Journal of Educational Statistics
Summer 1990, Vol. 15, No. 2, pp. 91–112

Multivariate Variance Component Analysis: An Application in Test Development

Nicholas T. Longford
Educational Testing Service

Key words: *latent trait, multivariate data, psychometric test, variance components*

This paper presents a multilevel variance component analysis of data from the pilot year of pretesting of an educational test. The multilevel nature of the data is induced by the clustering of students within colleges and by having multivariate observations (scores) on students. The presence of a multidimensional trait underlying the scores is formulated as a hypothesis about the full rank of a variance matrix.

The need for objective measurement of the quality of general education in United States colleges has been appreciated for some time. Colleges are facing increasing demand for provision of such information to state governments, accrediting boards, and organizations assisting in placement of high school students. Also, many college administrators are interested in comparisons of their freshman and upperclass cohorts, or in longitudinal comparisons of freshmen and upperclassmen. Awareness of the competition for the same freshman candidates has raised the importance of comparisons among colleges. Educational testing organizations in the United States have responded to this demand by instituting efforts to develop tests of general education outcomes.

In a typical educational admissions test, such as the SAT, ACT, or GRE, students are administered sets of items, the responses to which are scored and the scores reported back to students. Test developers usually attempt to set the number of items in the test close to a minimum that would make the test scores for the examinees sufficiently reliable. The testing time (i.e., the time period during which the examinees are exposed to the test items) should be set to a minimum in which the majority of students would have sufficient time to study and to respond to all the test items, so as to encroach to the smallest extent on the time available for (academic) instruction or on the free time of the examinees.

The author has benefited from discussions with Charlie Lewis. Comments of an associate editor and of two referees have helped to improve the presentation. Secretarial services of Liz Brophy, ETS, are acknowledged.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

If assessment of colleges is the prime objective, it would be a luxury to administer a test instrument that yields a reliable score for each individual examinee, because the within-college means of such scores would have unnecessarily high reliability. It would be preferable, for example, to administer discrete sections of the test to randomly selected subgroups of examinees (e.g., by spiralling).

In a more general context, designs in which examinees are administered different subsets of the test item pool (i.e., multiple matrix sampling) have been suggested by Lord (1962). Sirotnik and Wellington (1977) have described a comprehensive statistical method for analysis of data arising from such a design. Goldstein and James (1983) have derived efficient computational procedures for analysis of such data. These methods assume that the response patterns of the examinees are uncorrelated. In this paper I extend these methods to the case of correlated vectors of subscores. The correlation structure is induced by the random effect vectors associated with colleges. I set up a multivariate version of the standard random effects model with one-way layout along the lines of Goldstein and McDonald (1988), and for model fitting I used computational procedures for maximum likelihood estimation based on Longford (1987); for the technical details, see Longford (1989), and for further background, Searle (1971) and Harville (1977).

Another method for extracting maximum information within minimum testing time is to administer tests consisting of sections (subtests, testlets, or cells of items) that tap different, though related, aspects of examinees' knowledge, ability, or skills. An important issue with such tests is to establish discriminant validity of the subscores. The "true" subscores underlying the observed subscores are often highly correlated. If two true subscores, related to different domains of ability (subtests), are perfectly correlated, then the corresponding observed scores are merely two less reliable versions of the true score underlying the aggregate of the two subtests. Then it is preferable to provide only the observed score for the aggregate, thus simplifying the format of the score report.

With examinees clustered within colleges, it is useful to consider these issues separately for the scores of examinees within a college and for the true college means. For example, the true subscores underlying subtests A and B may be correlated .85 within each college in a population of colleges, but the within-college means of these true subscores may be perfectly correlated. Conversely, the true subscores may be perfectly correlated within each college, but the college means may have a lower correlation. Hence, a single correlation of the examinees' subscores is an incomplete summary of the relevant information contained in the subscores. If the clustering of examinees is more complex (e.g., students within grades within colleges within states), then each of these four layers of the clustering

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

hierarchy may have a different correlation structure of the true subscores (their means).

This paper presents a multivariate variance component analysis of data from the pilot stage of development of a test instrument aimed to provide information for colleges, or more generally groups of students, about several aspects of general education outcomes. The analysis addresses issues of discriminant validity of the subtests and other statistical issues that arise in the development of a test instrument with complex test specification.

The sections of the paper are organized as follows: details of the studied test instrument and statement of the principal research problem, overview of variance component models, definition of the substantively important dimension of the latent trait, data and procedures, results, and conclusions and summary.

