Research Report

Extension of the NAEP BGROUP Program to Higher Dimensions

Sandip Sinharay

Matthias von Davier

Research & Development

December 2005
RR-05-27

Extension of the NAEP BGROUP Program to Higher Dimensions

Sandip Sinharay and Matthias von Davier
ETS, Princeton, NJ

December 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

The reporting methods used in large scale assessments such as the National Assessment of Educational Progress (NAEP) rely on a *latent regression model*. The first component of the model consists of a p-scale IRT measurement model that defines the response probabilities on a set of cognitive items in p scales depending on a p-dimensional latent trait variable $\theta = (\theta_1, \ldots \theta_p)$. In the second component, the conditional distribution of this latent trait variable θ is modeled by a multivariate, multiple linear regression on a set of predictor variables, which are usually based on student, school and teacher variables in assessments such as NAEP.

To fit the latent regression model using the maximum (marginal) likelihood estimation technique, multivariate integrals have to be evaluated. In the computer program MGROUP used by ETS for fitting the latent regression model to data from NAEP and other programs, the integration is currently done either by numerical quadrature for problems up to two dimensions or by an approximation of the integral. CGROUP, the current operational version of the MGROUP program used in NAEP and other assessments since 1993, is based on Laplace approximation, which may not provide fully satisfactory results, especially if the number of items per scale is small (see, e.g., Thomas, 1993a, or von Davier & Sinharay, 2004). There is scope for improvement in the technique used.

This paper extends the NAEP BGROUP program to higher dimensions. Two real data analyses, one with a medium-sized data set and another with a large data set, show that the extension promises to be useful for fitting the NAEP model.

Key words: CGROUP, latent regression, MGROUP, NAEP

Acknowledgements

The authors thank Dan Eignor, Andreas Oranje, Amy Hauger, John Mazzeo, Shelby Haberman, and Neal Thomas for useful advice and Kim Fryer for her help with proofreading.

1. Introduction

National Assessment of Educational Progress (NAEP), the only regularly administered and congressionally mandated national assessment program (see, e.g., Beaton & Zwick, 1992), is an ongoing survey of the academic achievement of students in the United States in a number of subject areas such as reading, writing, and mathematics. For several reasons (e.g., von Davier & Sinharay, 2004; Mislevy, Johnson, & Muraki, 1992), NAEP reporting methods started using in 1984 a multilevel statistical model consisting of two components: (a) an item response theory (IRT) component at the first level and (b) a linear regression regression component at the second level (see, e.g., Beaton, 1987; Mislevy et al., 1992). Other large scale educational assessments such as the International Adult Literacy Study (IALS; Kirsch, 2001), Trends in Mathematics and Science Study (TIMSS; Martin & Kelly, 1996), and Progress in International Reading Literacy Study (PIRLS; Mullis, Martin, Gonzalez, & Kennedy, 2003) also adopted essentially the same model.

This model is often referred to as a *latent regression model*. An algorithm for estimating the parameters of this model is implemented in the MGROUP set of programs, which is an ETS product. MGROUP computes the maximum likelihood estimates of the parameters of the model using a version of the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) suggested by Mislevy (1984, 1985). The algorithm requires the values of the posterior mean and the posterior standard deviation (SD) of the proficiency variable $\boldsymbol{\theta}$ for each examinee, computation of which involves integration with respect to the multivariate $\boldsymbol{\theta}$. For problems up to two dimensions (subscales), the integration is computed using numerical quadrature implemented in the BGROUP version (Beaton, 1987) of the MGROUP program. For higher dimensions, no numerical integration routine is available and an approximation of the integral is used. The CGROUP version of MGROUP, the current operational procedure used in NAEP and other assessments for tests with more than two dimensions, is based on the Laplace approximation (Kass & Steffey, 1989) that ignores the higher-order derivatives of the examinee posterior distribution and may not provide accurate results, especially for higher dimensions. For example, a graphical plot for a data example in Thomas (1993a) shows that CGROUP overestimates the high examinee posterior variances for an assessment with two subscales (dimensions).

Similar results have been found in von Davier and Sinharay (2004). Further, it is not even known how accurate CGROUP results are for more than two dimensions, where the Laplace approximation may result in considerably inaccurate results. Under these circumstances, an operational program that can perform numerical quadrature for more than two dimensions and hence does not require any approximations may be of great help.

This paper examines a successful extension of the BGROUP version of MGROUP to more than two dimensions. Two real data examples, one with a medium-sized data set and another with a large data set, show that the results produced by the extension of BGROUP are often different from those produced by CGROUP.

Section 2 describes the current NAEP model and estimation procedure; included is a detailed description of the BGROUP procedure. Section 3 discusses the results from application of the extension of the BGROUP to two real data examples. Section 4 discusses the conclusions and future work.

2. The NAEP Statistical Model and Estimation Method

2.1 The Latent Regression Model

NAEP employs a latent regression model utilizing an IRT measurement model. Assume that the unique p-dimensional latent proficiency vector for examinee i is $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots \theta_{ip})'$. In operational NAEP assessments, p could be any integer between 1 and 5.

Let us denote the response vector to the test items for examinee i as $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{ip})$, where, $\boldsymbol{y}_{ik}$, a vector of responses, contributes information about θ_{ik}. The likelihood for an examinee is given by

$$l(\boldsymbol{\theta}_i) = \prod_{q=1}^{p} l_q(\boldsymbol{y}_{iq}|\theta_{iq}) \cdot \tag{1}$$

Each quantity $l_q(\boldsymbol{y}_{iq}|\theta_{iq})$ above is given by products of terms from a univariate IRT model; usually the terms are from the three-parameter logistic (3PL) model or generalized partial credit model (GPCM). For example, if the items measuring θ_{iq}s are all multiple-choice

items, $l_q(\boldsymbol{y}_{iq}|\theta_{iq})$ is given by

$$l_q(\boldsymbol{y}_{iq}|\theta_{iq}) = \prod_j p_{iqj}^{\boldsymbol{y}_{iqj}}(1-p_{iqj})^{1-\boldsymbol{y}_{iqj}},$$

where $p_{iqj} = c_{jq} + (1-c_{jq})(1+\mathrm{e}^{a_{jq}(b_{jq}-\theta_{iq})})^{-1}$, and $\boldsymbol{y}_{iq} = (\boldsymbol{y}_{iq1}, \boldsymbol{y}_{iq2}, \ldots, \boldsymbol{y}_{iqJ})$. For reasons to be discussed later, the dependence of (1) on the item parameters is suppressed.

Suppose $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots x_{im})$ are m fully measured demographic and educational characteristics corresponding to the examinee. Conditional on $\boldsymbol{x}_i$, the examinee proficiency vector $\boldsymbol{\theta}_i$ is assumed to follow a multivariate normal prior distribution, that is, $\boldsymbol{\theta}_i|\boldsymbol{x}_i \sim N(\boldsymbol{\Gamma}'\boldsymbol{x}_i, \boldsymbol{\Sigma})$. The mean parameter matrix $\boldsymbol{\Gamma}$ and the nonnegative definite variance matrix $\boldsymbol{\Sigma}$ are assumed to be the same for all examinee groups.

Under this setup, $L(\boldsymbol{\Gamma}, \boldsymbol{\Sigma}|\boldsymbol{X}, \boldsymbol{Y})$, the (marginal) likelihood function for $(\boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ based on the data $(\boldsymbol{X}, \boldsymbol{Y})$, is given by

$$L(\boldsymbol{\Gamma}, \boldsymbol{\Sigma}|\boldsymbol{X}, \boldsymbol{Y}) = \prod_{i=1}^{n} \int l_1(\boldsymbol{y}_{i1}|\theta_{i1}) \ldots l_p(\boldsymbol{y}_{ip}|\theta_{ip}) \phi(\boldsymbol{\theta}_i|\boldsymbol{\Gamma}'\boldsymbol{x}_i, \boldsymbol{\Sigma}) d\boldsymbol{\theta}_i, \tag{2}$$

where n is the number of examinees, and $\phi(.|.,.)$ is the multivariate normal density function.

