

■ **An Empirical Bayes Enhancement of
Mantel-Haenszel DIF Analysis for
Computer-Adaptive Tests**

**Rebecca Zwick
University of California, Santa Barbara**

**Dorothy T. Thayer
Educational Testing Service**

■ **Law School Admission Council
Computerized Testing Report 98-15
August 2003**

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 201 law schools in the United States and Canada.

Copyright© 2003 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any part or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, 661 Penn Street, Box 40, Newtown, PA 18940-0040

LSAT® and LSAC are registered marks of the Law School Admission Council, Inc.

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

Table of Contents

Executive Summary	1
Introduction	1
Bayesian Methods in Psychometrics and Educational Research	2
EB Enhancement of Mantel-Haenszel DIF Analysis	2
<i>Statistical Model for the EB DIF Approach</i>	3
<i>Validity Evidence</i>	5
Simulation Study.	6
<i>Simulation Conditions</i>	7
<i>Model and Parameters for Data Generation</i>	7
<i>Item Calibration and Development of Easy, Medium, and Hard Testlets</i>	8
<i>CAT Administration and Ability Estimation</i>	9
<i>Number of Replications Per Condition</i>	9
<i>DIF Analysis</i>	9
Analysis and Results	9
<i>Analysis of DATA From Preliminary Simulations</i>	9
<i>Results of Main Simulation</i>	11
Conclusion and Ideas for Future Research	20
References	21
Appendix	24

Executive Summary

The Law School Admission Council (LSAC) is currently investigating the feasibility of implementing a computer adaptive version of the Law School Admission Test (LSAT). The introduction of computer adaptive tests (CATs) requires that new approaches be developed for the analysis of item properties, including differential item functioning (DIF). DIF is said to occur when test takers from two demographic groups (say, men and women) perform differently on an item even after they have been matched in terms of overall ability. The presence of DIF may point to unintended sources of difficulty in the item (e.g., a math item may require sports knowledge that is more common in men than in women). A significant technical challenge in assessing DIF in CATs is the need to develop a method that will produce stable* results in small samples: Even if the total number of test takers for a CAT is large, the number of responses to some items may be very small.

Currently within the testing industry, the Mantel-Haenszel DIF analysis method is the most commonly used to detect DIF for paper-and-pencil tests. In fact, this is the analysis used for the LSAT. A body of statistical methods referred to as empirical Bayes (EB) methods are known to be capable of producing stable statistical results with fewer test takers. The present study investigated the applicability to CAT data of a DIF analysis method that involves an EB enhancement of the popular MH DIF analysis method.

The computerized LSAT test design assumed for this study was similar to that currently being evaluated for a potential computerized LSAT. Here, rather than being presented with a single test item at a time, test takers are presented with small groups of items, commonly referred to as testlets. The CAT pool for this research consisted of 10 five-item testlets at each of three difficulty levels. The item parameters, which are statistics that describe the various item characteristics such as item difficulty, were specified to resemble those typically observed for the LSAT. Using these item-level statistics, responses to the test questions were generated for simulated test takers. These simulations consisted of four conditions that varied in terms of group sample sizes and group ability distributions; both of these factors are known to affect the performance of DIF methods. Sample sizes for the two test taker groups were either 1,000 or 3,000 (before application of the CAT algorithm). The distribution of test taker ability for the two groups were either the same or differed by one standard deviation.

The results showed the performance of the EB DIF approach to be very promising, even in extremely small samples. The EB estimates tended to be closer to their target values than did the ordinary Mantel-Haenszel (MH) statistics; the EB statistics were also more highly correlated with the true DIF values than were the MH statistics.

Introduction

Because computer adaptive tests (CATs) typically involve smaller samples than paper-and-pencil tests for at least some items, standard differential item functioning (DIF) techniques may not provide results with adequate precision in this setting. Zwick, Thayer, and Lewis (1997, 1999, 2000) developed an empirical Bayes (EB) approach to Mantel-Haenszel (MH) DIF analysis which yields more stable and interpretable results in small samples than do conventional procedures and, therefore, seems well suited to adaptive testing conditions. The computations involve the MH indexes and their standard errors, along with an assumed prior distribution for the true DIF parameters. In an earlier study, the EB methods were extensively investigated through simulation study and were applied to paper-and-pencil tests. The current report describes work that was conducted to investigate the applicability of these methods to a large-scale computer adaptive admission test. The study, which is sponsored by the Law School Admission Council (LSAC), is part of a research program (Pashley, 1997) that is intended to investigate the feasibility of a computer adaptive Law School Admission Test (LSAT).