The Test Instrument

The test instrument, in this paper referred to by the acronym GENED, involves a matrix design, as depicted in Figure 1. There are three forms of the test, each with 48 items. The items within a form are cross-classified into three content areas (humanities, social sciences, and natural sciences; the rows in Figure 1) and four academic skills (reading, writing, critical thinking, and mathematics; the columns in Figure 1). The numbers in Figure 1 represent the numbers of items within each cell of one of the three forms of the test, and the row and column subtotals. Of the 48 items, there are 16 in each row and 12 in each column; however, the numbers of items within cells vary both within a form and from form to form. Each item has four alternative responses, one of which is correct.

The first stage of pretesting of the instrument took place in the academic year 1986–87 and involved about 1,000 students from 12 colleges. The original pool of 240 items has been reduced to 144 items, and several items have been reviewed and amended. The main purpose of the second stage of pretesting, the pilot year (1987–88), was to establish content and construct validity of the test and of its components (skills and content areas) to decide what kinds of comparisons of college-mean scores can be made and to settle the format of the report for participating colleges.

The test can be administered in two ways. The short form involves 1 hour of testing time, and each student has to respond to 48 items of one form. The three forms are spiralled (in the order A, B, C, A, B, . . .) so as to ensure approximately equal numbers of students and equal distribution of ability for all three forms within colleges. Note that the three forms are unlikely to be exactly parallel, because that would mean that they have matching characteristics either for each cell or at least for the three rows and the four columns of the design. However, because individual examinee scores are not reported, this stringent condition can be dispensed with.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

	COLLEGE-LEVEL READING	COLLEGE-LEVEL WRITING	CRITICAL THINKING	USING MATHEMATICAL DATA	Three Content Subscores Reported
HUMANITIES	5	5	4	2	Humanities Subscore 16
SOCIAL SCIENCES	4	4	4	4	Soc. Sci. Subscore 16
NATURAL SCIENCES	3	3	4	6	Nat. Sci. Subscore 16
Four Skill Subscores Reported	Reading Subscore 12	Writing Subscore 12	Cr. Think Subscore 12	Maths Subscore 12	TOTAL SCORE 48

FIGURE 1. *An example of the design of a form of the GENED test (the cells contain the numbers of items; the exact number of items per cell varies from form to form)*

The long form involves 3 hours of testing time, in which each student is administered three forms of the test in three separate 1-hour sessions. The order of administration of the forms within the long form is balanced to enable inference about fatigue and practice effects. The score report for the long form administration will contain both individual students' scores and college-average scores and subscores.

Results of the short form administration are intended for colleges, or groups of students, only. The college-average number of correct item responses would be reported for the entire test, for the three subject areas, and for the four skills. Suitable minimum numbers of students from a college must be recommended, for whom meaningful scores and subscores could be provided. Alternatively, the expected standard errors associated with these scores, as functions of the number of students in a college, would suffice. Colleges may administer the test to a random sample of students, and then the choice of the sampling size is important.

Another important point is to decide which subscore means to report,

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

and in what form. The subscore means are highly correlated, and therefore it is more informative to report certain contrasts. Suppose two colleges have substantially different values of a given contrast of true subscore means. Then this contrast of subscore means is declared to be worth reporting, for colleges with n examinees each, if in a future administration with n examinees from each college the observed subscore means are very likely to be significantly different, using the conventional t test.

Variance Component Models

Data from tests with multichoice items are often analyzed using the classical item-response methodology (IRT), implemented in the software packages LOGIST (Wingersky, Barton, & Lord, 1982), BILOG (Mislevy & Bock, 1983), and others. The IRT approach relies heavily on the unidimensionality of the underlying "examinee-ability" trait, whereas our interest is in establishing multidimensional traits underlying the responses of examinees.

Because we are interested in both the traits underlying student scores and college-mean scores, in the standard factor analytical approach we require a decomposition of the variance matrix of the subscores into the between-college and within-college components.

Variance component methods have been proposed for analysis of multivariate data involving clustering by Harville (1977) and, more recently, Goldstein (1987). Software for variance component analysis implementing the maximum likelihood algorithms of Goldstein (1986) and Longford (1987) enables simultaneous modeling of several features of the data that can be represented as levels of the nesting hierarchy.