2.2 NAEP Estimation Process and the MGROUP Program

NAEP uses a three-stage estimation process for fitting the above mentioned latent regression model and making inferences. The first stage, *scaling*, fits a simple IRT model (3PL model for multiple-choice items and the GPCM for constructed-response items) to the examinee response data and estimates the item parameters. The prior distribution used in this step is not $\boldsymbol{\theta}_i|\boldsymbol{x}_i \sim N(\boldsymbol{\Gamma}'\boldsymbol{x}_i, \boldsymbol{\Sigma})$ as described above, but is a discrete distribution over 41 quadrature points for each component of $\boldsymbol{\theta}$ so that the probabilities at the 41 points are estimated from the data; also the subscales are assumed to be independent a priori. The second stage, *conditioning*, assumes that the item parameters are known and equal to the estimates found in *scaling* and fits the model in (2) to the data (i.e., estimates $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ as a first part). In the second part of the *conditioning* step, *plausible values* for all examinees are obtained using the parameter estimates obtained in *scaling* and the first part of *conditioning*—the plausible values are used to estimate examinee subgroup averages. The third stage of the NAEP estimation process, called *variance estimation*, estimates the

variances corresponding to the examinee subgroup averages using a jackknife approach (see, e.g., Johnson & Jenkins, 2004). Our research will focus on the *conditioning* step and assume that the *scaling* has already been done (i.e., the item parameters are fixed); this is the reason we suppress the dependence of (1) on the item parameters.

Because we will be concerned with the *conditioning* step, the remaining part of the section provides a more detailed discussion of it. The first objective of this step is to estimate $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ from the data. If the $\boldsymbol{\theta}_i$s were known, the maximum likelihood estimators of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ would be

$$\hat{\mathbf{\Gamma}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_n\right)', \tag{3}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_i (\boldsymbol{\theta}_i - \mathbf{\Gamma}'\boldsymbol{x}_i)(\boldsymbol{\theta}_i - \mathbf{\Gamma}'\boldsymbol{x}_i)'\cdot \tag{4}$$

However, $\boldsymbol{\theta}_i$s are actually unknown. Mislevy (1984, 1985) shows that the maximum likelihood estimates of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ under unknown $\boldsymbol{\theta}_i$s can be obtained using an EM algorithm (Dempster et al., 1977). The EM algorithm iterates through a number of expectation steps (E-step) and maximization steps (M-step). The expression for $(\mathbf{\Gamma}_{t+1}, \mathbf{\Sigma}_{t+1})$, the updated value of the parameters in the t^{th} M-step, is obtained as:

$$\mathbf{\Gamma}_{t+1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left(\widetilde{\boldsymbol{\theta}_{1t}}, \widetilde{\boldsymbol{\theta}_{2t}}, \ldots \widetilde{\boldsymbol{\theta}_{nt}}\right)', \tag{5}$$

$$\mathbf{\Sigma}_{t+1} = \frac{1}{n}\left[\sum_i \mathrm{Var}(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t) + \sum_i (\widetilde{\boldsymbol{\theta}_{it}} - \mathbf{\Gamma}'_{t+1}\boldsymbol{x}_i)(\widetilde{\boldsymbol{\theta}_{it}} - \mathbf{\Gamma}'_{t+1}\boldsymbol{x}_i)'\right], \tag{6}$$

where $\widetilde{\boldsymbol{\theta}_{it}} = E(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t)$ is the posterior mean for examinee i given the preliminary parameter estimates of iteration t. The process is repeated until convergence of the estimates $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$.

Equations (5) and (6) require the values of the posterior means $E(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t)$ and the posterior variances $\mathrm{Var}(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t)$ for the examinees, which are given by

$$E(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t) \equiv \int \boldsymbol{\theta}_i g(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t) d\boldsymbol{\theta}_i, \text{ and} \tag{7}$$

$$\mathrm{Var}(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t) \equiv \int (\boldsymbol{\theta}_i - E(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t))(\boldsymbol{\theta}_i - E(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t))' \, g(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t) d\boldsymbol{\theta}_i, \tag{8}$$

where the examinee posterior distribution $g(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Gamma}_t, \boldsymbol{\Sigma}_t)$ is given by

$$g(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Gamma}_t, \boldsymbol{\Sigma}_t) \propto l_1(\boldsymbol{y}_{i1}|\theta_{i1}) \ldots l_p(\boldsymbol{y}_{ip}|\theta_{ip})\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}_t'\boldsymbol{x}_i, \boldsymbol{\Sigma}_t) \tag{9}$$

using (2). The proportionality constant in (9) is a function of $\boldsymbol{y}_i$, $\boldsymbol{\Gamma}_t$, and $\boldsymbol{\Sigma}_t$.

Correspondingly, the t^{th} E-step computes the two required quantities for all the examinees. The MGROUP set of programs at ETS perform the above mentioned EM algorithm.

The MGROUP program consists of two primary controlling routines called PHASE1 and PHASE2. The former does some preliminary processing while the latter directs the EM iterations. There are different versions of the MGROUP program depending on the method used to perform the E-step in PHASE2: NGROUP (Beaton, 1988) using Bayesian normal theory, BGROUP (which is used when the dimension of $\boldsymbol{\theta}_i$ is up to two) using numerical quadrature, CGROUP (Thomas, 1993a) using Laplace approximations, and Y-group (von Davier & Yu, 2003) using seemingly unrelated regression (SUR; Zellner, 1962).

2.3 The Limitations of the Current Estimation Method

The BGROUP version of MGROUP program is the gold standard in MGROUP. However, Thomas (1993b) mentioned that numerical quadrature is computationally unfeasible for applications with more than two subscales. When the dimension of $\boldsymbol{\theta}_i$ is larger than two, CGROUP is the most appropriate and used operationally in NAEP. This approach uses the Laplace approximation, which involves a Taylor-series expansion of an integrand while ignoring higher-order derivatives of examinee posterior distributions, of the posterior mean and variance. Details about the method can be found in Thomas (1993b, pp. 316-317). The Laplace method does not provide an unbiased estimate of the quantity it is approximating and may provide inaccurate results if higher order derivatives of the examinee posterior distributions (that the Laplace method assumes to be equal to zero) are not negligible. The error of approximation for each component of the mean and covariance of $\boldsymbol{\theta}_i$ is of order $O(\frac{1}{k^2})$ (e.g., Kass & Steffey, 1989), where k is the number of items measuring skill corresponding to the component. Because the number of items given to each examinee in large scale assessments such as NAEP is not too large (making k rather small), the error

in the Laplace approximation may become nonnegligible, especially for high-dimensional $\boldsymbol{\theta}_i$s. Further, if the posterior distribution of $\boldsymbol{\theta}_i$s is multimodal (which is not impossible, especially for a small number of items), the method can perform poorly. Therefore the CGROUP version of MGROUP is not entirely satisfactory. Figure 1 in Thomas (1993b), where the posterior variance estimates of 500 randomly selected examinees using BGROUP and CGROUP for two-dimensional $\boldsymbol{\theta}_i$ are plotted, shows that the CGROUP provides inflated variance estimates for examinees with large posterior variance (von Davier & Sinharay, 2004, observed a similar phenomenon). The departure may be more severe for $\boldsymbol{\theta}_i$s in higher dimensions. Thus, the current NAEP estimation methods leave room for improvement.

3. Extending BGROUP to Higher Dimensions

This section begins with a description of how the current NAEP BGROUP program calculates the posterior means and variances in the E-step of the EM algorithm and then proceeds to describe our extension of the BGROUP program.

3.1 Currently Used BGROUP E-step

Consider an assessment with one subscale ($p = 1$). Using notations introduced in (1), let the likelihood term for an examinee with proficiency θ be $l(\theta)$. The quadrature implemented for $p = 1$ evaluates the examinee posterior on a grid of m points $q_1, q_2, \ldots q_m$ on the θ-scale, computes the expectation of θ^k (for a scalar variance $\boldsymbol{\Sigma}_t$, mean vector $\boldsymbol{\Gamma}_t$, and background information vector $\boldsymbol{x}$) as

$$E(\theta^k|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Gamma}_t, \boldsymbol{\Sigma}_t) \approx \frac{\sum_{i=1}^{m} q_i^k l(q_i)\phi(q_i|\boldsymbol{\Gamma}_t'\boldsymbol{x}, \boldsymbol{\Sigma}_t)}{\sum_{i=1}^{m} l(q_i)\phi(q_i|\boldsymbol{\Gamma}_t'\boldsymbol{x}, \boldsymbol{\Sigma}_t)}. \tag{10}$$

The denominator in (10) estimates the normalizing constant of the examinee posterior density. This approach is described in Beaton (1987).