The study, which addressed the technical innovations necessary for application of the EB DIF method to CATs, was based on a simulated test that involved a pool of adaptively administered five-item testlets. A scaled score was computed for each test taker based on responses to five of these testlets; this score was used as the matching variable for DIF analysis. Following this, the EB elaboration of the MH procedure was applied. The resulting DIF statistics were compared to the true (generating) DIF.

We are grateful to Law School Admission Council for their sponsorship, to Joyce Wang for assistance in data analysis, and to Charlie Lewis for consultation. Certain results of this study were presented at the annual meeting of the American Educational Research Association, San Diego, 1998. Some general information on the empirical Bayes method in this report, including tabular material, has appeared in slightly different form in publications by Zwick, Thayer, and Lewis (1997, 1999, 2000).

* An estimation procedure is considered stable if the estimates tend to be close to their target values.

This report addresses the role of Bayesian methods in psychometrics and educational research. It also describes the EB DIF model and presents some of the issues that had to be considered in applying the EB methods to CATs; describes the simulation study that was designed to investigate the CAT DIF procedures; presents the analyses and results; and finally, provides a discussion of the findings and outlines our ideas for future research.

Bayesian Methods in Psychometrics and Educational Research

Bayesian and empirical Bayes methods have found wide application in psychometrics and in educational research. For example, in studies of test validity, a Bayesian or EB approach can yield estimated regression coefficients that are more stable than are the usual least squares coefficients by pooling information from multiple schools. This pooling is achieved by making an assumption about the prior distribution of the true regression parameters across schools. To estimate the regression parameters for a particular school, this prior is combined with the school-level regression model. The mean of the resulting posterior distribution¹ is typically taken as the point estimate of the regression parameter for that school. (An analogous approach can be used to pool data from multiple years.) When the parameters of the prior are estimated from the data, a method of this kind is called *empirical Bayes*. In a fully Bayesian approach, a distribution would be assumed for the parameters of the prior.

A particularly lucid introduction to EB methods is given by Casella (1985); useful descriptions of EB philosophies and estimation methods are given by Efron and Morris (1973) and Braun (1989). EB regression models have been used in test validity studies, beginning with Rubin's important paper (1980) on the LSAT. Braun, Jones, Rubin, and Thayer (1983) discussed a general model for EB regression, which has been applied in several validity studies, such as Zwick (1993). An EB survival model developed by Braun and later modified (Braun & Zwick, 1993) was used to study time to Ph.D. candidacy and time to degree by Zwick and Braun (1988) and Zwick (1991). Charles Lewis and Dorothy Thayer have investigated EB methods for test equating. Finally, many of the hierarchical models that are now popular in educational research can also be characterized as Bayes or EB models. In EB DIF analysis, information is pooled across items to produce DIF estimates that are more stable than the original DIF statistics.

EB Enhancement of Mantel-Haenszel DIF Analysis

The Mantel-Haenszel DIF analysis procedure of Holland and Thayer (1988) is a well-established method for assessing DIF. A $2 \times 2 \times K$ table of test taker data is constructed based on item performance (right or wrong), group membership (the *focal group*, which is of primary interest, or the *reference group*), and score on an overall proficiency measure (with K levels). The MH (Mantel & Haenszel, 1959) odds ratio estimate is then used to compare the two groups in terms of their odds of answering the item correctly, conditional on the proficiency measure. The MH index of DIF, *MH D-DIF*, is obtained by multiplying the natural log of the MH odds ratio estimate $\hat{\alpha}_{MH}$ by -2.35 ; the transformation of $\hat{\alpha}_{MH}$ places *MH D-DIF* on the ETS delta scale of item difficulty (Holland & Thayer, 1985). By convention, *MH D-DIF* is defined so as to be negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. Phillips and Holland (1987) derived an estimated standard error for $\ln(\hat{\alpha}_{MH})$; their result proved to be identical to that of Robins, Breslow, and Greenland (1986).

