

To appear in N. Flournoy, W. F. Rosenberger & W. K. Wong (Eds.) *New developments and applications in experimental design*. Berkeley: Institute of Mathematical Statistics.

OPTIMAL DESIGN FOR ITEM CALIBRATION IN COMPUTERIZED ADAPTIVE TESTING: THE 2PL CASE

STEVEN G. BUYSKE

ABSTRACT. Item Response Theory is the psychometric model used for standardized tests such as the Graduate Record Examination. A test-taker's answer on an item is modelled as a binary response with success probability depending on parameters for the test-taker and the item. The advent of computerized adaptive versions of these tests leads to sequential design problems. We show how the need for estimation of the item parameters with their ultimate use in mind leads to a locally L-optimal design criterion. A sequential implementation of the optimal design is presented, which is 52% more efficient than the most common current design.

1. INTRODUCTION

The basis for modern standardized testing is known as Item Response Theory [6]. The key idea is that each test question, generally called an item, is characterized by a few parameters, and each test-taker is characterized by a single parameter, generally called proficiency or ability. The probability that a given test-taker answers a given item correctly is given by a function of both the item's and the test-taker's parameters. Conditional on those parameters, the response on one item is independent of the responses to other items.

The model that we focus on in this paper is

$$P(\text{correct response} \mid a, b, \theta) = \frac{1}{1 + \exp(-a(\theta - b))}.$$

Here θ is the test-taker's proficiency, and a and b are item parameters. Generally a is called the discrimination and b the difficulty parameter. This model is known as the 2 Parameter Logistic, or 2PL, model. Another popular model, the 3PL model, includes a non-zero left asymptote, and is treated in [3].

Key words and phrases. online calibration, sequential design, 2PL, CAT.
Research supported in part by an ETS Psychometric Fellowship.

Historically, essentially equivalent paper-and-pencil tests were given at a fixed time to an extremely large number of people. In that case, the item and test-taker parameters could be jointly estimated by maximum likelihood methods. Current interest in standardized tests has shifted to Computerized Adaptive Tests (CAT). Because individuals can take the test at almost any time, on a CAT the items must be previously calibrated and the item parameters treated as known. For a given test-taker, an algorithm adaptively picks items so as to efficiently estimate the test-taker's proficiency subject to various content constraints. Prominent tests currently offered in a CAT format are the Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT), and Test of English as a Foreign Language (TOEFL).

While the theory of designs for estimating test-takers' proficiencies is reasonably mature (see [4] or [5] for some of the more statistical recent work), little work has been done on designs for calibrating test items. There appear to be just two existing designs to calibrate new items. The first is essentially an online version of paper-and-pencil calibration. Uncalibrated items, called pretest items, are selected at random for a given test-taker. After an item has been given to a specified number of test-takers, its parameters are estimated and the item is placed in production.

A second design has been proposed by Berger ([1], [2]) and by Jones and Jin [7]. This design is a sequential locally D-optimal design. Given current estimates of the parameters, test-takers whose $\hat{\theta}$'s give a probability $P(\hat{a}, \hat{b}, \hat{\theta})$ near .18 or .82 are chosen with equal weights. Their responses are used to update the estimates, and then new test-takers are chosen. The process continues until a fixed sample size is obtained.

In this paper we will propose a new design based on a criterion that incorporates the ultimate use of the items.

2. DESIGN CRITERIA

CAT algorithms sequentially adapt to each test-taker's proficiency estimate. A 2PL algorithm generally selects items with difficulty b near the current proficiency estimate $\hat{\theta}$, subject to certain content constraints on the items, because an item with $b = \theta$ maximizes the Fisher information for θ . A good algorithm will select items with low discrimination a early and high discrimination later [4], because high discrimination items carry higher Fisher information when b is near the true value of θ .

As noted in the introduction, in a CAT the items must be calibrated before they are used. We can think of the test-takers used in calibration as the calibration test-takers, as opposed to the production test-takers, for whom the item will be used to help determine their proficiency estimates. The design

question is how to pick the calibration test-takers. Since the principal reason to calibrate the items is to be able to estimate the proficiencies of future (production) test-takers, it seems natural to insist that our criterion should be to give the best results in estimating the proficiencies of the production test-takers.

The general criterion leads to two explicit criteria. First, the production θ estimates should be unbiased. Since the item difficulty b and the test-taker's proficiency θ are measured on the same arbitrary scale, it is important that the production θ estimates be unbiased so as to avoid drift in the scale. Note that we are *not* requiring that the item parameter estimates be unbiased.

The second explicit criterion is that the variance of $\hat{\theta}$ due to calibration errors be minimized. Again, this is not the same as requiring some minimal measure of variance for the item parameter estimates.