For illustration, consider first an additive model for the scores in the 12 cells of the design:

$$Y_{ijkh} = \mu_{kh} + a_{ikh} + b_{ijkh} + \epsilon_{ijkh}. \tag{1}$$

The term μ_{kh} represents the overall mean score for the cell (k, h), the college-specific term a_{ikh} represents the deviation of the *true* college mean score $\mu_{kh} + a_{ikh}$ from the overall mean score for the cell (k, h), and the examinee term b_{ijkh} stands for the deviation of the examinee's true score from the true mean score for his or her college. The random term ϵ_{ijkh} is included to account for the aggregate of the random influences associated with momentary variation in human performance and recall, external influences, guessing, and so on, and for model inadequacy.

Differences in the difficulties as well as in the numbers of items in the cells of the three forms can be accommodated by form-specific parameters μ_{khf} $(f = 1, 2, 3)$ in (1). We will assume, however, that the college and student deviations, a_{ikh} and b_{ijkh}, are form-independent.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

Because each cell contains only 2–6 items, the assumptions of normality in (1) are difficult to justify, even after a transformation of the subscores. We circumvent these issues by aggregating the subscores to the level of content areas and skills. Because the development for content area and skill subscores is completely analogous, I describe in detail only the latter.

For the skill subscores we have an analysis-of-variance type model,

$$Y_{ijk\cdot} = M_{kf} + A_{ik} + B_{ijk} + \epsilon_{ijk\cdot}, \tag{2}$$

where $M_{kf} = \Sigma_h \mu_{khf}$, $A_{ik} = \Sigma_h a_{ikh}$, $B_{ijk} = \Sigma_h b_{ijkh}$, and $\epsilon_{ijk\cdot} = \Sigma_h \epsilon_{ijkh}$. The assumption of normality of these subscores, or of their suitable transformations, is now more palatable than in the model (1). Assume that the college terms $\mathbf{A}_i = (A_{i1}, A_{i2}, A_{i3}, A_{i4})$, the examinee terms $\mathbf{B}_{ij} = (B_{ij1}, B_{ij2}, B_{ij3}, B_{ij4})$, and ϵ_{ijk} form three mutually independent random samples:

$$\begin{aligned} \mathbf{A}_i &\sim N_4(\mathbf{0}, \mathbf{\Omega}_A), \\ \mathbf{B}_{ij} &\sim N_4(\mathbf{0}, \mathbf{\Omega}_B), \\ \epsilon_{ijk\cdot} &\sim N(0, \sigma^2). \end{aligned} \tag{3}$$

The variance matrix for the four skill subscores of an examinee is equal to $\sigma^2\mathbf{I} + \mathbf{\Omega}_B + \mathbf{\Omega}_A$, and the conditional variance matrix, given the college-level random terms $\mathbf{A}_i$ (i.e., the within-college variance matrix), is $\mathbf{\Omega}_B^* = \sigma^2\mathbf{I} + \mathbf{\Omega}_B$. In absence of any replication for the subscores the components σ^2 and $\mathbf{\Omega}_B$ are confounded, and therefore σ^2 cannot be estimated without some data in addition to $\{Y_{ijk\cdot}\}$.

An upper bound for σ^2 can be obtained from the eigenvalue decomposition of $\mathbf{\Omega}_B^*$; we have $\sigma^2 \leq e_1$, where e_1 is the smallest eigenvalue of $\mathbf{\Omega}_B^*$. If $\sigma^2 < e_1$, then $\mathbf{\Omega}_B = \mathbf{\Omega}_B^* - \sigma^2\mathbf{I}$ is of full rank; otherwise, the rank of $\mathbf{\Omega}_B$ is equal to the number of eigenvalues greater than e_1. Thus, an estimate of $\mathbf{\Omega}_B^*$ can provide an estimate of the upper bound for σ^2. A different "guestimate" can be obtained by considering the variation associated with binomial outcomes (0/1 responses to 12 items). As a crude approximation for the upper bound for σ^2, we have $12 \times (½)^2 = 3$. Note, however, that the responses contributing to a subscore are probably positively correlated, even after partialling out the college and examinee terms. Reliable inference about σ^2 could be obtained only from an experiment involving replication, embedded in the administation of the test.