The quadrature implemented for p=2 evaluates the examinee posterior on a grid of $m \times m$ points, formed by $q_{11}, q_{12}, \ldots q_{1m}$ on the θ_1-scale and $q_{21}, q_{22}, \ldots q_{2m}$ on the θ_2-scale, and computes the expectation of $\theta_1^k\theta_2^l$ (for a variance matrix $\boldsymbol{\Sigma}_t$, mean matrix $\boldsymbol{\Gamma}_t$, and

background information vector $\boldsymbol{x}$) as

$$E(\theta_1^k\theta_2^l|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Gamma}_t,\boldsymbol{\Sigma}_t) \approx \frac{\sum_{i=1}^m\sum_{j=1}^m q_{1i}^k q_{2j}^l l(q_{1i},q_{2j})\phi(q_{1i},q_{2j}|\boldsymbol{\Gamma}_t'\boldsymbol{x},\boldsymbol{\Sigma}_t)}{\sum_{i=1}^m\sum_{j=1}^m l(q_{1i},q_{2j})\phi(q_{1i},q_{2j}|\boldsymbol{\Gamma}_t'\boldsymbol{x},\boldsymbol{\Sigma}_t)}. \tag{11}$$

Thomas (1993b) mentioned that numerical quadrature is computationally unfeasible for applications with more than two subscales. However, with the recent advance in speed of computing, it is possible to apply the same process to more than two dimensions.

3.2 Details of Our Implementation

The quadrature implemented for p dimensions evaluates the examinee posterior on a grid of m^p points, formed by $q_{11}, q_{12}, \ldots q_{1m}$ on the θ_1-scale, $q_{21}, q_{22}, \ldots q_{2m}$ on the θ_2-scale, $\ldots$ $q_{p1}, q_{p2}, \ldots q_{pm}$ on the θ_p-scale, and approximates (for a variance matrix $\boldsymbol{\Sigma}_t$, mean matrix $\boldsymbol{\Gamma}_t$ and background information vector $\boldsymbol{x}$) $E(\theta_1^{k_1}\theta_2^{k_2}\ldots\theta_p^{k_p}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Gamma}_t,\boldsymbol{\Sigma}_t)$ as

$$\frac{\sum_{i_1=1}^m\ldots\sum_{i_p=1}^m q_{1i_1}^{k_1}\ldots q_{pi_p}^{k_p} l(q_{1i_1},\ldots q_{pi_p})\phi(q_{1i_1},\ldots,q_{pi_p}|\boldsymbol{\Gamma}_t'\boldsymbol{x},\boldsymbol{\Sigma}_t)}{\sum_{i_1=1}^m\ldots\sum_{i_p=1}^m l(q_{1i_1},\ldots q_{pi_p})\phi(q_{1i_1},\ldots,q_{pi_p}|\boldsymbol{\Gamma}_t'\boldsymbol{x},\boldsymbol{\Sigma}_t)}. \tag{12}$$

This brute-force approach is computationally costly and results in a long runtime for a multivariate problem. Even in the case of a three-dimensional problem, the number of grid points on which the computation of the posterior distribution for each examinee is required runs as large as $41^3 = 68921$ (because of the use of 41 quadrature points per dimension). In higher dimensional problems such as the NAEP math assessment, the number of subscales is up to 5, resulting in far more points to evaluate than would be feasible when using a brute-force approach.

To overcome this problem, our implementation includes a more efficient approach that makes use of the fact that the likelihood is factored (Thomas, 1993a) in latent regressions assuming simple structure for the measurement model. In that case, using (1),

$$l(q_{1i_1}, q_{2i_2}, \ldots q_{pi_p}) = \prod_k l_k(qki_k),$$

where $l_k()$ denotes the one-dimensional likelihood function corresponding to dimension k. This form of the likelihood yields values that are numerically indistinguishable from zero if at least one of the product terms vanishes. This means that all grid coordinates for which at

least one $l_k(qki_k)$ is zero may be ignored. Finally, using (12), the following approximation of $E(\theta_1^{k_1}\theta_2^{k_2}\ldots\theta_p^{k_p}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Gamma}_t,\boldsymbol{\Sigma}_t)$ is

$$\frac{\sum_{l_1(q_{1i_1})\ldots l_p(q_{pi_p})\neq 0} q_{1i_1}^{k_1}\ldots q_{pi_p}^{k_p} l_1(q_{1i_1})\ldots l_p(q_{pi_p})\phi(q_{1i_1},\ldots,q_{pi_p}|\boldsymbol{\Gamma}_t'\boldsymbol{x},\boldsymbol{\Sigma}_t)}{\sum_{l_1(q_{1i_1})\ldots l_p(q_{pi_p})\neq 0} l(q_{1i_1},\ldots q_{pi_p})\phi(q_{1i_1},\ldots,q_{pi_p}|\boldsymbol{\Gamma}_t'\boldsymbol{x},\boldsymbol{\Sigma}_t)}. \tag{13}$$

Further optimization is possible if a similar approach is taken to inform the algorithm whether grid coordinates need to be evaluated based on the prior

$$\phi(q_{1i_1},\ldots,q_{pi_p}|\boldsymbol{\Gamma}_t'\boldsymbol{x},\boldsymbol{\Sigma}_t)<\epsilon,$$

which results in additional gains in speed. This is true particularly for high-dimensional problems that contain very highly correlated dimensions such as the NAEP math assessment. In these cases, grid coordinates that contain very dissimilar q_{ki_k} are associated with very low prior density values, which leads to values that are ignorable in the integral evaluation. This additional optimization needs some more computation per grid coordinate, but may save a lot of floating point multiplications in cases where dimensions are highly correlated or the prior variances are small compared to the range of the integration intervals.

Comparisons of estimates obtained using the unoptimized brute-force version and the first-level optimization using the above mentioned vanishing marginal likelihood rule showed no noticeable differences for small three-dimensional test cases. Comparisons of the results obtained using the unoptimized brute-force version and the second-level optimization using the vanishing prior density showed very small differences in posterior means and resulted in no noticeable differences in regression estimates or group statistics.

Therefore, the p-dimensional runs were carried out using the second-level optimizations in order to cut down the time needed to evaluate the 41^p integral by not evaluating ignorable terms of the integral.

3.3 Results for the 2002 NAEP Reading Assessment at Grade 12

We ran our program extending BGROUP on small test data sets and compared the results with those obtained from the operational programs. For test data sets with 1 and 2 subscales (i.e., $p = 1, 2$), the results from our program matched those from the operational versions (BGROUP and CGROUP) very closely.

Next, we applied our program to data from the 2002 NAEP reading assessment at grade 12 (see, for example, http://nces.ed.gov/nationsreportcard/reading/results2002). Each of 14,724 students was asked either two 25-minute blocks of questions or one 50-minute block of questions; each block contains at least one passage and related set of approximately 10 to 12 comprehension questions (combination of four-option multiple-choice and constructed response). Three subskills of reading are assessed: (a) reading for literary experience, (b) reading for information, and (c) reading to perform a task. Thus, this is an example where our work may be beneficial because for such three-subscale assessments, CGROUP is the only currently available version of MGROUP for operational analysis.

On a PC with a 2.2 GHZ Pentium 4 processor with 512 MB RAM running Linux, the program takes approximately 36 hours to converge. To reduce run time, we also ran the extended BGROUP program using the CGROUP estimates as starting values—it took about 12 hours. Results are practically indistinguishable whether we use the CGROUP estimates as starting values or not.

We also ran the extended BGROUP using 101 quadrature points per dimension—the error of approximation of the numerical quadrature formula (13) is expected to be almost negligible for such a large number of quadrature points—thus, the estimates from this run provide the gold standard that both CGROUP and the extended BGROUP with 41 quadrature points per dimension attempt to approximate. The estimates from extended BGROUP with 41 quadrature points per dimension are very close to the gold standard (results not shown)—this provides proof that the extended BGROUP program with 41 points performs adequately. The following discussion compares results from the CGROUP and the extended BGROUP (with 41 quadrature points per dimension).

Figure 1 compares the estimated regression coefficients (i.e., estimates of components of $\mathbf{\Gamma}$) for CGROUP and extended BGROUP for the three subskills. The differences between the two methods are negligible; the maximum difference is in the third decimal place.

