,c)(q_j) - P(\tilde{a},\tilde{b},\tilde{c})(q_j))^2} \ ,$$

where $(\tilde{a},\tilde{b},\tilde{c})$ is the estimation of "true" parameters (a,b,c) by some package in a particular simulation experiment; q_j, $j = 1,\ldots,50$ are equidistant points in ability domain $[-3.0, 3.0]$, and weights are normally distributed, i. e. $W \cdot w_j \in N(0,1)$; $\sum_{j=1}^{T} w_j = 1$, where W is a scaling coefficient. In Table 2 and Figure 4 we show that, from the point of view of distances between ICC curves, both algorithms perform more or less equally.

(Table 2 and Figure 4 about here).

This is because the influence of guessing parameter is strong where the density of the examinee population is small. From this simulation experiment, we see that the performance of both packages is about the same for $M = 1{,}500$, and variability of item parameters is minimal and stabilizing. Based on

that, we will assume that calibration of a seeded item requires at least 1500 examinee answers per item in "real" on-line calibration with CAT-ASVAB.

In the case of the CAT-ASVAB, very often we have violation of normality in examinee ability distribution due to seasonal and geographical location differences. To simulate this situation we consider two types of artificial populations. In the first type, we mix 750 examinees with normal-normal ability distribution with 750 examinees with ability distribution $N(-0.8, 1.0)$. After mixing, we get not-normal ability distributed population of examinees with mean of ability equal –0.4 and SD equal ≈ 1.15. We call this population "less able" (to the test). In the same mode, we make an "more able" population with mean +0.4 and the same SE≈ 1.15. In both cases, we apply previously described simulation for the same three items of CAT-ASVAB Form 1 AR. We find the variances of estimation of 3-PL parameters are about the same as for the normal case (described above); the main differences are in biases of parameter estimations. Those biases are shown in Figure 5.

(Figure 5 about here.)

As we can see, BilogMG begins to be significantly biased in estimation of difficulty parameters, overestimates them for the "less able" population, and underestimates for the "more able" population. As a result, the average weighted distance between estimated ICCs and "true" ICCs significantly increases for BilogMG (Figure 6). On the other hand, the bias increases for DMAP are not significant with respect to the normal case. Let us note that in real life we can only assume of existence of "more" or "less" able group of examinees in the particular site of test taking. Due to that we can not apply a different test group design provided by BilogMG.

(Figure 6 about here.)

The difference in precision (biases) of estimation of b parameters between DMAP and BilogMG can be explained by the heavy dependence in BilogMG architecture on assumption of normality of prior ability distribution. DMAP being a version of joint likelihood maximization method is not depend on the examinee ability distribution. We found another case of differences in precision of parameter estimations – case of not one-dimensional test. In this case variability of estimation of a and b by DMAP were considerably lesser than variability provided by BilogMG estimation. However, both algorithms do not have a mechanism to compensate not one dimensionality effect, and improving by DMAP estimation may be connected with not much reliance on assumption of type of prior ability distribution. Result of experiments for multidimensional is given in Appendix 4.

5. Conclusion

We have demonstrated that the above described DMAP algorithm has about the same precision as the BilogMG algorithm in calibrating items from the CAT-ASVAB seeded design. More than that, in "special" circumstances, such as the absence of normality in prior distribution of examinee ability or the multi-dimensionality in item content, BilogMG loses its precision, but DMAP does not. This is because BilogMG is a marginal algorithm, with normality, to some extent, built in by the application of computation joint distribution through quadrature points. The other "weak" part of BilogMG is the application of only the Newton-Raphson algorithm as the main engine for local sub-optimization. As we have already mentioned, this tool will not pursue constrained optimization. However, from the point of view of maximization of joint likelihood, BilogMG and DMAP use different types of heuristics, so their solutions in different initial circumstances can be better or worse, depending on many "internal" conditions.

Appendix 1

DMAP Algorithm Ability Estimation.

Typically, we begin from standard normal prior $N(0,1)$ and then tune m and s to get faster convergence for this particular examinee (of course, we are not trying to change the scale of whole population at that moment). If we begin from $N(0,1)$, to get the maximizing $\bar{q}_k$ we consider log-likelihood which has its derivative due to (1) as:

$$\frac{d}{dq}\log(L(\bar{u}_k,q)) = -q + \sum_{i=1}^{k}\left(\frac{u_i}{P_i(q)} - \frac{1-u_i}{1-P_i(q)}\right)\cdot\frac{d}{dq}P_i(q) = -q + \sum_{i=1}^{k} R_i(q)$$

where

$$R_i(q) = \begin{cases} \dfrac{(1-c_i)\cdot\exp(l_i(q))\cdot D\cdot a_i}{(1+\exp(l_i(q)))\cdot(1+c_i\cdot\exp(l_i(q)))}; & for\ u_i = 1 \\ \dfrac{-D\cdot a_i}{(1+\exp(l_i(q)))}; & for\ u_i = 0 \end{cases} \qquad (7)$$