Because the skill subscores, or their averages, are highly correlated, it is advantageous to consider a nonsingular linear transformation of the subscores. I choose the transformation

$$\begin{aligned} A_{i1}^* &= (A_{i1} + A_{i2} + A_{i3} + A_{i4})/4, \\ A_{ik}^* &= A_{ik} - A_{i1} \qquad \text{for } k = 2, 3, 4, \end{aligned}$$

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

that is, $\mathbf{A}_i^* = \mathbf{A}_i \mathbf{D}_4$, for college-level random effects, where

$$\mathbf{D}_4 = \begin{bmatrix} 1/4 & -1 & -1 & -1 \\ 1/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/4 & 0 & 0 & 1 \end{bmatrix},$$

and $\mathbf{B}_{ij}^* = \mathbf{B}_{ij} \mathbf{D}_4$ for examinee-level random effects. I will refer to this transformation, given by the matrix $\mathbf{D}_4$, as the *centered* transformation (parameterization). The first components of $\mathbf{A}_i^*$ and $\mathbf{B}_{ij}^*$ are associated with the score average, and the others with the respective contrasts of writing, critical thinking, and mathematics versus reading. The variance matrices for these random effects are $\mathbf{\Sigma}_A = \mathbf{D}_4^T \mathbf{\Omega}_A \mathbf{D}_4$ and $\mathbf{\Sigma}_B = \mathbf{D}_4^T \mathbf{\Omega}_B \mathbf{D}_4$. The inverse transformation is given by the matrix

$$\mathbf{C}_4 = \mathbf{D}_4^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix},$$

$(\mathbf{A}_i = \mathbf{A}_i^* \mathbf{C}_4, \mathbf{B}_{ij} = \mathbf{B}_{ij}^* \mathbf{C}_4)$.

Explanatory variables, such as background information, can be directly incorporated within the systematic (fixed) part M_{kf} in model (2). We can consider variables defined for examinees (age, sex, major, high school academic achievement, and so on) and for colleges (public/private, total enrollment, and so on). If the test is administered to distinct subgroups of students within a college, such as freshmen and upperclassmen, or the four grades, the model (2) can be extended by another level of clustering (examinees within grades within colleges) to

$$Y_{igjk\bullet} = M_{kgf} + A_{ik} + G_{igk} + B_{igjk} + \epsilon_{igjk\bullet}, \tag{4}$$

where the assumptions (3) are supplemented by $\mathbf{G}_{ig} = (G_{ig1}, \ldots, G_{ig4}) \sim N_4(\mathbf{0}, \mathbf{\Omega}_G)$. Systematic differences among the grades are accommodated within M_{kgf}, and the within-college deviations from the mean differences among the grades are represented by the random vector $\mathbf{G}_{ig}$. Models for longitudinal data are obtained by introducing a level of clustering between the examinee and subscore levels (subscores within occasions for examinees). Similarly, repeated administrations across years correspond to the clustering of students within college–years within colleges (assuming no migration).

The number of colleges taking part in the pilot year is limited; the participation is voluntary and recruitment of colleges requires extensive efforts. Therefore, only a small number of college-level variables can be

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

used to model the fixed part M_{kf} (M_{kgf}). There is more scope for incorporating examinee-level variables, but it should be borne in mind that such variables may be almost constant within colleges. If several college-level variables, or variables with small within-college variation, are used, then the variance matrix $\mathbf{\Sigma}_A$ may vanish.

Voluntary participation raises several issues related to sampling. The hope is that the colleges taking part in the pilot year are a representative sample of the hypothetical population of colleges that will administer the test in future. However, the composition of colleges interested in future operational administration may change because of various circumstances, including the results of the pilot year analysis. Another issue concerns participation and motivation of the designated students within participating colleges. In the short form administration the examinee's name is not entered on the response form, and so he or she is aware that no substantial personal stake is involved. The problem arises to a lesser extent in the long form administration, where individual examinees' scores are reported. This problem relates not only to the pilot year data but also to future operation because differential motivation of the examinees may seriously influence the intended comparisons.