The results of an MH DIF analysis typically include the *MH D-DIF* index, along with its estimated standard error. In making decisions about whether to discard items or flag them for review, however, testing companies may rely instead on categorical ratings of the severity of DIF. Several testing companies have adopted a system developed by ETS for categorizing the severity of DIF based on both the magnitude of the *MH D-DIF* index and the statistical significance of the results (see Zieky, 1993). According to this classification scheme, a "C" categorization, which represents moderate to large DIF, requires that the absolute value of *MH D-DIF* be at least 1.5 and be significantly greater than 1 (at $\alpha = .05$). A "B" categorization, which indicates slight to moderate DIF, requires that *MH D-DIF* be significantly different from zero (at $\alpha = .05$) and that the absolute value of *MH D-DIF* be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or the C categories are labeled "A" items, which are considered to have negligible DIF. Items that fall in the C category are subjected to further scrutiny and may be eliminated from tests. For most purposes, it is useful to distinguish between negative DIF (DIF against the focal group, by convention) and positive DIF (DIF against the reference group). This distinction yields a total of five DIF classifications: C-, B-, A, B+, and C+. We make use of this five-way categorization in our work, though we sometimes refer for convenience to the "A, B, and C categories."

¹ In Bayesian analysis, "posterior" means "following the collection of data."

Zwick, Thayer, and Lewis (1997, 1999, 2000) developed an empirical Bayes (EB) approach to DIF analysis and classification which yields more stable results in small samples than do conventional procedures and is therefore likely to be advantageous in adaptive testing conditions. An assumption is made about the prior distribution of DIF parameters across items. The prior is combined with the item's DIF results to estimate a posterior distribution; the posterior mean serves as the EB point estimate of the DIF parameter for that item.

In addition to offering an alternative point estimate of DIF, the EB method provides a version of the A, B, and C DIF classification system. Two related problems associated with the traditional classification approach are that (1) when sample sizes are small, the DIF category is unstable and may vary substantially from one test administration to another and (2) attaching an A, B, or C label to an item may convey the mistaken notion that an item's DIF category is deterministic. The EB approach yields an estimate of the *probability* that the true DIF for an item falls into the A, B, and C categories, based on an estimate of the posterior distribution of DIF parameters. The estimated A, B, and C probabilities can be regarded as representing our state of knowledge about the true DIF category for the item.

A possible advantage of the EB method of probabilistic DIF classification is that it may convey information about the sampling variability of DIF results in a more comprehensible way than do the current procedures. This alternative way of representing the variability of DIF findings lends itself well to graphical display. Pie charts can be used effectively to represent the posterior probabilities associated with the A, B, and C categories, as shown in the section, "Properties of EB Point Estimates." The EB methods can be modified easily if the current rules used to assign items to categories are adjusted or if other hypothesis-testing approaches are substituted for the Mantel-Haenszel procedure.

The EB approach to DIF analysis is related to three areas of previous research. A precursor to the method was developed in the context of a simulation study of DIF methods for computer adaptive tests conducted by Zwick, Thayer, and Wingersky (1994a, 1994b, 1995). Also, the variance component analysis of DIF developed by Longford and his colleagues (Longford, 1995, chapter 5; Longford, Holland, & Thayer, 1993) can be described as an EB approach. Finally, a Bayesian conceptualization of DIF was described by Holland in ETS internal documents (January 27, 1987; February 11, 1987).

Modification of the EB DIF methods for the LSAT CAT context required that we address several issues. First, we needed to take into account LSAC's interest in considering a CAT that was adaptive on the testlet level, rather than the item level; this required us to design a simulation that would involve testlet-based CAT administration. We then had to decide whether the matching of test takers for DIF analysis should be based on a score that took the testlet structure into account. Another determination we had to make was what set of items to use in estimating the prior distribution for the EB procedures, and whether our procedures for estimating the parameters of the prior, which had been developed for nonadaptive tests, needed modification for application to CATs. Finally, we had to determine whether the EB method, previously tested on samples no smaller than 200 test takers for the reference group and 50 for the focal group, could be applied successfully with even smaller samples. These issues are addressed in subsequent sections.

Statistical Model for the EB DIF Approach

The EB DIF method uses the observed values of *MH D-DIF* and *SE (MH D-DIF)*, along with an assumed prior distribution, to obtain the *posterior* distribution of true Mantel-Haenszel DIF parameters. The model can be expressed as follows. (The notation changes from *MH D-DIF* to *MH* and from *SE(MH D-DIF)* to *SE (MH)* to make the presentation less cumbersome.) We know that $\ln(\hat{\alpha}_{MH})$ has an asymptotic normal distribution (Agresti, 1990). Therefore, it is reasonable to assume that

$$MH_i | \theta_i \sim N(\theta_i, \sigma_i^2) \quad (1)$$

where MH_i is the MH statistic for item i , σ_i^2 is the sampling variance of the MH statistic, and $E(MH_i) = \theta_i$ represents the unknown parameter value corresponding to MH_i . In our computations, we assume that the sampling variance is known; that is, we set σ_i^2 equal to the observed estimate of the squared standard error, $SE^2(MH_i)$. The effect of ignoring the error associated with the estimation of $SE(MH_i)$ was judged to be minimal by Longford (1995); this was confirmed in analyses conducted by Zwick, Thayer, and Lewis (1997; 1999).