To find a zero of log-likelihood derivative, in the case when the log-likelihood maximum is reached inside of domain segment $[q_{\min}, q_{\max}]$, we must solve the "fixed-point" problem for function $\sum_{i=1}^{k} R_i(q)$, i.e., find a solution of the equation:

$$q = \sum_{i=1}^{k} R_i(q). \qquad (8)$$

Solution of this type of equation is heavily studied in computational mathematics literature (Blum, 1972; Ramsay, 1975), but the fastest solution can be reached in the case of monotone functions $R_i(q)$ which

we have here, at least in the area of solution of (8). From (7) it follows that, in the case of $u_i = 1$, we have $R_i(q) > 0$, and $R_i(q) \to 0$ if $q \to \pm\infty$. The function $R_i(q)$ is uni-modal and reaching maximum at the point $\bar{q}_i = b_i - \frac{\ln(\sqrt{3}c_i)}{D \cdot a_i}$; i. e. $\bar{q}_i$ is big enough, if c_i is small enough. In the other words, if c_i is small enough the function $R_i(q)$ is monotone increasing in the area of solution. On the other hand, in the case of $u_i = 0$, the function $R_i(q) < 0$, and $R_i(q) \to 0$ if $q \to -\infty$, and $R_i(q) \to -D \cdot a_i$ if $q \to +\infty$. The function $R_i(q)$ is monotone increasing in this case with point of inflection on $\bar{q}_i = b_i + \frac{\ln 3}{D \cdot a_i}$. Out of this follows, that $q > \sum_{j=1}^{k} R_{i_j}(q)$ for $q = q_{\max}$, and $q < \sum_{j=1}^{k} R_{i_j}(q)$ for $q = q_{\min}$, if $q_{\max}$ is large enough and $q_{\min}$ is small enough and for $k = 1$ (i. e. at the beginning of the CAT exam) equation (8) has unique solution. Therefore, depending on whether the answer is right or wrong, the first solution of (8), which defines the DMAP estimation of examinee ability after the first item administered by CAT, can be found by dichotomy, or bisection, from the "right side" if the answer is correct, or "left side" in the opposite case. Under right side, we mean beginning the process of checking if the inequality

$$q_{\max} \le \sum_{i=1}^{k} R_i(q_{\max}) \qquad (9)$$

holds. From (9) it follows that in the case, when (9) holds, the derivative of $\frac{d \log(L(u_k, q))}{dq}$ is negative in all our domains, so the maximizing latent ability $\bar{q}_1 = q_{\min}$; in this case the process can be continued to

the next item. If the above inequality is not true, we check the left side condition $q_{\min} \geq \sum_{i=1}^{k} R_i(q_{\min})$ to see if maximization is reached on the right border of the domain. After checking borders we are sure that at least one solution of (8) is inside the segment $[q_{\min}, q_{\max}]$, and it can be found by the following dichotomy process: Let $\tilde{q}_{\min} = q_{\min}$ and $\tilde{q}_{\max} = q_{\max}$ define $\tilde{q} = \tilde{q}_{\min} + 0.5 \cdot (\tilde{q}_{\max} - \tilde{q}_{\min})$. If $\tilde{q} < \sum_{j=1}^{k} R_{i_j}(\tilde{q})$, then $\tilde{q}_{\min} = \tilde{q}$ and $\tilde{q}_{\max} = \tilde{q}$ in the case of opposite inequality. The process continues until $\tilde{q}_{\max} - \tilde{q}_{\min} > d$, where d is a given precision of computation. The algorithm convergence rate is $\frac{1}{2^n}$, where is n is the number of iterations, i. e. size of the area where solution of (8) is located contracted as $\frac{1}{2^n}$ when process is continue, which provides rather big speed of convergence. On our standard IBM compatible Pentium PC with speed 166 MHZ it usually take lesser than a second to get a solution of (8).

As it is shown by Samejima (1973), the log-likelihood function (2) is not, generally speaking, unimodal so (8) can have more than one solution, but the second solution is usually out of the border of the "normal" domain. Our algorithm is designed to hunt for more than one solution of (8) checking the sign of difference between left and right sides of (8) on rather tight net of ability values. However, after more than 1,000,000 applications of the algorithm to the simulated or real life test situation, we were not able to find a second solution of (8) in the considered domain [-3.0, +3.0].

From the properties of (7) it follows, independently of the first answer, if the answer on the second item is correct, the root of the equation (8) will be moved to the right, and it can be found by

dichotomy beginning from the right side. If the answer on the second item in the sequence is wrong, the root of (8) will be moved to the left, and it can be found by dichotomy from the left side. This phenomenon is due to the property $R_i(q) > 0$ in the case of a correct answer, and $R_i(q) < 0$ in the case of a wrong answer. This phenomenon reduces the domain of searching of maximizing likelihood ability while the test is developing adaptively. Let us note that this phenomena corresponds to the application of CAT exam to an examinee.

In Figure 9, we present the case of a test where the first item is answered correctly and the second wrongly. The darker curve corresponds to the function $R_1(q)$ for the first correct answer, and the lighter curve corresponds to the summation $R_1(q) + R_2(q)$ for the first two items when the first was answered correctly and the second wrongly. The intersection of the straight line and the graph of the function $R_1(q) + R_2(q)$ gives the DMAP estimation of theta for the test length of two.

(Figure 9 about here.)

Appendix 2

DMAP Algorithm Parameter Estimation.