First we look at the unbiasedness criterion. Errors in estimating the item parameters will generally lead to bias in the estimates of the production θ 's. Following [8], suppose that $\Psi(\mathbf{a}, \mathbf{b}, \mathbf{Y}, \theta) = 0$ is the estimating equation for θ , where \mathbf{a} is the vector of discrimination parameters for the production items seen by the test-taker, \mathbf{b} is similarly the vector of difficulty parameters, \mathbf{Y} is the vector of responses, and θ is the test-taker's proficiency. The maximum likelihood estimator $\hat{\theta}$ is the solution to $\Psi(\mathbf{a}, \mathbf{b}, \mathbf{Y}, \hat{\theta}) = 0$. Now if $\epsilon_{\mathbf{a}}$ and $\epsilon_{\mathbf{b}}$ are random vectors such that $E\epsilon_{\mathbf{a}} = E\epsilon_{\mathbf{b}} = 0$ and $E\epsilon_{\mathbf{a}}\epsilon_{\mathbf{a}}^t = E\epsilon_{\mathbf{b}}\epsilon_{\mathbf{b}}^t = I$, where I is the identity matrix, then we can write $\hat{\theta}(\sigma)$ as the solution to $\Psi(\mathbf{a} + \sigma\epsilon_{\mathbf{a}}, \mathbf{b} + \sigma\epsilon_{\mathbf{b}}, \mathbf{Y}, \hat{\theta}(\sigma)) = 0$. In general, $E(\hat{\theta}(\sigma)) \neq \theta$. However, it is not difficult to show that for an item used for a test-taker with $\theta = b + \delta$ we have

$$E(\hat{\theta}(\sigma)) = \theta + \sigma^2 \sum_j \left(\text{Cov}(\hat{a}_j, \hat{b}_j)(1/a_j + O(\delta^2)) + \text{Var}(\hat{a}_j)O(\delta^3) + \text{Var}(\hat{b}_j)O(\delta) \right) + O(\sigma^3). \quad (1)$$

As the distribution of δ will be approximately symmetric about zero, the odd order terms in δ will vanish when we take the expected value over the entire test-taking population. Thus if $\text{Cov}(\hat{a}, \hat{b}) = 0$, we will have second order unbiasedness. Since $\text{Cov}(\hat{a}, \hat{b})$ has a numerator equal to

$$\sum_i a(\theta_i - b) \frac{\exp(-a(\theta_i - b))}{(1 + \exp(-a(\theta_i - b)))^2},$$

if the calibration θ 's are symmetric about b then $\text{Cov}(\hat{a}, \hat{b}) = 0$.

To understand the item calibration component of the variance of $\hat{\theta}$, suppose item (a_1, b_1) has been calibrated by test-takers with proficiencies θ_1 ,

$\dots, \theta_i, \dots, \theta_{m-1}$. The item is now placed in production and exposed to a test taker with proficiency θ_m . The test-taker's proficiency will be estimated on the basis of the responses to items $(a_1, b_1), \dots, (a_j, b_j), \dots, (a_n, b_n)$. The score function for θ_m is thus

$$V = \sum_{j=1}^n a_j (Y_{m,j} - P(a_j, b_j, \theta_m)),$$

while the score function for (a_1, b_1) is

$$W = \sum_{i=1}^m \begin{bmatrix} a_1 \\ (\theta_i - b_1) \end{bmatrix} (Y_{i,1} - P(a_1, b_1, \theta_i)).$$

Write $I_i(a_1 b_1, a_1 b_1)$ for the variance of a single summand of W , and write $I(a_1 b_1, a_1 b_1) = \sum_i I_i(a_1 b_1, a_1 b_1) = \text{Var}(W)$, $I(\theta_m, \theta_m) = \text{Var}(V)$, and $I(\theta_m, a_1 b_1) = \text{Cov}(V, W)$.

Since we will not be using $Y_{m,1}$ to refine our estimate of (a_1, b_1) , we can consider (a_1, b_1) as incidental. If we suppose that the parameters of all the other items are known perfectly, then the information for θ_m is

$$I(\theta_m, \theta_m) - I(\theta_m, a_1 b_1) I(a_1 b_1, a_1 b_1)^{-1} I(\theta_m, a_1 b_1)^T. \quad (2)$$