We assume the following prior distribution for θ_i :

$$\theta_i \sim N(\mu, \tau^2) \quad (2)$$

where μ is the across-item mean of the DIF parameters θ_i and τ^2 is the across-item variance. Estimation of μ and τ^2 is discussed in the section, "Estimation of μ and τ^2 ."

The posterior distribution of θ_i , given the observed statistic, MH_i , can be expressed as

$$f(\theta_i | MH_i) \propto f(MH_i | \theta_i) f(\theta_i) \quad (3)$$

Standard Bayesian calculations (see, e.g., Novick & Jackson, 1974) show that this distribution is normal with mean and variance given by

$$E(\theta_i | MH_i) = W_i MH_i + (1 - W_i) \mu \quad (4)$$

and

$$Var(\theta_i | MH_i) = W_i \sigma_i^2, \quad (5)$$

where

$$W_i = \frac{\tau^2}{\sigma_i^2 + \tau^2} \quad (6)$$

The means and variances for the model, prior, and posterior distributions are summarized in Table 1. The posterior mean is a *shrinkage* estimator of Mantel-Haenszel DIF, obtained by substituting estimates of μ and τ^2 for the corresponding parameters in equations 4–6 and setting σ_i^2 equal to $SE^2(MH_i)$. The larger the value of σ_i^2 , the more the EB estimation procedure “shrinks” the observed MH value toward the prior mean (often zero or close to zero, as described in the section, Estimation of μ and τ^2). On the other hand, as σ_i^2 approaches zero, W_i approaches 1, and the posterior mean approaches the observed MH_i value.

TABLE 1
Means and variances of key distributions

Distribution	Mean	Variance
Model $f(MH_i \theta_i)$	θ_i (unknown)	σ_i^2 (treated as known and equal to $SE^2(MH_i)$)
Prior $f(\theta_i)$	μ	τ^2
Posterior $f(\theta_i MH_i)$	$W_i MH_i + (1 - W_i) \mu$	$W_i \sigma_i^2$

Note. $W_i = \frac{\tau^2}{\sigma_i^2 + \tau^2}$

Obtaining the posterior probabilities associated with the five possible true DIF categories (C-, B-, A, B+, and C+) is accomplished by considering a normal distribution with mean and variance equal to the estimates of the posterior mean and variance (equations 4 and 5), respectively. The magnitude criteria presented in the section, EB Enhancement of Mantel-Haenszel DIF Analysis, are then applied. For example, to estimate the probability that the true DIF category is C-, the area under this normal density function, which is to the left of -1.5, is obtained. (Since the goal of the procedure is to estimate the distribution of DIF *parameters*, the statistical significance criteria for C- status, described in that section are not relevant here.)

In summary, the steps in the EB DIF procedure are as follows:

1. For the item of interest, estimate the values of MH_i and $SE(MH_i)$. Assume the distribution of MH_i is normal, with unknown mean θ_i and known standard deviation $SE(MH_i)$.
2. Assume that the *prior distribution* for the true DIF parameter, θ_i , is normal. Use the observed mean and variance of *MH D-DIF* statistics across an appropriately defined set of items (a test section or form in the case of paper-and-pencil tests), along with an estimate of the across-item average of $SE^2(MH_i)$, to estimate the parameters of the prior (see Equation 7). The prior distribution is the same for the entire set of items.

3. Based on a and b , use standard Bayesian theory to estimate, for each item, the *posterior distribution* of the true DIF parameter θ_i , given the observed statistics. The posterior mean is the EB point estimate of DIF.
4. By applying the magnitude criteria associated with the DIF classifications to the posterior distribution, estimate the probabilities that the true DIF for the item is in each of the five DIF categories.²

Estimation of μ and τ^2

An aspect of the EB procedure that needs further explanation is the determination of reasonable values for the prior mean μ and the prior variance τ^2 . When MH DIF analyses are conducted using number-right score as the matching variable, the MH DIF statistics are constrained to sum to approximately zero over the set of items. Therefore, in some applications, setting μ equal to zero may be appropriate. However, we have chosen to estimate μ as well as τ^2 from the current dataset; this less restrictive approach is appropriate under a wider range of circumstances, including analyses in which the matching variable is external to the test under investigation. Our estimates of μ and τ^2 were

$$\hat{\mu} = \text{Average}(MH_i) \quad \text{and} \quad \hat{\tau}^2 = \text{Var}(MH_i) - \text{Average}(SE_i^2(MH_i)) \quad (7)$$

where $\text{Var}(MH_i)$ is the observed across-item variance of the (MH_i) statistics.