We are interested in constrained maximization on the given parallelepiped-domain: $(\bar{a},\bar{b},\bar{c}) \in [a_{\min}, a_{\max}] * [b_{\min}, b_{\max}] * [c_{\min}, c_{\max}]$. The DMAP algorithm described below will check the border of this domain parallelepiped before going to the internal point. But if we assume the maximizing solution in (6) is reached on an inside point of the domain, we must find a solution of equalities:

$$\frac{\partial \ln L}{\partial a}(\hat{a},\hat{b},\hat{c}) = \frac{\partial \ln L}{\partial b}(\hat{a},\hat{b},\hat{c}) = \frac{\partial \ln L}{\partial c}(\hat{a},\hat{b},\hat{c}) = 0. \qquad (10)$$

Then, from (10), we will have:

$$\frac{\partial \ln L}{\partial c} = \sum \left(\frac{u_m}{P(\bar{a},\bar{b},\bar{c})(q_m)} - \frac{1-u_m}{1-P(\bar{a},\bar{b},\bar{c})(q_m)}\right) \cdot \frac{\partial P}{\partial c}(\bar{a},\bar{b},\bar{c})(q_m)$$

However, from definition (1), the function $\frac{\partial P}{\partial c}(a,b,c)(q)$ does not depend on c. Using this fact, we have:

$$\frac{\partial^2 \ln L}{\partial^2 c} = \sum_{m=1}^{M} -\left(\frac{u_m}{P^2(\bar{a},\bar{b},\bar{c})(q_m)} + \frac{1-u_m}{(1-P(\bar{a},\bar{b},\bar{c})(q_m))^2}\right) \cdot \left(\frac{\partial P}{\partial c}(\bar{a},\bar{b},\bar{c})(q_m)\right)^2 < 0.$$

From this we can state that for fixed parameters $(\bar{a},\bar{b})$, the function $\ln L(\bar{a},\bar{b},\bar{c})$ is convex on c, and therefore the function $\frac{\partial \ln L}{\partial c}(\bar{a},\bar{b},\bar{c})$ is monotone, decreasing on c. As in the case of estimation of ability, if $\ln L(\bar{a},\bar{b},\bar{c})$ is not reaching maximum on the border of the segment $[c_{\min}, c_{\max}]$, its maximum

is reached in the root of the function $\frac{\partial \ln L}{\partial c}(\bar{a},\bar{b},\bar{c})$ which can be found by a dichotomy process.

Below, we describe in more detail how this work could be done in that case.

Let's introduce a function $F_m = 1+\exp(d\cdot a\cdot(q_m - b)); \quad m = 1,\cdots, M$, then

$\frac{\partial P(\bar{a},\bar{b},\bar{c})}{\partial c}(q_m) = \frac{1}{F_m}$, and $P(\bar{a},\bar{b},\bar{c})(q_m) = 1+\frac{\bar{c}-1}{F_m}$. After some algebra we will have:

$$\frac{\partial \ln L}{\partial c} = \sum_{m=1}^{M}\left(\frac{u_m}{F_m+\bar{c}-1} - \frac{1-u_m}{1-\bar{c}}\right) = \sum_{m=1}^{M}\frac{u_m}{F_m+\bar{c}-1} - \frac{N}{1-\bar{c}} \qquad (11)$$

where N is the total number of wrong answers on the seeded item in the test. If $N = 0$, i.e., there are no wrong answers, $u_m = 1,\ m = 1,\cdots, M$ for $c = 1$ (case of "perfect guessing"), we will have $\frac{\partial \ln L}{\partial c} = \sum_{m=1}^{M}\frac{1}{F_m} > 0$, which, due to monotone decreasing nature of function $\frac{\partial \ln L}{\partial c}$, means that $\frac{\partial \ln L}{\partial c} > 0$ for all c, and so the log-likelihood function $\ln L$ is monotone, increasing function and reaching maximum on the right end $\bar{\bar{c}} = 1$. If $N > 0$ so there is examinee m_0 such that $u_{m_0} = 0$, then $\frac{\partial \ln L}{\partial c} \to -\infty$ when $c \to 1$ and behavior of the function $\ln L$ depends on the behavior $\frac{\partial \ln L}{\partial c}$ on the left end $c = 0$. If $c = 0$; then

$$\frac{\partial \ln L}{\partial c} = \sum_{m=1}^{M} u_m \cdot \frac{1}{F_m - 1} - N = \sum_{m=1}^{M} u_m \cdot \frac{F_m}{F_m - 1} - M = \sum_{m=1}^{M}\frac{u_m}{P_m} - M \qquad (12)$$