Table 1 shows the residual variance estimates $\hat{\mathbf{\Sigma}}$ from extended BGROUP and, for convenience, the difference in these estimates from CGROUP and extended BGROUP. The differences of results produced by the two methods are negligible. However, all the

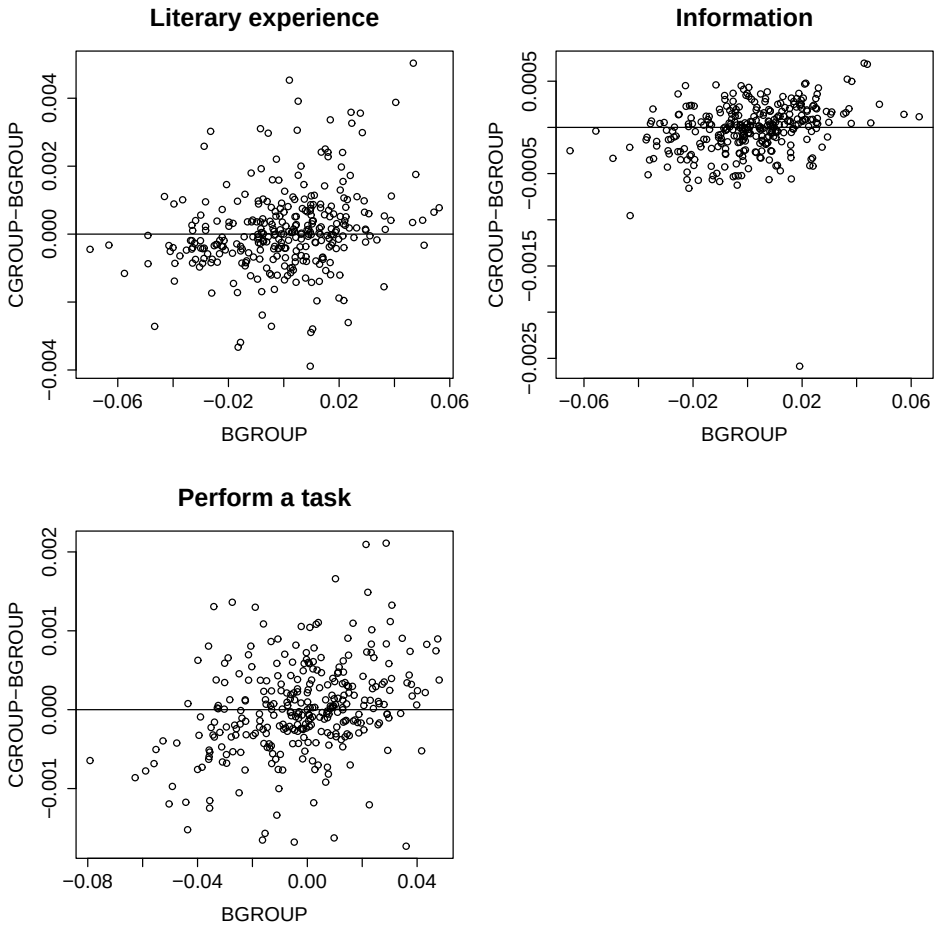

***Figure* 1. Comparison of regression coefficients from CGROUP and extended BGROUP for the 2002 NAEP reading assessment at grade 12.**

three variance component estimates and the three covariance estimates are slightly lower for BGROUP than the corresponding CGROUP estimates.

Figures 2 and 3 compare the marginal posterior means and standard deviations (SDs) of 1,000 randomly chosen examinees for CGROUP and extended BGROUP. Both figures have a plot for each subscale.

Figure 2 also shows the differences roughly in the scale reported by NAEP. In operational NAEP, a weighted average of the scores in the three subscales (with weights 0.35, 0.45, and 0.20 for literacy, information, and perform, respectively) is reported. NAEP uses a complicated linking procedure involving data for grades 4, 8, and 12—this usually

Table 1.

Residual Variances, Covariances, and Correlations for the 2002 NAEP Reading Assessment at Grade 12

	BGROUP			CGROUP-BGROUP		
	Literary	Information	Perform	Literary	Information	Perform
Literary	0.448	0.365	0.333	0.008	0.010	0.006
Information	0.784	0.483	0.356	0.007	0.008	0.008
Perform	0.712	0.733	0.488	-0.004	-0.001	0.014

Note. Residual variances are shown on main diagonals, covariances on upper off-diagonals, and correlations on lower off-diagonals.

converts the composite score to a scale with mean of approximately 300 and an SD of approximately 35. For the 2002 NAEP reading assessment at grade 12, the reported mean of the composite was 287 and the SD of the composite was 35. We do not use the rigorous NAEP linking procedure here, but instead use a simpler alternative; we compute the weighted average of the posterior means in the three subscales for each examinee using the weights 0.35, 0.45, and 0.20, as used in NAEP. Then we apply a linear transformation of the resulting weighted average to convert it to a scale with a mean of 287 and an SD of 35.

Results produced by the CGROUP version are mostly close to those produced by the extended BGROUP version. The CGROUP routine has a tendency to overestimate high posterior means and underestimate low posterior means (the extent of underestimation being more severe, especially for a few examinees). The CGROUP routine slightly overestimates the extreme posterior SDs, a phenomenon that was observed by Thomas (1993a) and von Davier and Sinharay (2004). The lowest point in all of the plots in Figure 2 belongs to the same examinee; the same is true for the next three lowest points in all the plots. The lowest three points in all the plots in Figure 3 belong to three examinees who are not outliers in Figure 2.

Table 2 compares the subgroup means and SDs (in parentheses) from extended BGROUP and the difference in these values from CGROUP and extended BGROUP—there