By conditional independence, $I(\theta_m, a_1 b_1)$ depends only on θ_m, a_1 , and b_1 , while $I(a_1 b_1, a_1 b_1)^{-1}$ is closely approximated by the variance matrix for (\hat{a}_1, \hat{b}_1) arising from the calibration (i.e., $(\sum_{i=1}^{m-1} I_i(a_1 b_1, a_1 b_1))^{-1}$). Since

$$\begin{aligned} & I(\theta_m, a_1 b_1) \left(\sum_{i=1}^{m-1} I_i(a_1 b_1, a_1 b_1) \right)^{-1} I(\theta_m, a_1 b_1)^T \\ &= \text{tr} \left[\left(\sum_{i=1}^{m-1} I_i(a_1 b_1, a_1 b_1) \right)^{-1} I(\theta_m, a_1 b_1)^T I(\theta_m, a_1 b_1) \right], \end{aligned}$$

where tr denotes trace, represents the information lost due to calibration, this is the function of the design that we want to minimize. Actually, the calculations are somewhat easier if we reduce the proportional information loss, namely

$$\text{tr} \left[\left(\sum_{i=1}^{m-1} I_i(a_1 b_1, a_1 b_1) \right)^{-1} I(\theta_m, a_1 b_1)^T I(\theta_m, a_1 b_1) / I^1(\theta_m, \theta_m) \right],$$

where $I^1(\theta_m, \theta_m)$ is the component of $I(\theta_m, \theta_m)$ contributed by (a_1, b_1) .

Unfortunately, this expression depends on θ , which is the proficiency of a future test-taker who is given the item when it is in production. Thus this criterion for calibrating the item depends on exactly how it will be used. To cover a plausible range of possible uses, we will integrate the expression against a prior for $P = P(\text{correct response} | a, b, \theta)$. A beta (ν, ν) prior

works well and is consistent with simulation results. Because for a fixed item the prior on P is equivalent to a prior on θ , and because θ and b are on the same scale, obtaining our criterion by integrating P and θ out is formally equivalent to obtaining a Bayes design with a prior on b and a unit mass prior on a .

Dropping subscripts, we have the integrated proportional information loss equals

$$\int \text{tr} \left[\left(\sum_{i=1}^{m-1} I_i(ab, ab) \right)^{-1} I(\theta, ab)^T I(\theta, ab) / I(\theta, \theta)^1 \right] p(\theta) d\theta \quad (3)$$

$$=: \text{tr} \left[\left(\sum_{i=1}^{m-1} I_i(ab, ab) \right)^{-1} T \right],$$

where $T = \int I(\theta, ab)^T I(\theta, ab) / I(\theta, \theta)^1 p(\theta) d\theta$. This shows that the criterion can be considered as an L -optimality criterion. Interestingly, T can be shown to be equal to the Fisher information for (a, b) from the test-takers after the item is in production. Thus the criterion is in some sense the information gained but not used in production divided by the information gained and used during calibration. Some calculation shows that

$$\text{tr} \left[\left(\sum_{i=1}^{m-1} I_i(ab, ab) \right)^{-1} T \right] \propto \left(\frac{1}{a^2} \text{Var} \hat{a} + a^2 \text{const}(\nu) \text{Var} \hat{b} \right), \quad (4)$$

where a range of values for the constant is given in Table 1. An appropriate choice of ν depends on the size of the pool of items and on the number of item content constraints. Generally, a large pool suggests a larger ν , giving a narrow spread around the optimal value of $P = .5$, while many constraints suggest a smaller value for ν . For the rest of this paper, we will use a value of $\nu = 4$, or a criterion of

$$\frac{1}{a^2} \text{Var} \hat{a} + 2.26a^2 \text{Var} \hat{b}$$

to minimize. The final design is not very sensitive to ν .

Given this criterion one can find, and verify through the General Equivalence Theorem, that the local optimal design puts equal weight at θ_1 and θ_2 , where

$$P(a, b, \theta_1) = .25 \quad \text{and} \quad P(a, b, \theta_2) = .75.$$

This can be compared to the local D-optimal design, which puts equal weight at $\tilde{\theta}_1$ and $\tilde{\theta}_2$, where

$$P(a, b, \tilde{\theta}_1) = .18 \quad \text{and} \quad P(a, b, \tilde{\theta}_2) = .82.$$

ν	1	2	4	6	8
const(ν)	.76	1.27	2.26	3.26	4.25
upper design pt	.80	.78	.75	.73	.72

TABLE 1. Numerical values depending on the prior parameter ν .

Because the proposed design points are not as far out in the tails as the D-optimal design, the proposed design should be less dependent on the correctness of the model. Additionally, from the test-taker's point of view, the proposed design is less extreme, which is important psychologically.

Thus, the proposed calibration procedure is

1. Estimate a and b .
2. Pick θ_1 's and θ_2 's with equal weight, where θ_i is defined by

$$P(a, b, \theta_1) = .25 \quad \text{and} \quad P(a, b, \theta_2) = .75.$$

3. Use the resulting responses to re-estimate a and b .
4. Repeat until pre-specified sample size is reached, or use a stopping rule based on the optimization criterion. A rule of repeat until $\widehat{\text{Var}}(\hat{a})/\hat{a}^2 + 2.26\hat{a}^2\widehat{\text{Var}}(\hat{b}) < \text{cutoff}$ is effective in simulations.