where $P_m = P(\bar{a},\bar{b},\bar{c})(q_m)$. From (12) it follows, if $\sum_{m=1}^{M}\frac{u_m}{P_m} - M < 0$, then the likelihood function is monotone, decreasing and reaching maximum on the left end $\bar{\bar{c}} = 0$. If $\sum_{m=1}^{M}\frac{u_m}{P_m} - M \geq 0$, we will have one root for function $\frac{\partial \ln L}{\partial c}(\bar{a},\bar{b},\bar{c})$ which can be found by dichotomy. This root $\bar{\bar{c}} = c(a,b)$ will provide the searched likelihood maximum for fixed parameters a, b. Utilizing this, we implement a search through the dense net of points $(\bar{a}_j,\bar{b}_j)$, $j = 1,\cdots,N$, where $(\bar{a}_j,\bar{b}_j) \in A\times B$, computing the likelihood $L(\bar{a}_j,\bar{b}_j,c(\bar{a}_j,\bar{b}_j))$ and getting approximate maximization, for which the precision depends on the density of the net. This search ensures that we did not miss "essential" for maximization regions of parameter space. The scope of search can be considerably decreased if we use a convexity of the function $L(\bar{a},\bar{b},c(\bar{a},\bar{b}))$ on $\bar{b}$ for fixed $\bar{a} \in [a_{\min}, a_{\max}]$ (provided in the Appendix 3) under some approximation. Again, after more than 1,000,000 experiments, we can state that this approximation is holding in our case, i.e., the function $L(\bar{a},\bar{b},c(\bar{a},\bar{b}))$ is convex on $\bar{b}$.

Appendix 3

Convexity by Other Parameters

As we show, for fixed (a,b) the log-likelihood function $\ln L(a,b,c)$ is convex on c and reaches its maximum inside the prescribed segment $[c_{\min}, c_{\max}]$ or on its border. We now consider the case when the function $\ln L(a,b,c)$ reaches its maximum on c inside the above domain-segment. In this case there is a function $c = c(a,b)$ such that

$$\frac{\partial \ln L(a,b,c(a,b))}{\partial c} \equiv 0 . \qquad (13)$$

Because all considered functions are analytical under some regularity conditions (Kantorovich, 1968), the function $c = c(a,b)$ is also analytical, so it has all the derivatives. Let us present our 3PL function in the form:

$$P(a,b,c)(q) = c + (1-c)\cdot P_0(a,b)(q), \qquad (14)$$

where $P_0(a,b)(J) = \frac{\exp(\bar{l}(a,b,q))}{1+\exp(\bar{l}(a,b,q))}$, i.e., $P_0(a,b)(J)$ is a 2PL ICC in the considered case (Here $\bar{l}(a,b,q) = D\cdot a\cdot(q-b)$). Using (14) we can rewrite identity (13) in the form:

$$\frac{\partial \ln L(a,b,c(a,b))}{\partial c} \equiv \sum_{m=1}^{M} \left(\frac{u_m^*}{P(a,b,c(a,b))(q_m)} - \frac{1-u_m^*}{1-P(a,b,c(a,b))(q_m)}\right)(1-P_0(a,b)(q_m)) \equiv 0 . \qquad (15)$$

Then for the derivative of $\ln L(a,b,c(a,b))$ with respect to b we have:

$$\frac{\partial \ln L(a,b,c(a,b))}{\partial b} = \sum_{m=1}^{M} \left(\frac{u_m^*}{P(a,b,c(a,b))(q_m)} - \frac{1-u_m^*}{1-P(a,b,c(a,b))(q_m)} \right) \cdot \frac{\partial P(a,b,c(a,b))(q_m)}{\partial b} \quad \text{and}$$

$$\frac{\partial^2 \ln L(a,b,c(a,b))}{\partial b^2} = \sum_{m=1}^{M} -\left(\frac{u_m^*}{(P(a,b,c(a,b))(q_m))^2} + \frac{1-u_m^*}{(1-P(a,b,c(a,b))(q_m))^2} \right) \cdot \left(\frac{\partial P(a,b,c(a,b))(q_m)}{\partial b} \right)^2$$

$$+ \sum_{m=1}^{M} \left(\frac{u_m^*}{P(a,b,c(a,b))(q_m)} - \frac{1-u_m^*}{1-P(a,b,c(a,b))(q_m)} \right) \cdot \frac{\partial^2 P(a,b,c(a,b))(q_m)}{\partial b^2}$$

The first sum in this expression has a negative value. To work with the second sum, let us consider the expression for the second derivative $\frac{\partial^2 P(a,b,c(a,b))(q)}{\partial b^2}$. Taking a derivative of (14) we have:

$$\frac{\partial P(a,b,c(a,b))(q)}{\partial b} = \frac{\partial c(a,b)}{\partial b} \cdot (1 - P_0(a,b)(q)) - (1 - c(a,b)) \cdot D \cdot a \cdot P_0(a,b)(q) \cdot (1 - P_0(a,b)(q))$$

.

From this expression we get:

$$\frac{\partial^2 P(a,b,c(a,b))(q)}{\partial b^2} = (1 - P_0(a,b)(q)) \cdot \frac{\partial^2 c(a,b)}{\partial b^2} + 2 \cdot D \cdot a \cdot P_0(a,b)(q)(1 - P_0(a,b)(q)) \cdot \frac{\partial c(a,b)}{\partial b} +$$

$$+ 2 \cdot (1 - c(a,b)) \cdot (D \cdot a)^2 \cdot P_0(a,b)(q)(1 - 2P_0(a,b)(q)). \qquad (16)$$

Our approximation assumption is that $((1 - P_0(a,b)(q)) >> P_0(a,b)(q) \cdot (1 - P_0(a,b)(q))$ which means that probabilities product is considerably less that probability by itself. This assumption looks like a little bit too heavy but it works in practical computations just fine. In our opinion this assumption is rather close to the one of neglecting terms with higher powers in a Tailor presentation of a function comparing with the term of first and may be zero power. Applying this assumption to (16) will lead to

$$\frac{\partial^2 P(a,b,c(a,b))(q)}{\partial b^2} \approx \frac{\partial^2 c(a,b)}{\partial b^2} \cdot (1 - P_0(a,b)(q))$$

which, together with identity (15), will get us to the conclusion that under this approximation

$\frac{\partial^2 L(a,b,c(a,b))}{\partial b^2}$ is negative, so the function $L(a,b,c(a,b))$ is convex on b for fixed a. The same type of consideration can be given about convexity of $L(a,b,c(a,b))$ with respect to a for fixed b under the same approximation assumption.