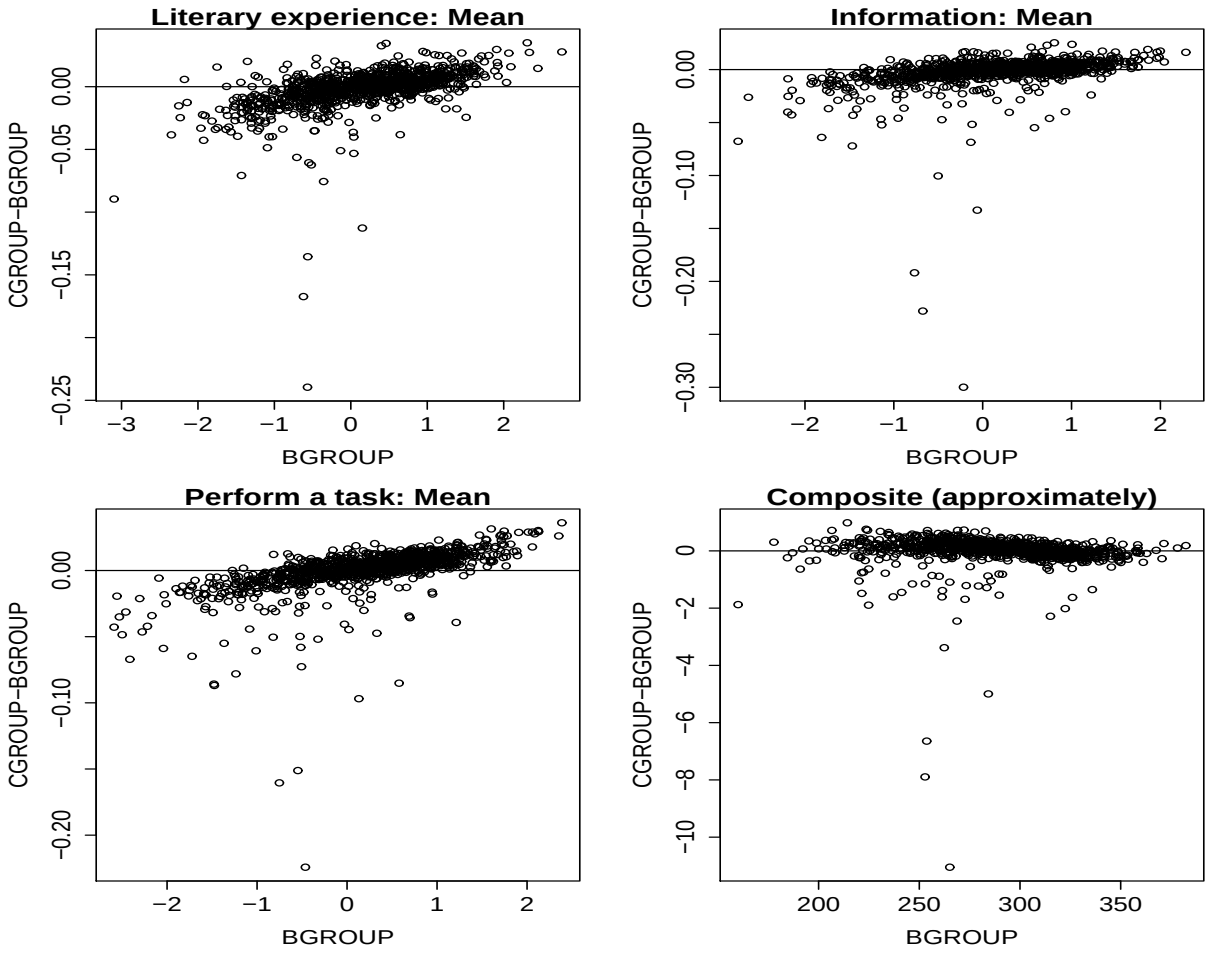

***Figure* 2. Comparison of posterior means from CGROUP and extended BGROUP for the 2002 NAEP reading assessment at grade 12.**

seems to be little difference between the two methods from this aspect; however, the BGROUP means are larger than or equal to the CGROUP means except for one entry. This supports Figure 2, where the extent of underestimation of low posterior means by CGROUP is larger than the extent of overestimation of high posterior means. The BGROUP SDs are all slightly less than the CGROUP SDs, which is consistent with Figure 3.

3.4 Results for the 2002 NAEP Reading Assessment at Grade 8

The extended BGROUP program gave acceptable results for a moderately large data set, so we applied it to an assessment with larger sample size, specif-

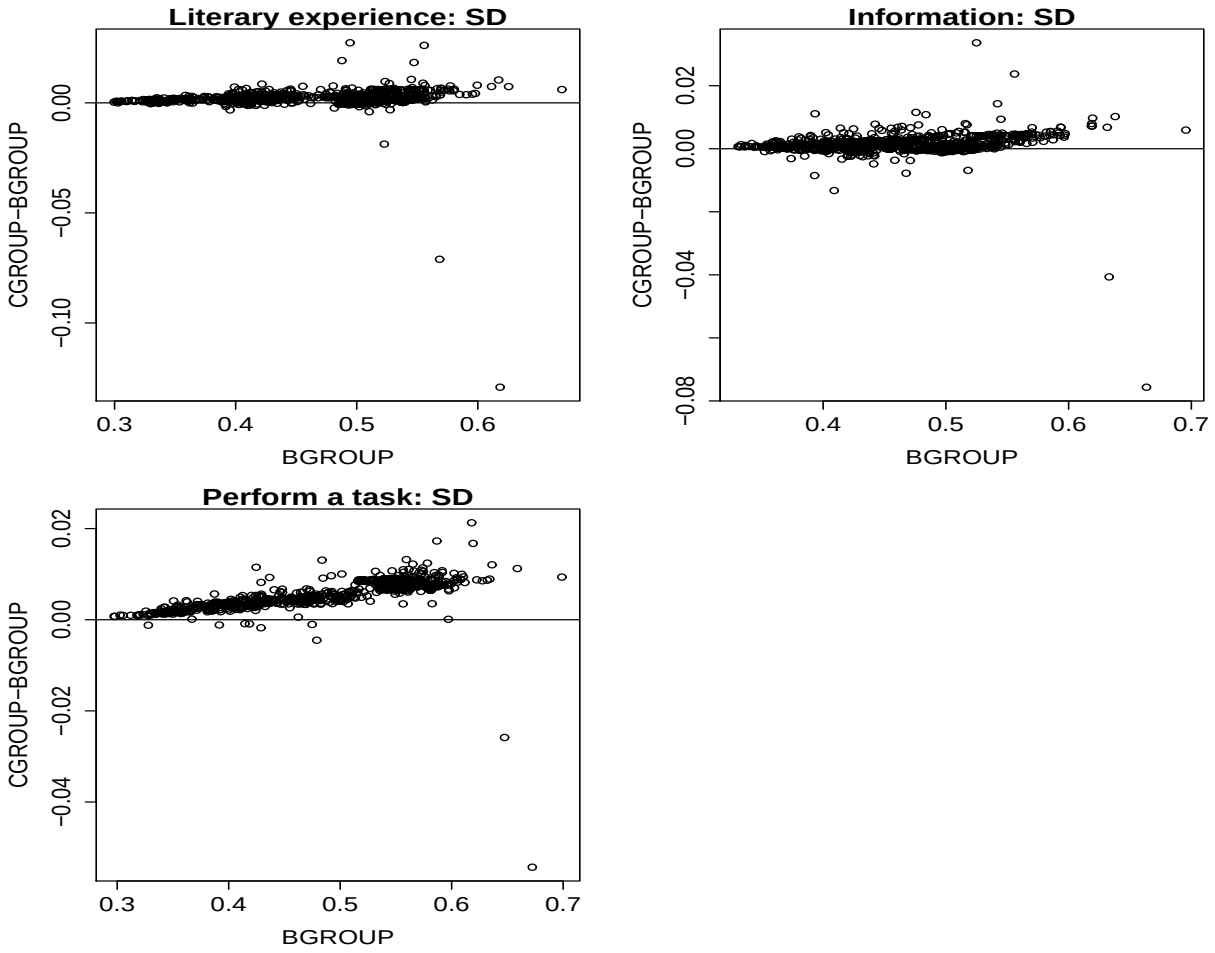

***Figure* 3. Comparison of posterior SDs from CGROUP and extended BGROUP for the 2002 NAEP reading assessment at grade 12.**

ically, to data from the 2002 NAEP reading assessment at grade 8 (see, e.g., http://nces.ed.gov/nationsreportcard/reading/results2002). Altogether, about 115,000 students took the test, which was similar in structure to the 2002 NAEP reading assessment at grade 12.

We ran the extended BGROUP program with CGROUP estimates as starting values—the program took approximately 48 hours (94 iterations of the EM algorithm) to converge on a PC with a 2.2 GHZ Pentium 4 processor and 512 MB RAM running Linux.

Figure 4 compares the regression coefficients for CGROUP and extended BGROUP for the three subskills. The differences between the two methods are negligible; the

Table 2.

Comparison of Subgroup Estimates From Extended BGROUP and CGROUP for the 2002 NAEP Reading Assessment at Grade 12

	BGROUP			CGROUP-BGROUP		
Subgroup	Literary	Information	Perform	Literary	Information	Perform
Overall	0.015	0.027	0.019	-0.003	-0.002	-0.002
	(0.959)	(0.950)	(1.003)	(0.010)	(0.007)	(0.013)
Male	-0.174	-0.152	-0.239	-0.004	-0.003	-0.004
	(0.946)	(0.961)	(0.991)	(0.011)	(0.008)	(0.013)
Female	0.198	0.200	0.268	-0.001	-0.001	0.002
	(0.935)	(0.907)	(0.951)	(0.009)	(0.006)	(0.011)
White	0.195	0.198	0.174	-0.001	-0.001	0.000
	(0.910)	(0.903)	(0.961)	(0.008)	(0.006)	(0.011)
Black	-0.516	-0.429	-0.453	-0.007	-0.004	-0.005
	(0.890)	(0.888)	(0.940)	(0.013)	(0.007)	(0.013)
Hispanic	-0.387	-0.357	-0.262	-0.008	-0.006	-0.005
	(0.971)	(0.986)	(1.051)	(0.016)	(0.010)	(0.014)
Asian	0.023	-0.013	-0.052	-0.005	-0.003	-0.003
	(0.907)	(0.910)	(1.010)	(0.010)	(0.007)	(0.013)
American	-0.094	-0.276	-0.358	-0.002	-0.004	-0.008
Indian	(0.960)	(0.977)	(1.027)	(0.010)	(0.010)	(0.015)

maximum difference is in the third decimal place.