We note that because the optimality criterion is not linear in the design points, if all of the design points were not placed optimally for the current estimate, then the new design points given in step (2) above will not be optimal; the optimal points should be calculated based on the previous points. The increased computation does not seem worth the slight loss of efficiency, however.

3. SIMULATION RESULTS

In this section we present some simulation results. Ten thousand items were generated with $a \sim \text{unif}(.5, 2.5)$ and $b \sim N(0, 1)$. Each item was calibrated with seven different methods: the standard calibration, meaning 400 θ 's chosen randomly from $N(0, 1)$; the proposed design, namely the procedure outlined in the previous section, using the stopping rule; the D-optimal design, but using the stopping rule instead of a fixed sample size (note that the stopping rule is based on the proposed criterion, not the D-optimal criterion, so that this design is actually intermediate between the D-optimal design and the proposed design); the standard calibration, but using the stopping rule instead of a fixed sample size; and finally both the proposed design and the D-optimal design with a step size of 30, and with the θ 's picked at the design points plus a $N(0, (.25)^2)$ variable. Except for

Design	Log(Criterion) using MSE in place of variance		Calibration Sample Size	
	Mean	St Dev	Mean	St Dev
Standard	-3.4	1.5	400	0
Proposed	-3.4	1.3	263	2.1
D-optimal	-3.5	1.4	291	17
Standard with stopping	-3.2	1.5	352	202
Proposed with jitter	-3.4	1.3	276	12
D-optimal with jitter	-3.5	1.4	305	21

TABLE 2. Calibration results using various designs. “Standard with stopping” refers to the standard design but using the stopping rule instead of having a fixed length. The designs with jitter have a larger step size and the design points are $N(0, (.25)^2)$ distributed around the optimal points.

these last two designs, all of the sequential designs have a step size of 2. In all designs with a stopping rule, the cutoff was 0.065.

Table 2 summarizes the results of the calibration simulations. For the criterion function, the mean squared error version is used. That is, $\text{mean}(\hat{a} - a)^2$ is used in place of $\text{Var} \hat{a}$, and so on.

Figures 1, 2, and 3 graphically compare results from 1500 items for the various designs. Figure 1 shows a versus the error in estimating b . The cone shape illustrates how the proposed design generates smaller errors than the standard design in \hat{b} for higher values of a ; the other designs are intermediate. Figure 2 shows that the standard design generates larger errors in \hat{b} for more extreme values of b than the proposed; again the other designs are intermediate. Ironically, in a high-stakes test, errors in estimating b for large values of b cause θ estimation errors exactly where the stakes are highest. Figure 3 shows, on a log scale, the distribution of the criterion function in terms of b .

Table 3 shows the results of using the calibrations to estimate proficiencies. For each design, θ was estimated for 1000 $N(0, 1)$ simulated test-takers, and then 1000 test-takers with $\theta = 0$ and $\theta = 2$ using a 30 item test. The bias and mean squared error is given for each design, as well as for “perfect calibration,” when the item parameters are known perfectly. Figure 4 shows the distribution of errors in $\hat{\theta}$ as a function of θ . The important point here is that the different designs give essentially similar results

Design	$\theta \sim N(0, 1)$		$\theta = 0$		$\theta = 2$	
	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$
Perfect	-0.003	0.24	-0.012	0.24	0.006	0.26
Standard	-0.009	0.26	-0.002	0.25	0.062	0.27
Proposed	-0.013	0.25	0.00007	0.25	0.018	0.26
D-optimal	-0.009	0.25	-0.0002	0.25	0.052	0.26
Standard with stopping	-0.003	0.25	-0.010	0.25	0.038	0.27

TABLE 3. Proficiency estimation using calibrated items. The “perfect” design refers to proficiency estimation using the actual parameter values of the items. The designs performed about equally well, although the proposed, D -optimal, and standard-with-stopping designs required calibration samples just 66%, 73%, and 88% as large as the standard design.

but the proposed, D -optimal, and standard-with-stopping designs required calibration samples just 66%, 73%, and 88% as large as the standard design.

4. IMPLEMENTATION CONSIDERATIONS

One interesting aspect, from the design point of view, of the item calibration problem is that one cannot actually pick the optimal design points θ_i . Test-takers, each with a different proficiency θ_i , show up sequentially and at random. Additionally, at any one time there are a number of different items to calibrate, and each test-taker needs to help calibrate a specified number of items. One scheme for handling this aspect would be that for each test-taker who needs to calibrate k items, the algorithm would pick the k items for which the test-taker’s proficiency θ is closest to the desired design points. In this case, “closest” would refer not to the θ metric but to the probability metric. Simulation results indicate that little efficiency is lost when the calibration θ ’s are not exactly on the design points. A design using a similar approach is implemented in [3].