Appendix 4

Case of Not-One-Dimensional Test

In the CAT-ASVAB test is one essentially not one-dimensional by its content , General Science, which consists of three subtests: Physical Science, Biological Science, and Chemical Science. Statistically non one-dimensionality of GS test was shown also by factor-analysis results (Zimowski & Bock, 1987). To simulate the application of this test, we assume that every simulee has three abilities for every subtest, which are normal-normal distributed but highly correlated with a coefficient of correlation equal 0.8. Thus, the matrix of correlation for General Science abilities in this population looks like

$R = \begin{pmatrix} 1.0 & 0.8 & 0.8 \\ 0.8 & 1.0 & 0.8 \\ 0.8 & 0.8 & 1.0 \end{pmatrix}$. We would like to get a three-dimensional ability vector $\tilde{q} = (\tilde{q}_1, \tilde{q}_2, \tilde{q}_3)$ such that every component of it will have a normal distribution with mean 0, and the correlation matrix between components will be equal R. To do this, we make a Cholesky decomposition of R, i.e., present it in the form $R = A^T * A$ where A^T matrix transposes to matrix A, the square root of R and $A = Q * diag(\sqrt{l_i})$ where Q is a three-dimensional orthogonal matrix. In our case $l_1 = l_2 = 0.2$ and $l_3 = 2.6$, and $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}$. Then, if vector $\bar{q} = (q_1, q_2, q_3)$ consists of three independent identically distributed components belonging to $N(0,1)$, vector $\tilde{q} = \bar{q} * A^T = (\tilde{q}_1, \tilde{q}_2, \tilde{q}_3)$ will have the desired multi-dimensional distribution (Bickel & Doksum, 1977). Thus, if a simulee gets a

Physical Science item, we use $\tilde{q}_1$ ability to get the response for that item; if item is Biological, we use $\tilde{q}_2$; and if it is Chemical, we use $\tilde{q}_3$.

In this three-dimensional situation, we choose for simulation three representative items for each science: one "easy" item $b < -1.4$, one "normal" item $-0.3 < b < -0.3$, and one "hard" item $b > 1.7$ (altogether we choose nine items for the General Science test). This design was made with purpose to eliminate effect of "difficulty" on variability of parameter estimations. As before, we run the simulation twenty times, changing random seeds and using 1,500 simulees in every run. Our results show that both packages are not significantly biased in parameter estimation, but there are increases in variance estimation, compared with a one-dimensional test. These increases are shown in the Figure 7.

(Figure 7 about here.)

As we can see, the largest and most significant increase is in the variances of estimating difficulty parameters by BilogMG. Further, with BilogMG, we have a significant increase in weighted distance between the estimated and "true" ICCs, especially for "normal" items (Figure 8). On the other hand, the increase in the variances of estimating difficulty parameters by DMAP is not significant relative to the normal case.

(Figure 8 about here.)

References

Assessment System Corporation (1988). User's Manual for the MicroCAT testing system, version 3. St. Paul, MN.

Baker, F. D. (1992). Item response theory. New York: Marcel Dekker Inc.

Bickel, P. J., & Doksum, K. A. (1977). Mathematical Statistics. San Francisco, CA: Holden-Day, Inc.

Blum, E. K. (1972). Numerical analysis and computation: Theory and Practice. Reading, MA: Addison Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, Psychometrika, 46, #4, 443-458.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized Adaptive Testing (pp. 141-145). Washington, DC: American Psychological Association.

Kantorovich, L. V. (1968). Functional analysis and applied mathematics. Washington, DC: NBS.

Krass, I. A. (1997, June). Getting more precision on computer adaptive testing. Paper presented at the 62nd Annual meeting of Psychometric Society, University of Tennessee, Knoxville, TN.

Krass, I. A. (1998, April). Application of direct optimization for On-Line calibration in computer adaptive testing. Paper presented at the 1998nd Annual meeting of National Council on Measurement in Education, San Diego, CA.

Levine, M. V. (1984). An introduction to multilinear formula scoring theory. (Office of Naval Research Report 84-4). Champaign, IL: University of Chicago.

Levine, M. V. Williams B. A. (1998). Development and Evaluation of Online Calibration Procedures. (Algorithm Design and Measurement Services, Inc. TCN # 96-216) Champaign, IL.

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

McLachlan, G. J., & Krishnan, T. (1997). The EM algorithm and Extensions. New York: John Wiley & Sons.

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

Ramsay, J. O. (1975). Solving implicit equations in psychometric data analysis, Psychometrika, 40, 337-360.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 38, 221-233.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized Adaptive Testing (pp. 131-140). Washington, D C: American Psychological Association.