Dimensionality of the Observable Trait

I subscribe to the approach of Goldstein and McDonald (1988) for factor analysis with multilevel data, but in the present context it is necessary to distinguish between statistical significance and substantive importance. I adopt a definition of the underlying multidimensional trait for a normally distributed random vector, similar to that used in factor analysis. By an m-dimensional normal trait I mean any m-variate normal vector with mean $\mathbf{0}$ and a nonsingular variance matrix. A normal vector $\mathbf{z}$ is said to have an underlying m-dimensional trait if it can be formed by a nonsingular linear transformation of the m-dimensional trait, that is,

$$\mathbf{z} = \mathbf{T} + \mathbf{U}\boldsymbol{\delta}, \tag{5}$$

where $\boldsymbol{\delta} \sim N_m(\mathbf{0}, \mathbf{\Psi})$, $\mathbf{\Psi}$ is nonsingular, $\mathbf{U}$ is a matrix of rank m, and $\mathbf{T}$ is a vector of constants. The matrices $\mathbf{U}$ and $\mathbf{\Psi}$ are not defined uniquely, and the analyst can make a choice that facilitates a suitable interpretation. In many situations it is useful to restrict the variance matrix $\mathbf{\Psi}$ to a unit matrix or a diagonal matrix, but this restriction is purely formal. Two random vectors are said to have the same trait if the corresponding matrices $\mathbf{U}$ in their representations (5) span the same linear space. In particular, two random vectors of the same length with nonsingular variance matrices have the same trait.

An important issue in the development of the GENED test is in establishing a four-dimensional trait underlying the four skill subscores (and a

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

three-dimensional trait for the subject-area subscores). It is important to consider separately examinee-level and college-level traits (and grade-level traits, if applicable). Inference about these traits is related to the eigenvalue decompositions of the corresponding variance matrices $\mathbf{\Omega}_B$ and $\mathbf{\Omega}_A$ (and $\mathbf{\Omega}_G$), or, equivalently, those of $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_A$ (and $\mathbf{\Sigma}_G$).

For example, if $\mathbf{\Sigma}_A = 0$, then there are no differences among the college-level true subscore means, and consequently there is no point in reporting any college-level subscore means. If $(\mathbf{\Sigma}_A)_{11} > 0$, but all the other elements of $\mathbf{\Sigma}_A$ vanish, then the total score represents the sole dimension of the trait underlying college-level variation in the subscores. Another hypothesis is that reading, writing, and critical thinking represent a single college-level trait, and therefore they could be fully described by a single subscore; reporting the three separate subscores would not be justified. My analysis strongly supports this hypothesis.

An important caveat must be attached to all the conclusions. My analysis assumes that the labels for the rows (subject areas) and the columns (skills) of the test are appropriate. That is, items within a cell of the design, a combination of a subject area and skill, do not contain elements of other content or knowledge areas, or at least that, within certain generally accepted bounds, knowledge of the subject and possession of the skill constitute (apart from guessing) the major component in the process by which an examinee makes a response choice to the items in the cell.

I emphasize that the main objective of the pretest study is to make inference not about the colleges participating in the pretest, but about the distribution of the reported scores, subscores, and various contrasts of scores in the *future* operation. For example, reporting a score that a college wants to compare with the score for another college, or the average of a group of colleges, would be of limited use if most such comparisons would be statistically not significant. The higher the number of examinees from a college, the higher the utility of the mean subscores because they are less affected by the "noise" due to examinee-level variation. An appropriate definition of an *observable* trait should reflect this dependence on the administration size.

More formally, suppose we have two colleges with n examinees each, and they are to be compared in a future administration on a contrast of skill subscores given by a 4×1 vector $\mathbf{a}$. Suppose the *true* contrasts $\mathbf{a}^T\mathbf{A}_1$ and $\mathbf{a}^T\mathbf{A}_2$ for the two colleges differ by the amount expected if the two colleges were drawn from the population of colleges at random, that is,

$$\mathbf{a}^T(\mathbf{A}_1 - \mathbf{A}_2) = \sqrt{2\mathbf{a}^T\mathbf{\Omega}_A\mathbf{a}}. \tag{6}$$

Conditionally on (6), the difference of the *observed* contrasts, $\mathbf{a}^T(\mathbf{Y}_1 - \mathbf{Y}_2)$, has a normal distribution with mean $\mathbf{a}^T(\mathbf{A}_1 - \mathbf{A}_2)$ and variance $2\mathbf{a}^T\mathbf{\Omega}_2^*\mathbf{a}/n$. We can interpret (6) as "colleges differ a great deal" on the contrast $\mathbf{a}$. It

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

is desired that the (future) observed contrasts be significantly different (at level α_1),

$$\mathbf{a}^T(\mathbf{Y}_1 - \mathbf{Y}_2) \geq \Phi^{-1}(1 - \alpha_1) \cdot \sqrt{2\mathbf{a}^T \boldsymbol{\Omega}_B^* \mathbf{a}/n}, \tag{7}$$

with high conditional probability:

$$\Pr\{(7)|(6)\} = \alpha_2,$$

where α_1 and α_2 are suitable constraints, and Φ is the distribution function for $N(0, 1)$. The latter equation is easy to solve with respect to n, and it yields

$$n \geq c \cdot \frac{\mathbf{a}^T \boldsymbol{\Omega}_B^* \mathbf{a}}{\mathbf{a}^T \boldsymbol{\Omega}_A \mathbf{a}}, \tag{8}$$

where $c = \{\Phi^{-1}(1 - \alpha_1) + \Phi^{-1}(\alpha_2)\}^2$.