Table 3 shows the residual variance estimates $\hat{\boldsymbol{\Sigma}}$ from extended BGROUP and the difference between CGROUP and extended BGROUP. The difference between the two methods is negligible. However, all of the three variance component estimates and the three covariance estimates are slightly lower for BGROUP than the corresponding CGROUP estimates. The three correlation estimates are slightly higher for BGROUP than the corresponding CGROUP estimates.

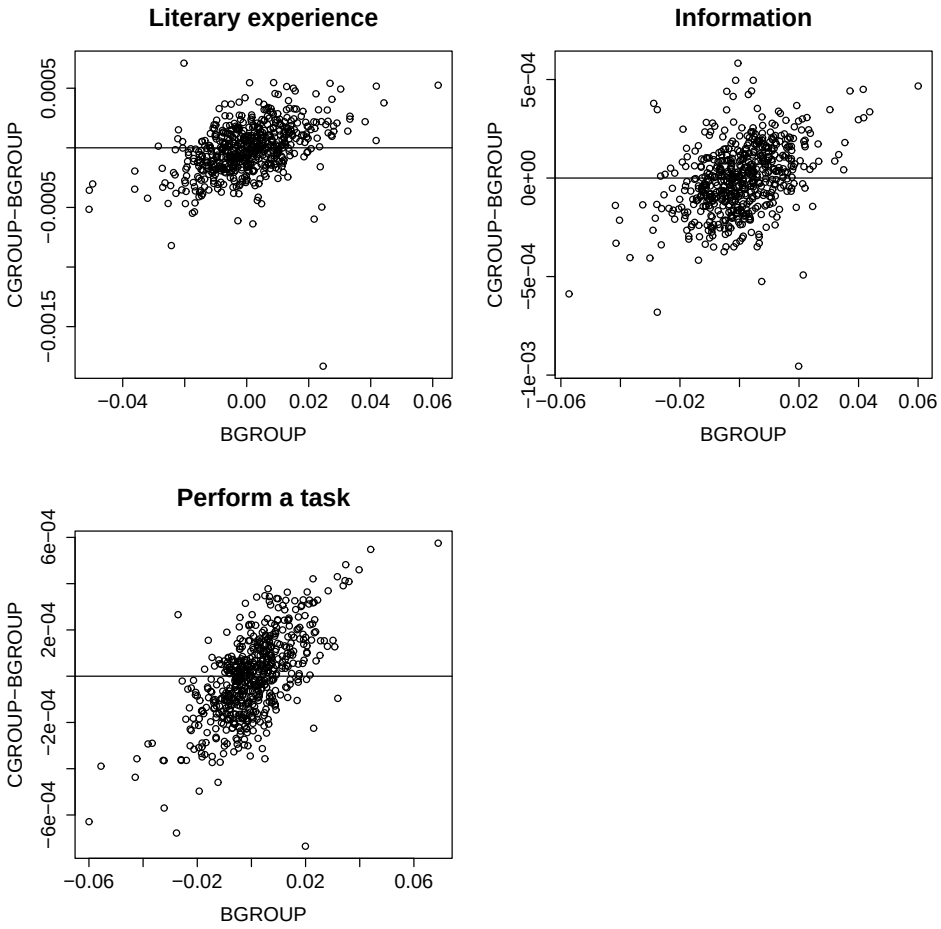

***Figure* 4. Comparison of regression coefficients from CGROUP and extended BGROUP for the 2002 NAEP reading assessment at grade 8.**

Figures 5 and 6 compare the marginal posterior means and SDs of 1,000 randomly chosen examinees for CGROUP and extended BGROUP for the three subskills. Figure 5 also shows the differences roughly in the same scale as reported by NAEP. We compute the weighted average of the posterior means in the three subscales for each examinee using the weights 0.4, 0.4, and 0.2, as used in NAEP. Then we applied a linear transformation of the resulting weighted average to convert it to a scale with a mean of 264 and an SD of 35 (the reported values of the composite for the 2002 NAEP reading assessment at grade 8). Results are very similar to those for the previous example (Figures 2 and 3); for example, CGROUP slightly overestimates the extreme posterior SDs.

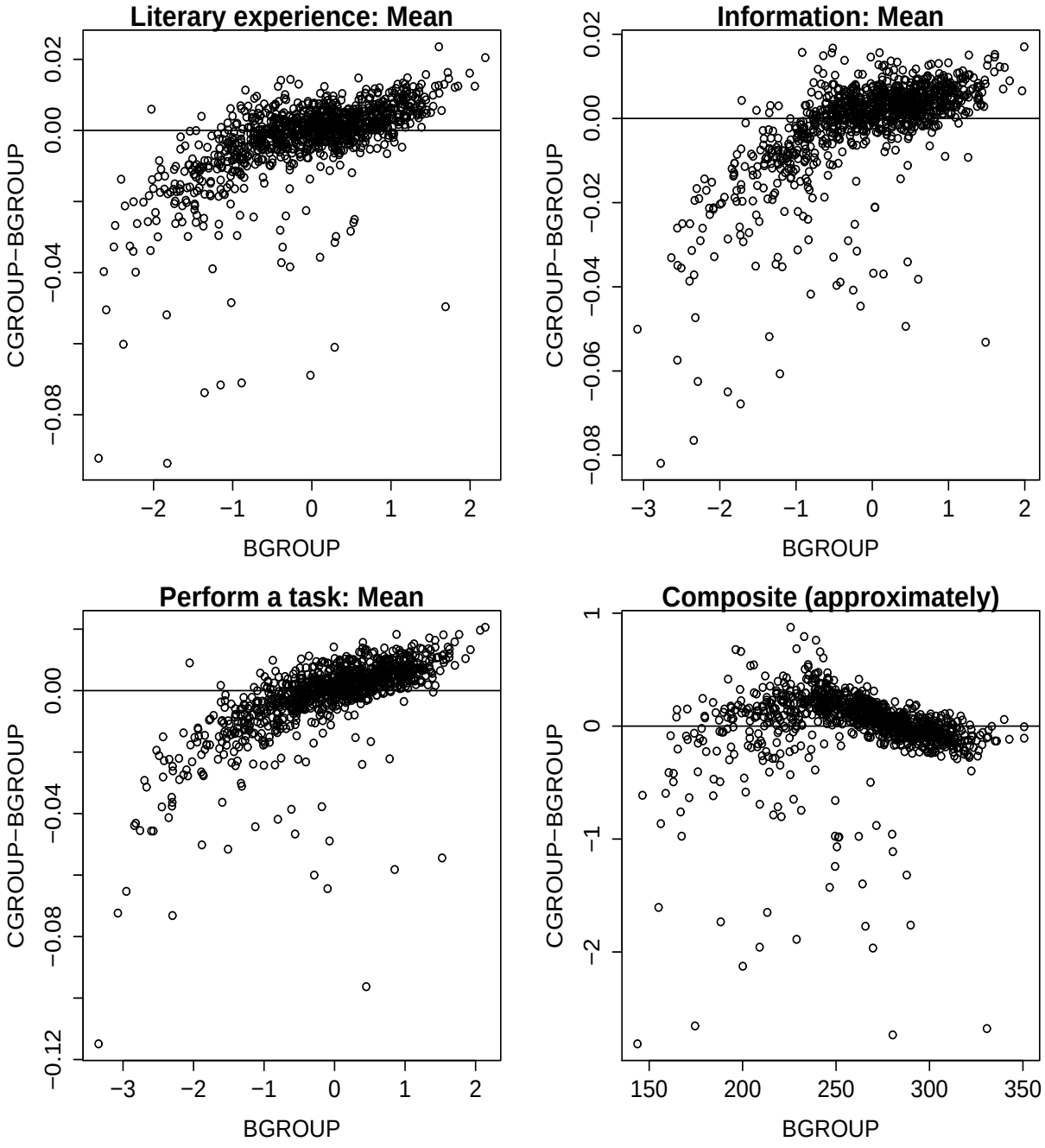

***Figure* 5. Comparison of posterior means from CGROUP and extended BGROUP for the 2002 NAEP reading assessment at grade 8.**