Thissen, D., & Steinberg. L. (1984). A model for multiple choice items. Psychometrika, 49, 501-519.

Wilson, T. W., Wood, R, Gibbons R. (1991). TESTFACT – Test Schoring, and Item Factor Analysis. Scientific Software International, Chicago

Wingersky, M. S., Barton, M. A., & Lord F. M. (1982). LOGIST User's Guide. Princeton, NJ: Educational Testing Service

Zimowski, M. F., & Bock, R. D. (1987). Full-information item factor analysis from the ASVAB CAT pool. (Methodology Research Center Report #87-1), Chicago: University of Chicago.

Zimowski, M. F., Muraki, E., Mislevy, R. J., Bock, R. D. (1966) BILOG-MG. Multiple-group IRT Analysis and Test Maintenance for Binary items. Scientific Software International, Chicago

Table 1. Variances of 3PL parameters in the "Normal" simulation

	A-parameter		B-parameter		C-parameter	
	BLG	DMAP	BLG	DMAP	BLG	DMAP
2000	0.0378	0.022	0.023	0.0154	0.0007	0.0041
1500	0.0231	0.0395	0.0231	0.0122	0.0004	0.0043
1000	0.0308	0.0428	0.0237	0.0169	0.0005	0.0051
750	0.0342	0.0428	0.025	0.0262	0.0006	0.0067
500	0.0668	0.0923	0.0318	0.0246	0.0003	0.0072
300	0.0942	0.2462	0.0428	0.0419	0.0006	0.0071

Table 2. Average distances between ICCs

	BLG	DMAP
2000	0.0221	0.0185
1500	0.0227	0.0196
1000	0.0235	0.0247
750	0.0254	0.0258
500	0.0322	0.0292
300	0.0399	0.0416

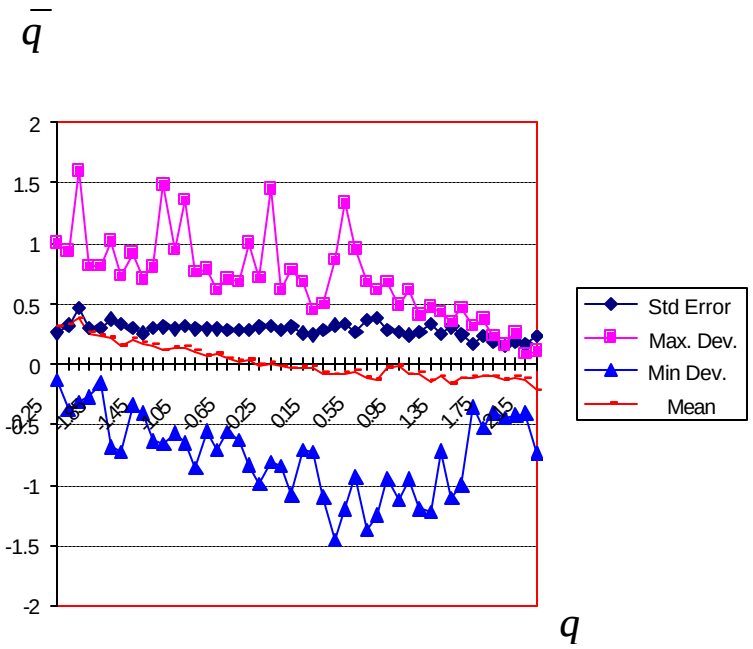


FIGURE 1

Results of AR simulation after standard Bayesian implementation.

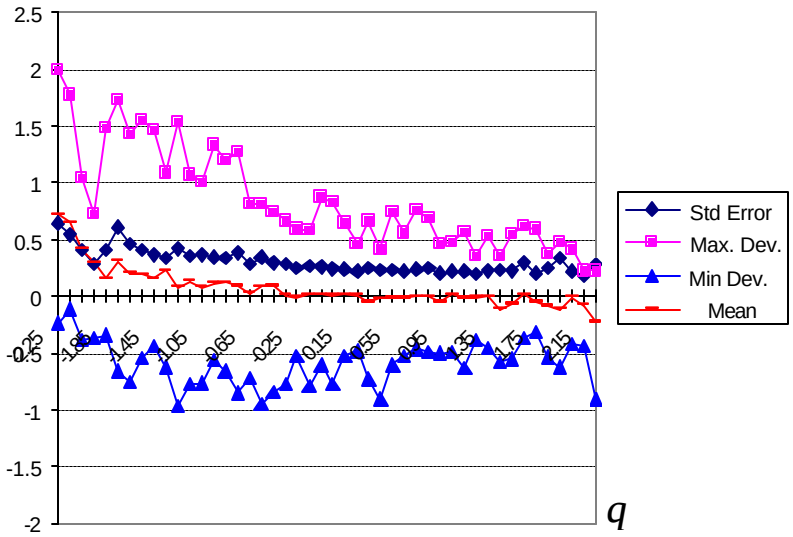
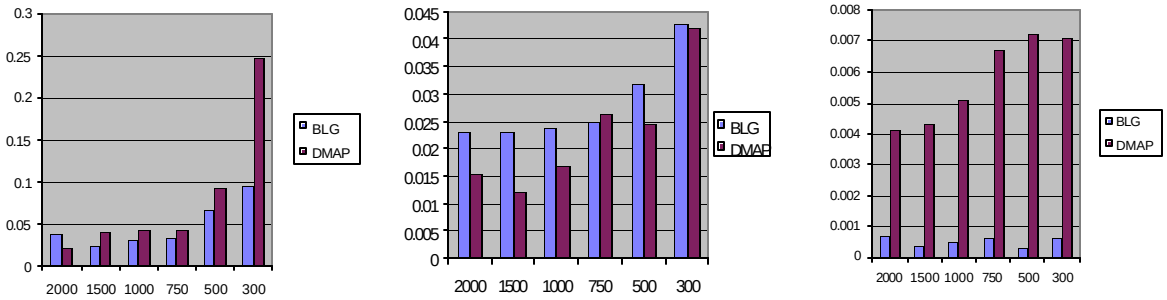