Another issue is that the θ s, although treated as known in this paper, are of course estimates. Since the 2PL model is based on the logistic function, the functional measurement error maximum likelihood estimator introduced in [9] can be used to reduce bias, as was done in [7].

Some computerized testing programs calibrate new items in a separate section, while others seed pretest items in among the production items. If an experimental section is used, and if it comes after the relevant production section, there is no difficulty. If an experimental section is used and

it comes after some sections, but not the relevant one, then the fact that an individual's proficiencies on different aspects of a test are highly correlated can be used. If the experimental section comes first, then calibration items at the very beginning of the calibration process can be selected at random. Finally, if pretest items are seeded in among production items, then items early in their calibration process can be used early in an individual test, when the proficiency is poorly estimated. When an item is later in its calibration process, so that its parameters are better calibrated, it can be used later in an individual test when the proficiency is better estimated.

5. DISCUSSION

While computerized adaptive testing was initiated to increase the efficiency of proficiency estimation, it also opens the possibility of increased efficiency of item parameter estimation. The calibration design called "standard" here is simply the random selection of test-takers for each item; this method is in general use. A D-optimal sequential design has also been proposed. In this paper we have proposed a criterion for calibrating items based on their ultimate purpose, namely proficiency estimation. This criterion can be used in two ways. The first way is to use the criterion to pick the optimal design measure. The proposed design picks points that are less extreme than the D-optimal design, a feature that should be more comfortable to test-takers and, because the tails of the quantile response function are most sensitive to the model, more robust to model mis-specification. The second, and probably more important, way to use the criterion is as a measure of how far advanced the calibration of a specific item is. Merely adding the stopping rule to the standard design results in a method that is equally effective for proficiency estimation but 14% more efficient in terms of calibration cost. Even better improvements are possible using the stopping rule with the D-optimal or proposed design points, which are, respectively, 37% and 52% more efficient than the standard design with the same effectiveness for proficiency estimation.

ACKNOWLEDGMENTS.

The author would like to thank his dissertation advisor, Zhiliang Ying, for introducing him to Item Response Theory and CAT.

REFERENCES

- [1] Berger, M. P. F. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika* **57** 521–538.
- [2] Berger, M. P. F. (1994). D-optimal sequential sampling design for item response theory models. *Journal of Educational Statistics* **19** 43–56.

- [3] Buyske, S. G. (1998). Item calibration In computerized adaptive testing using minimal information loss. Preprint.
- [4] Chang, H. H., and Ying, Z. (1996) A global information approach to computerized adaptive testing. *Applied Psychological Measurement* **20** 213–229.
- [5] Chang, H. H., and Ying, Z. (1997). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics* to appear.
- [6] Hambleton, R. K., and Swaminathan, H. (1985). *Item Response Theory*, Kluwer-Nijhoff, Boston.
- [7] Jones, D. H., and Jin, Z. (1994). Optimal sequential designs for on-line item estimation. *Psychometrika* **59** 59–75.
- [8] Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* **72** 583–592.
- [9] Stefanski, L. A., and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics* **13** 1335–1351.
- [10] Stocking, M. L. (1994). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika* **55** 461–475.

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
PISCATAWAY, NJ 08855
EMAIL: buyske@stat.rutgers.edu

STATISTICS DEPARTMENT, RUTGERS UNIVERSITY, PISCATAWAY, NJ 08855 U.S.A.
E-mail address: buyske@stat.rutgers.edu