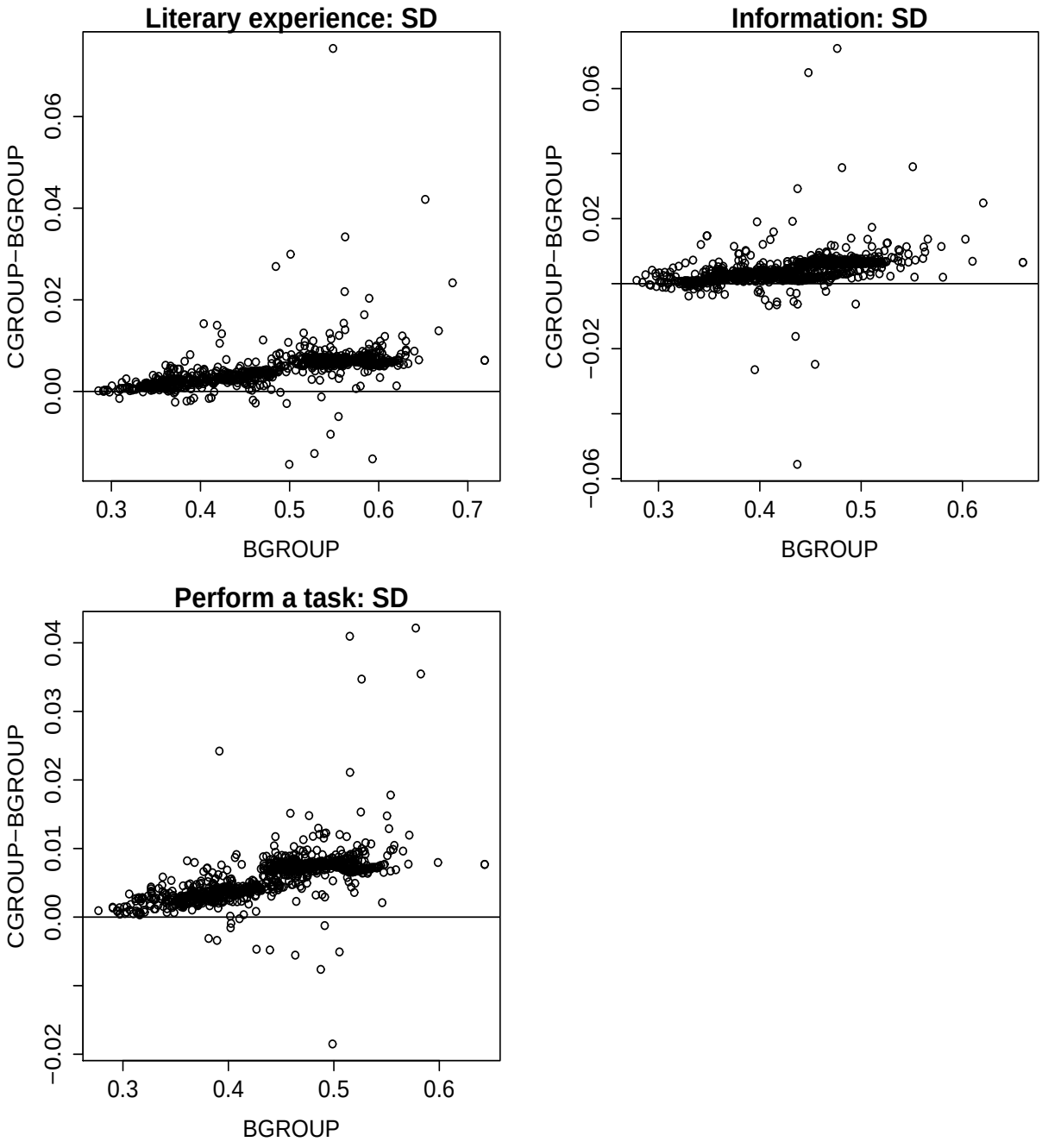

***Figure* 6. Comparison of posterior SDs from CGROUP and extended BGROUP for the 2002 NAEP reading assessment at grade 8.**

Table 3.

Residual Variances, Covariances, and Correlations for the 2002 NAEP Reading Assessment at Grade 8

	BGROUP			CGROUP-BGROUP		
	Literary	Information	Perform	Literary	Information	Perform
Literary	0.517	0.363	0.339	0.010	0.004	0.004
Information	0.765	0.435	0.339	-0.006	0.009	0.005
Perform	0.733	0.800	0.413	-0.007	-0.006	0.010

Note. Residual variances are shown on main diagonals, covariances on upper off-diagonals, and correlations on lower off-diagonals.

Table 4 compares the subgroup means and SDs (in parentheses) from BGROUP and CGROUP for relevant subgroups—there seems to be little difference between the two methods in this aspect as well. The BGROUP means are larger than or equal to the CGROUP mean for all but three entries and the BGROUP SDs are all slightly less than the CGROUP SDs.

4. Conclusions

CGROUP is the current operational method used in large-scale assessments such as NAEP. Though CGROUP provides more accurate results than its predecessor (N-group), it is not without problems, as demonstrated by Thomas (1993a) and von Davier and Sinharay (2004). In particular, CGROUP is found to inflate variance estimates for examinees with large posterior variances. Currently, there is no entirely satisfactory alternative to CGROUP.

As this work shows, an extension of the BGROUP routine to more than two dimensions provides a viable alternative to CGROUP. CGROUP was found to overestimate the posterior SDs of examinees (and hence to overestimate the SDs of population subgroups); CGROUP also was found to mostly underestimate low posterior means, mostly overestimate high posterior means, and mostly underestimate the population subgroup

Table 4.

Comparison of Subgroup Estimates From Extended BGROUP and CGROUP for the 2002 NAEP Reading Assessment at grade 8

	BGROUP			CGROUP-BGROUP		
Subgroup	Literary	Information	Perform	Literary	Information	Perform
Overall	0.027	0.022	0.022	-0.002	-0.001	-0.001
	(0.984)	(0.949)	(0.970)	(0.009)	(0.008)	(0.010)
Male	-0.116	-0.078	-0.133	-0.003	-0.002	-0.002
	(0.983)	(0.962)	(0.967)	(0.009)	(0.009)	(0.010)
Female	0.170	0.123	0.178	0.000	0.000	0.001
	(0.965)	(0.925)	(0.947)	(0.008)	(0.007)	(0.009)
White	0.286	0.267	0.315	0.000	0.001	0.002
	(0.897)	(0.852)	(0.849)	(0.007)	(0.006)	(0.008)
Black	-0.497	-0.447	-0.516	-0.006	-0.004	-0.005
	(0.935)	(0.897)	(0.903)	(0.011)	(0.009)	(0.011)
Hispanic	-0.408	-0.434	-0.516	-0.005	-0.004	-0.005
	(0.982)	(0.981)	(0.982)	(0.010)	(0.011)	(0.012)
Asian	0.137	0.200	0.124	-0.001	0.000	0.000
	(0.959)	(0.944)	(0.940)	(0.008)	(0.008)	(0.010)
American	-0.330	-0.299	-0.250	-0.004	-0.002	-0.002
Indian	(0.946)	(0.922)	(0.953)	(0.009)	(0.008)	(0.011)

means in the two examples here; this phenomenon has not been reported yet in literature. One problem with the extension of BGROUP is run time. Currently, the extension takes much longer to run than what can be afforded operationally. However, the program can be used to check the accuracy of the CGROUP results in a secondary analysis. In an attempt to make the extended BGROUP routine operational, we plan to apply a rescaling of integrals (Haberman, 2003) in future to reduce the run time of the extended BGROUP program—the idea is elaborated in the appendix.

References

Beaton, A. (1987). *The NAEP 1983-84 technical report.* Princeton, NJ: ETS.

Beaton, A. E. (1988). *Expanding the new design: The NAEP 1985-86 technical report.* Princeton, NJ: ETS.

Beaton, A., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics, 17*, 95-109.

Das, I. (2000). Robustness optimization for constrained nonlinear programming problems. *Engineering Optimization, 32,*(5), 585-618. Available from http://www.caam.rice.edu/ indra/researchwork.html

von Davier, M., & Sinharay, S. (2004). *Application of the stochastic EM method to latent regression models (ETS RR-04-34).* Princeton, NJ: ETS.

von Davier, M., & Yu, H. T. (2003). *Recovery of population characteristics from sparse matrix samples of simulated item responses.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

Davis, P. J., & Rabinowitz, P. (1967). *Numerical integration.* Waltham, MA: Blaisdell.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B, 39*, 1-38.

Haberman, S. H. (2003). *Numerical integration.* Presentation to the ETS Statistical Methods Club, Princeton, NJ.