FIGURE 2.

Results of AR simulation after DMAP implementation.



Variation of a. Variation of b. Variation of c.

FIGURE 3.

Variances of 3-PL parameters in the "Normal" simulation.

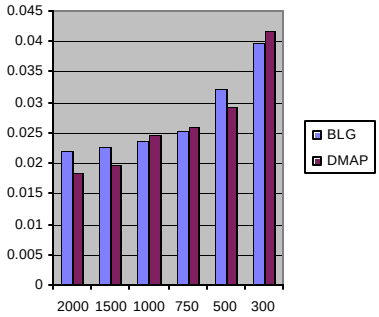


FIGURE 4.

Average distances between ICCs.

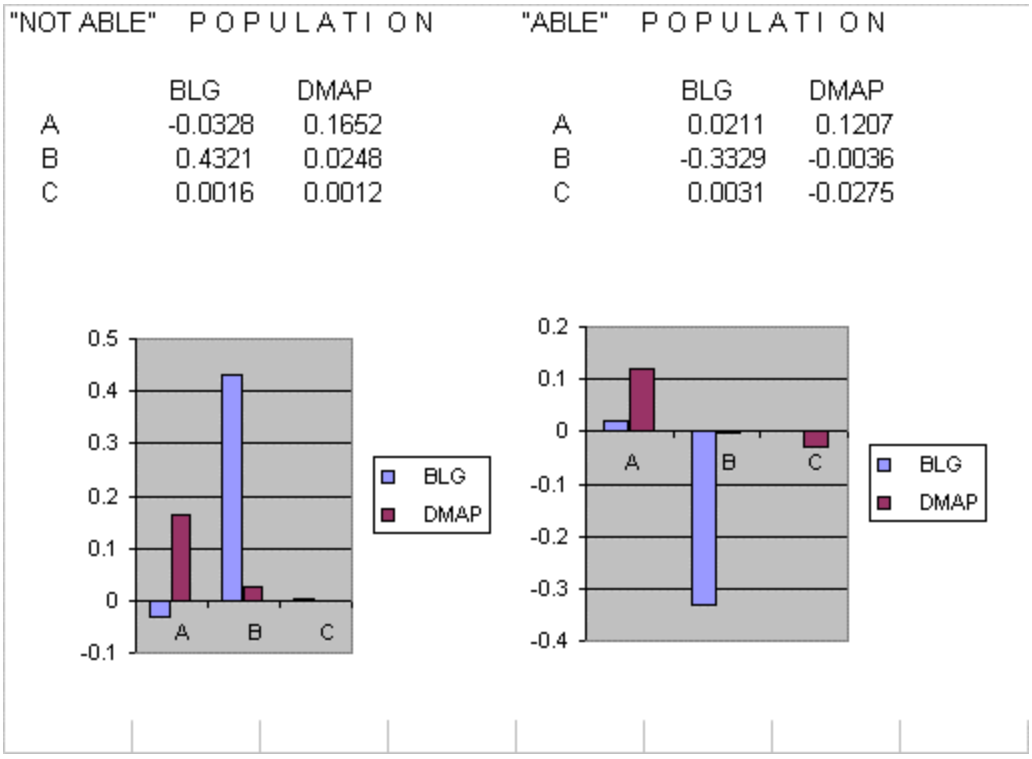


FIGURE 5.

Biases in the case of not “Normal” population.