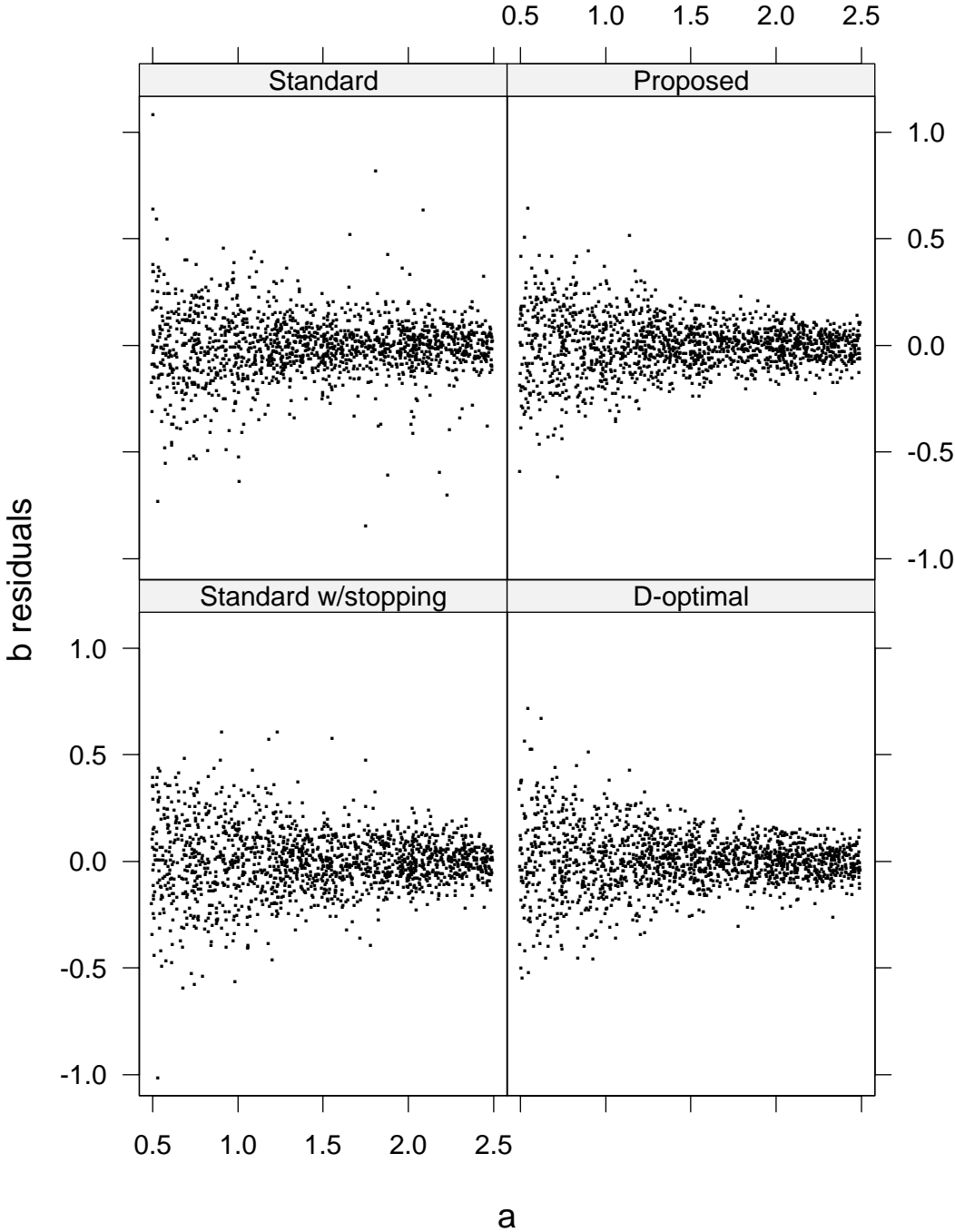
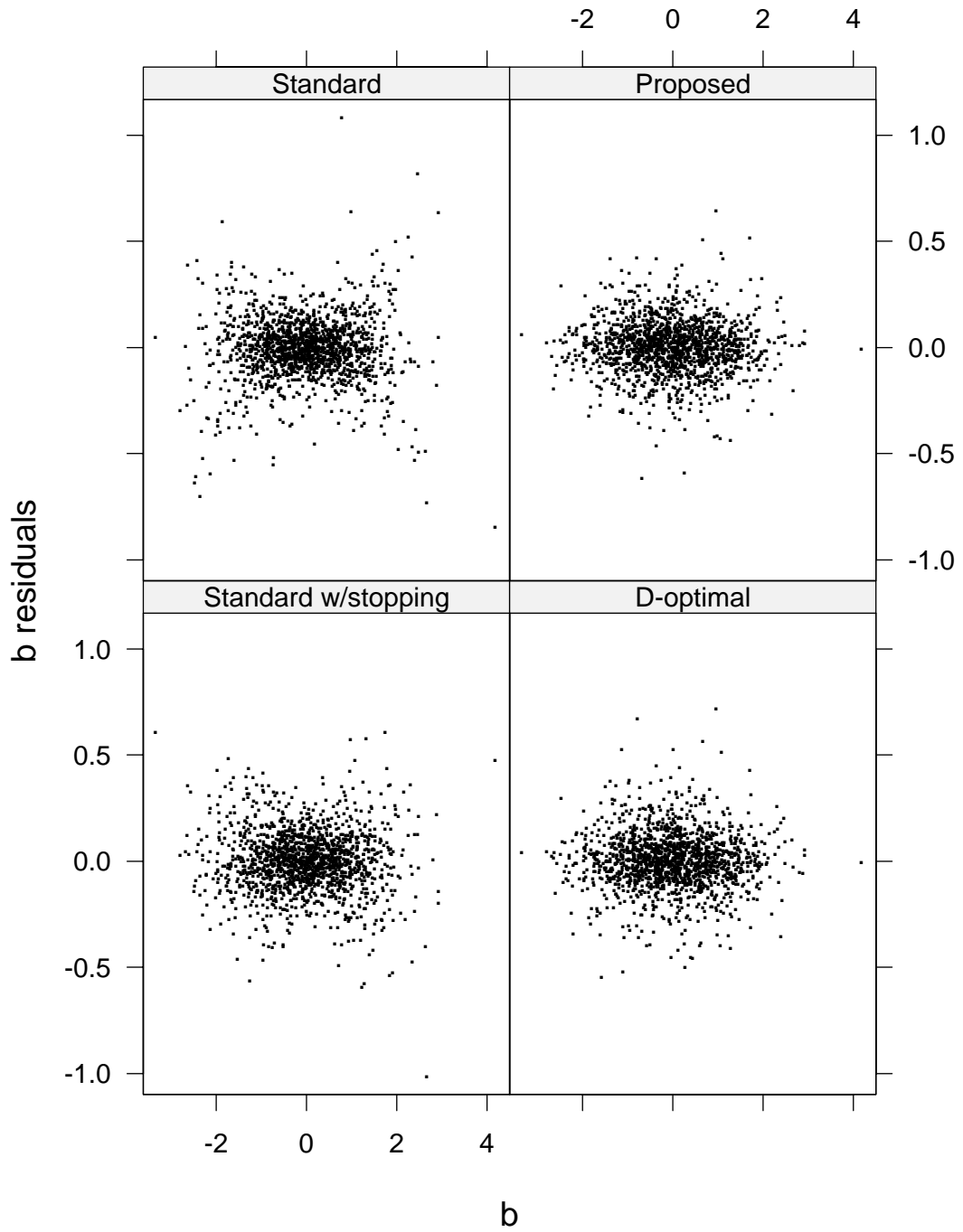