Johnson, M. S. (2002). *A Bayesian hierarchical model for multidimensional performance assessments.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Johnson, M. S., & Jenkins. F. (2004). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress.* Manuscript in progress.

Kass, R., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models. *Journal of the American Statistical Association, 84*, 717-726.

Kirsch, I. (2001) *The international adult literacy survey (IALS): Understanding what was measured* (ETS RR-01-25). ETS: Princeton, NJ.

Martin, M. O., & Kelly D. L. (1996). *Third International Mathematics and Science Study - Technical report volume I: Design and development.* Chestnut Hill, MA: Boston College.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika, 44,* 358-381.

Mislevy, R. (1985). Estimation of latent group effects, *Journal of the American Statistical Association, 80,* 993-997.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*, 131-154.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's Study of Reading Literacy Achievement in Primary Schools.* Chestnut Hill, MA: Boston College.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society, 31*(3), 214-225.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., & Naylor, J. C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician, 36*, 75-82.

Thomas, N. (1993a). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*(3), 309-322.

Thomas, N. (1993b). *The E-step of the MGROUP EM algorithm* (ETS TR-93-37). Princeton, NJ: ETS.

Zellner, A. (1962). An efficient method for estimating seemingly unrelated regressions and tests for aggregate bias, *Journal of the American Statistical Association, 57,* 348-368.

Appendix

Application of Rescaling of Integrals

Future work will include application of rescaling of the integral involved in BGROUP, which should reduce its run time significantly. The Gauss-Hermite integration technique approximates an integral of the form $\int_{-\infty}^{\infty} f(z)\exp(-z^2)dz$ as

$$\int_{-\infty}^{\infty} f(z)\exp(-z^2)dz \approx \sum_{i=1}^{I} \omega_i f(z_i), \quad \text{for } \omega_i = \frac{2^{I-1}I!\sqrt{\pi}}{I^2\{H_{I-1}(z_i)\}^2}, \tag{14}$$

where z_i is the ith zero of the Hermite polynomial $H_I(z)$ (see, e.g., Davis & Rabinowitz, 1967). Tables of z_i, ω_i, and so on are available (e.g., Davis & Rabinowitz, 1967).

For multidimensional integration, such as for $p-$dimensional $\boldsymbol{z}$, (14) can be generalized using the cartesian product rule (see, e.g., Naylor & Smith, 1982; Smith, Skene, Shaw, & Naylor, 1987) as

$$\int_{-\infty}^{\infty} f(\boldsymbol{z})\exp(-\boldsymbol{z}'\boldsymbol{z})d\boldsymbol{z} \approx \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_p=1}^{I_p} \omega_{1,i_1}\omega_{2,i_2}\ldots\omega_{p,i_p} f(z_{1,i_1}, z_{2,i_2}, \ldots z_{p,i_p}), \tag{15}$$

where $\omega_{1,i_1}, \omega_{2,i_2}, \ldots \omega_{p,i_p}$ are obtained as in (14), the univariate case.

Haberman (2003) discussed an example of rescaling an unidimensional integral where the goal is to obtain an estimate of

$$\int_{-\infty}^{\infty} b(z)\exp(c(z))\exp(-z^2)dz. \tag{16}$$

Suppose $c(z)$ is maximized at z_0. Then one may write

$$c(z) = c(z_0) + \frac{1}{2}c''(z_0)(z - z_0)^2 + \delta(z),$$

where $\delta(z) \equiv c(z) - c(z_0) - \frac{1}{2}c''(z_0)(z - z_0)^2$ Then one may express (16) as

$$\exp(k)\int_{-\infty}^{\infty} h(u)\exp(-u^2)du \tag{17}$$

for some k and an $h(u)$ that is much less variable than the original integrand $b(z)\exp(c(z))$ and approaches 0 very rapidly as u becomes more distant from 0. An application of the Gauss-Hermite integration with few points (e.g., Davis & Rabinowitz, 1967) is enough to achieve a high level of accuracy, which would not be possible with (16). Naylor and

Smith (1982) use a similar idea involving transformation of the original variables of integration in an attempt to make the resulting density close to the standard multivariate normal density.

Let us consider estimation of $E(f(\boldsymbol{\theta})|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Gamma}_t,\boldsymbol{\Sigma}_t)$ in MGROUP. For example, $f(\boldsymbol{\theta})$ is the same as $\boldsymbol{\theta}$ in (7), that is, while calculating expectation of $\boldsymbol{\theta}$. Let us denote $\boldsymbol{\theta}_0$ to be the mode of $l(\boldsymbol{\theta})$. We have

$$E(f(\boldsymbol{\theta})|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Gamma}_t,\boldsymbol{\Sigma}_t) \equiv \frac{\int f(\boldsymbol{\theta})l(\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}'\boldsymbol{x},\boldsymbol{\Sigma})d\boldsymbol{\theta}}{\int l(\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}'\boldsymbol{x},\boldsymbol{\Sigma})d\boldsymbol{\theta}} \tag{18}$$

To compute (18) using numerical quadrature, one requires several quadrature points (41 points are used in current operational BGROUP), mainly because of the variability of the integrand over the range of integration. However, it is possible to rescale the integral so that a few quadrature points might be enough to estimate the integral to an acceptable level of accuracy.

Applying the idea in Haberman (2003), the denominator in (18) can be written as

$$\begin{aligned}
&\int f(\boldsymbol{\theta})l(\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}'\boldsymbol{x},\boldsymbol{\Sigma})d\boldsymbol{\theta} \\
\equiv\ &\int f(\boldsymbol{\theta})\exp\{u(\boldsymbol{\theta})\}\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}'\boldsymbol{x},\boldsymbol{\Sigma})d\boldsymbol{\theta}, \quad \text{for } \exp\{u(\boldsymbol{\theta})\} \equiv l(\boldsymbol{\theta}) \\
\equiv\ &\int f(\boldsymbol{\theta})\exp\{u(\boldsymbol{\theta}_0)+\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'u''(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\Delta(\boldsymbol{\theta})\}\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}'\boldsymbol{x},\boldsymbol{\Sigma})d\boldsymbol{\theta} \text{ as } u'(\boldsymbol{\theta}_0)\equiv 0 \\
\equiv\ &e^{u(\boldsymbol{\theta}_0)}\int f(\boldsymbol{\theta})\exp(\Delta(\boldsymbol{\theta}))\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\}d\boldsymbol{\theta}
\end{aligned}$$

where $\Delta(\boldsymbol{\theta}) = u(\boldsymbol{\theta}) - u(\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'u''(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$,

$$\boldsymbol{\mu} = \left\{\boldsymbol{\Sigma}^{-1} - u''(\boldsymbol{\theta}_0)\right\}^{-1}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}'\boldsymbol{x} - u''(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0\right\}, \qquad \boldsymbol{S} = \left\{\boldsymbol{\Sigma}^{-1} - u''(\boldsymbol{\theta}_0)\right\}^{-1}.$$

On the application of a transformation $\boldsymbol{\psi} = \frac{1}{\sqrt{2}}\boldsymbol{S}^{-1/2}(\boldsymbol{\theta}-\boldsymbol{\mu})$, the above integral becomes

$$\sqrt{2}|\boldsymbol{S}^{1/2}|e^{u(\boldsymbol{\theta}_0)}\int f\left(\boldsymbol{\mu}+\boldsymbol{S}^{1/2}\boldsymbol{\psi}\sqrt{2}\right)\exp\left\{\Delta\left(\boldsymbol{\mu}+\boldsymbol{S}^{1/2}\boldsymbol{\psi}\sqrt{2}\right)\right\}\exp\left\{-\boldsymbol{\psi}'\boldsymbol{\psi}\right\}d\boldsymbol{\psi}.$$

Now one can apply the multiple Gauss-Hermite integration given by (15) to the above. The quantity $\exp\{-\boldsymbol{\psi}'\boldsymbol{\psi}\}$ forms the density of an independent normal vector and the quantity $\left\{\Delta\left(\boldsymbol{\mu}+\boldsymbol{S}^{1/2}\boldsymbol{\psi}\sqrt{2}\right)\right\}$ is much less variable (and close to zero) over the range of $\boldsymbol{\theta}$ than is $u(\boldsymbol{\theta})$. Therefore, multiple Gauss-Hermite integration with few points per dimension should provide adequate accuracy and precision.