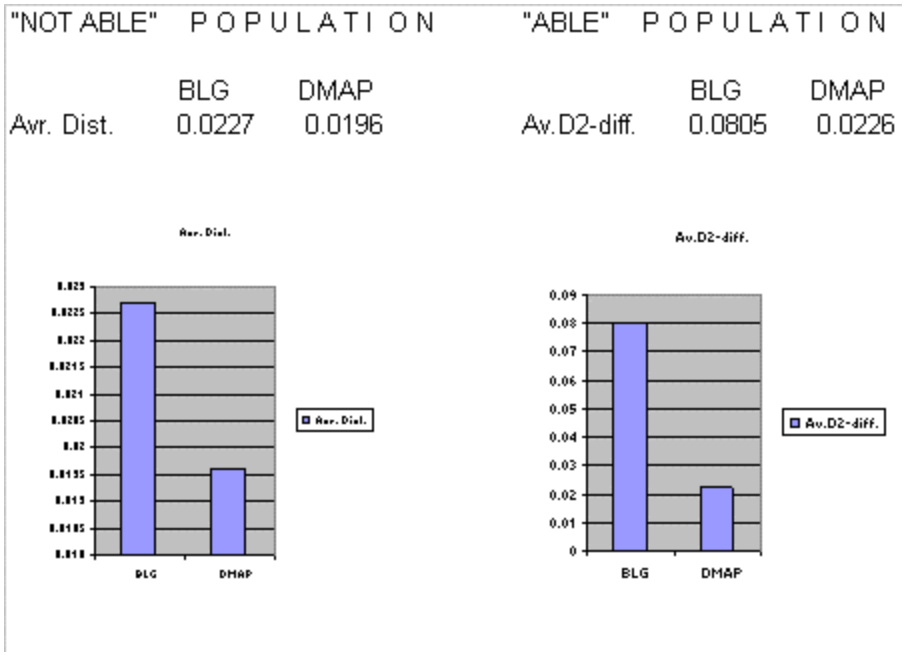


FIGURE 6.

Weighted ICCs differences in the case of not “Normal” population.

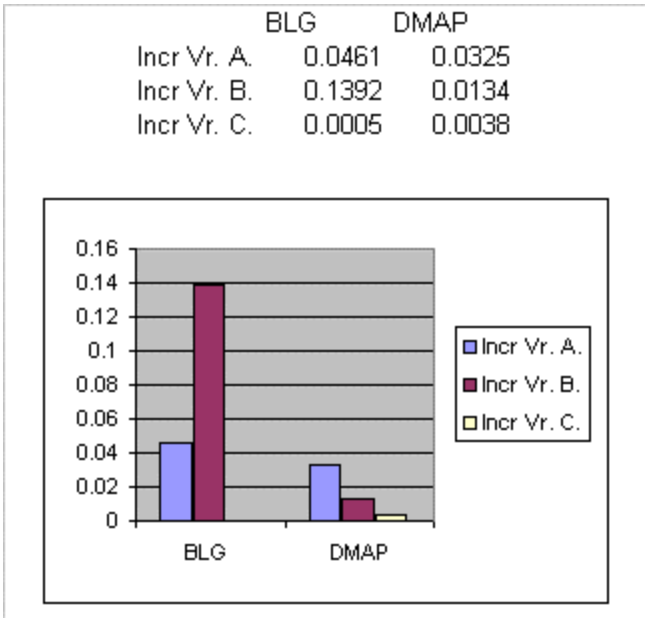


FIGURE 7.

Increases of variances in three-dimensional case.

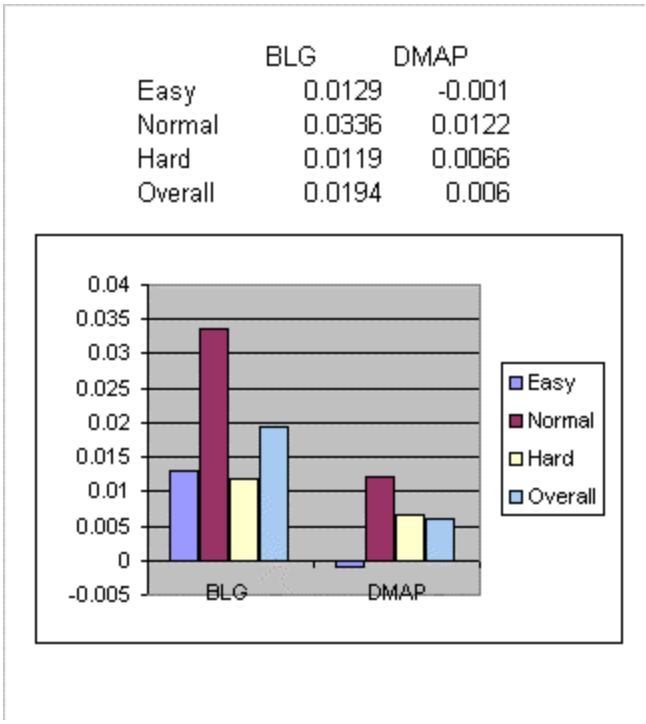


FIGURE 8.

Increase in distances between ICCs.

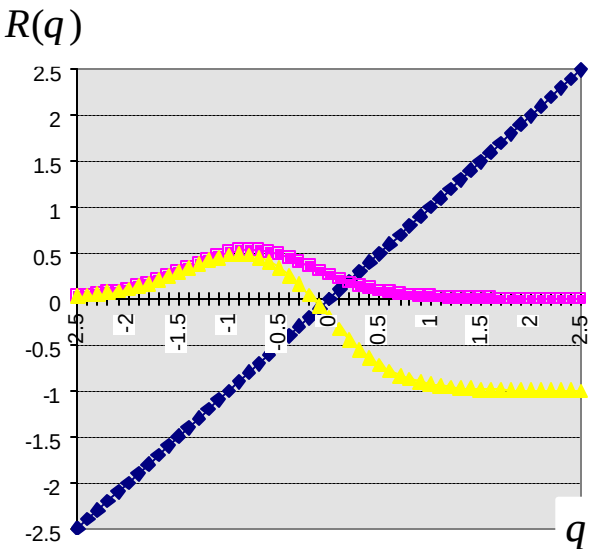


FIGURE 9.

The case of a test of length two, where the first item was answered correctly and the second wrongly.