In our initial work, we justified these estimators as follows: Suppose that $MH_i = \theta_i + e_i$ as in the simplest version of the model of Longford, Holland, and Thayer (1993), where e_i is an error term with $E(e_i | \theta_i) = 0$ and $\text{Var}(e_i | \theta_i) = \sigma_i^2$, $i = 1, 2, \dots, n$. Suppose further that $\text{Cov}(e_i, e_j | \theta_i, \theta_j) = 0$ for $i \neq j$. Then $E(MH_i) = E(\theta_i) = \mu$ and $\text{Var}(MH_i) = \tau^2 + \sigma_i^2$, where $\sigma^2 = E(\sigma_i^2)$. These estimators can still be justified in the case of adaptive testing, where it may be unreasonable to assume that the σ_i^2 values have a common expectation. As described by Hoaglin, Mosteller, and Tukey (1991, p. 205), we can define σ^2 as $\sum_i \sigma_i^2 / n$ and estimate σ^2 as in the equal-variance case without losing much precision. We use the observed mean of the MH_i statistics as our estimate of μ . In estimating τ^2 , we use $\text{Average}(SE^2(MH_i))$ to estimate σ^2 and use the observed across-item variance of the MH_i statistics as an estimate of $\text{Var}(MH_i)$. That is, in the formulation in Equation 7, the across-item prior variance of the true DIF values, τ^2 is estimated by deflating the estimated across-item variance of the DIF statistics by an amount equal to the average of the estimated item-level sampling variances. Similar estimators have been independently proposed by Camilli and his colleagues (e.g., Camilli & Penfield, 1997).

In the present study, we had to consider how to adapt our procedures for estimating μ and τ^2 for the CAT context. One determination we had to make was what set of items to use in estimating the prior distribution for the EB procedures; this is not entirely clear in the context of a testlet-based CAT. After considering several alternatives, we decided that the best way of preserving the advantages of the EB analysis without introducing unwieldy computational procedures was to estimate the parameters of the prior using data from all items in the pool.

In addition to deciding what set of items should be used in estimating the prior parameters, we needed to determine whether our former procedure for estimating τ^2 would perform well in the CAT context, where MH standard errors can vary considerably across items. In an unpublished simulation study, Dorothy Thayer and Charles Lewis compared our usual estimate of τ^2 (from Equation 7) to the iteratively obtained estimate of Longford (Longford, 1995; Longford, Holland, & Thayer, 1993) and to the approach developed by Camilli and his colleagues (see Camilli & Penfield, 1997). The simulation results showed that the seemingly unsophisticated estimate of Equation 7 performed best in a variety of circumstances. Estimation of μ and τ^2 in the current study is discussed further in the section, Comparison of $\hat{\mu}$ and $\hat{\tau}^2$ to Their Theoretical Values.

Validity Evidence

Zwick, Thayer, and Lewis (1997) conducted extensive validity studies of the EB DIF procedures, only a few of which are described here. Using simulated data, root mean square residuals (*RMSRs*) were computed to measure the deviation between DIF statistics and the true (generating) DIF, defined in the section, *True DIF values and True DIF categories for Simulation Items*. As expected, the EB point estimates had smaller *RMSRs* than did the ordinary *MH D-DIF* statistics; this advantage was greater in small samples. (See Casella, 1985, for a good intuitive explanation of the stability of EB estimates.) Application to actual test taker data for two administrations of the same test form showed that the Time 1 EB estimates were better predictors of the Time 2 MH_i statistics (i.e., had smaller *RMSRs*) than were the Time 1 MH_i statistics. Calibration plots

² Using computations only slightly more complex than those above, the EB approach can also be used to estimate the probability that an item will be classified as an A, B, or C in future administrations, based on the posterior predictive distribution (see Zwick, Thayer, & Lewis, 1997, 1999). The EB estimation procedures can also serve as the basis for a DIF detection procedure that uses loss functions (see Zwick, Thayer, & Lewis, 2000).