FIGURE 1. Errors in \hat{b} by a .

FIGURE 2. Errors in \hat{b} by b .

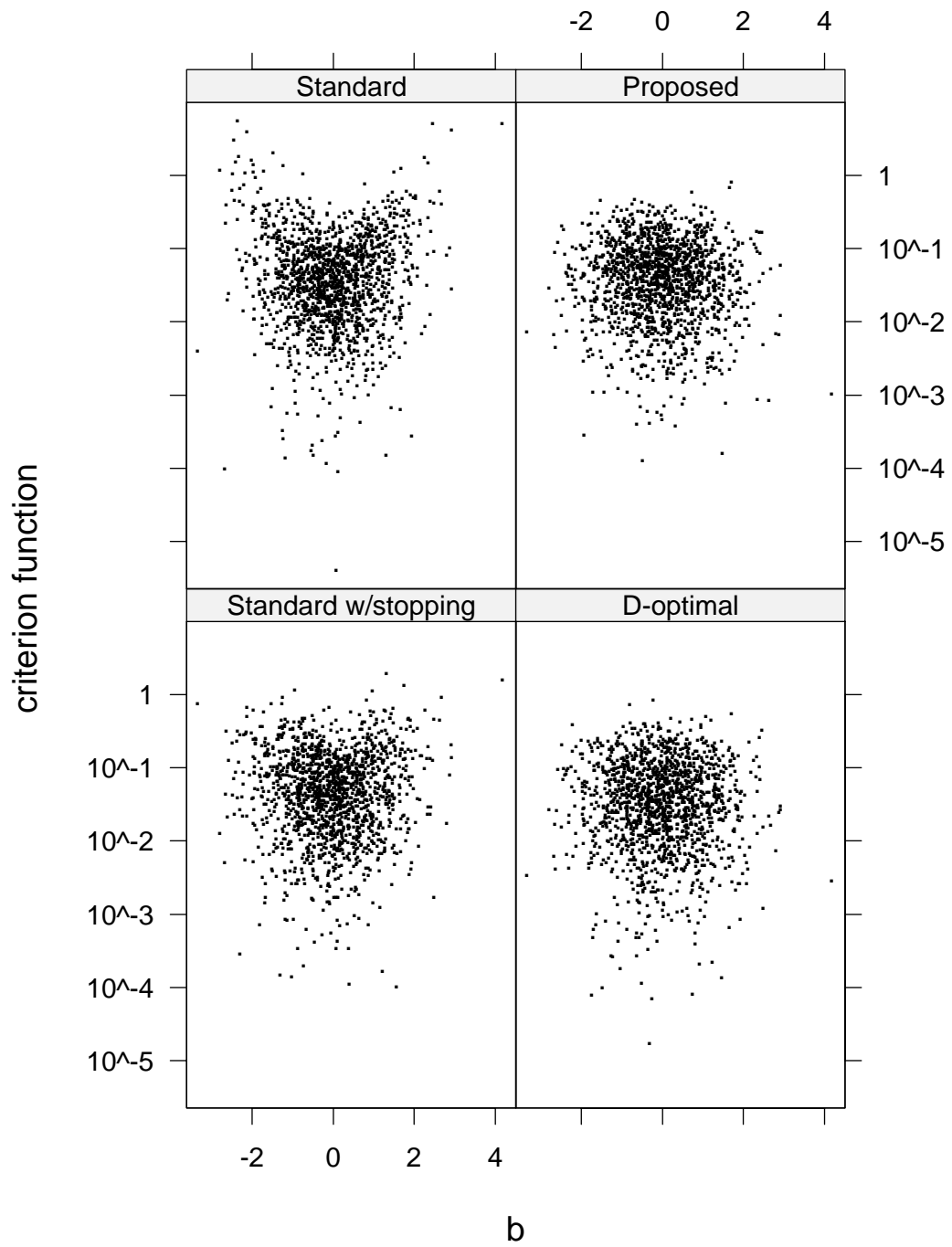
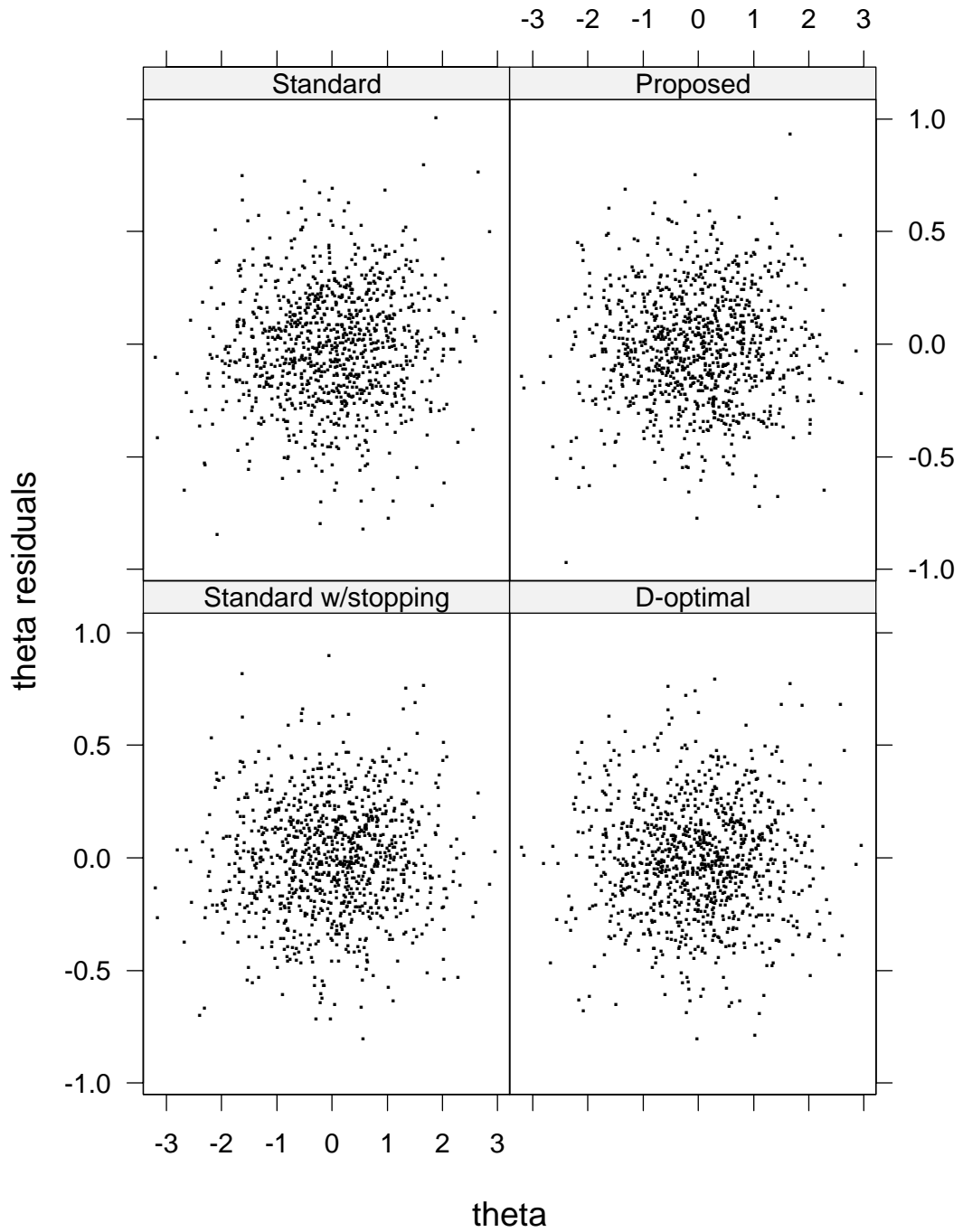


FIGURE 3. Values of the criterion function by b . Smaller is better.

FIGURE 4. Errors in θ estimation following calibration.