For example, for the choice $\alpha_1 = .10$ and $\alpha_2 = .90$, we have $c = 6.57$. For a given contrast $\mathbf{a}$ I will refer to the smallest integer satisfying (8) as the *critical* administration size, and if this integer is smaller than the size of the administration in a substantial number of the colleges participating in the pilot-year data (or expected sizes in future administrations), then I declare this contrast as *observable*. The criterion (8) can be interpreted as follows: A contrast is observable if the between-college component of its variation, $\mathbf{a}^T \boldsymbol{\Omega}_A \mathbf{a}$, is greater than a certain multiple of the within-college component, $\mathbf{a}^T \boldsymbol{\Omega}_B^* \mathbf{a}$. The multiplicative factor is inversely proportional to the number of students in the college.

The desirable outcome for the testing program would be that all dimensions of the (four-dimensional) trait for the skill subscores are observable. The choice of the representation for the college-level trait may influence the critical administration sizes for the dimensions of the trait, because the definition (8) involves the examinee-level variance matrix $\boldsymbol{\Omega}_B^*$. However, the substantive conclusions about the observable traits, arrived at from my analysis, turn out not to depend on this choice because $\mathbf{a}^T \boldsymbol{\Omega}_A \mathbf{a}$ varies much more than $\mathbf{a}^T \boldsymbol{\Omega}_B^* \mathbf{a}$ on the sphere $\mathbf{a}^T \mathbf{a} = 1$.

Data and Procedures

Approximately 11,000 students from 52 colleges have taken part in the pilot-year administration of the test. The short form has been used in 34 colleges with 6,131 students, and the long form in 18 colleges with 4,835 students. In the short form administration motivation of examinees was a major problem; many examinees apparently gave up or proceeded by guessing at random after the first few items. In correspondence with the plans of the program directors, I have excluded from my analysis all the examinees who have scored 10 points or less (out of 48). More than 16% (972) of the examinees were excluded. Because provision of the GENED

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

test and its scoring would be a commercial undertaking, the testing organization would be committed to provide the college means of the subscores even in the presence of substantial attrition, which may drastically affect their validity. Therefore, excluding the colleges with substantial attrition from the analysis might result in bias of the predicted variation of the outcomes in the future operation. The attrition rate in the long form administration was negligible. In the short form administration the vast majority of the examinees were freshmen (91%), and only three colleges had 50 or more upperclass examinees. In the long form administration there were seven colleges in which the test was administered to both freshmen and upperclassmen, in six colleges only freshmen were examined, and in the remaining five colleges only upperclassmen took part. Details of the administration sizes are given in Tables 1 and 2.

Separate analyses were carried out for short form and long form data and for skill subscores and subject-area subscores within forms. Variance component models were fitted using the Fisher scoring algorithm of Longford (1987) implemented in the software VARCL (Longford, 1988). The centered parameterization was used, and instead of the variance parameters their square roots (sigmas) were estimated. Therefore, the standard errors for these sigmas are quoted throughout.

For model fitting in VARCL I declare, for the analysis of skill subscores in the short form, $4 \times 5{,}159$ elementary observations (subscores) within 5,159 examinees, within 34 colleges. For the elementary observations we have the categorical explanatory variable, indicating the subscore type (4 categories), and for the examinees the categorical explanatory variable, indicating the form of the test (3 categories). In order to allow a general pattern of the between-form differences in the population subscore means, the form-by-subscore interaction is considered. The between-college and between-examinee covariance structures are declared as associated with the subscore type. Model fitting is iterative, but the number of iterations required in the models fitted was always less than 12. The model parameters are the location parameters associated with the form-by-subscore interaction (using the centered parameterization), an imputed value for the elementary level variance σ^2, and the variance and covariance elements of the matrices $\mathbf{\Sigma}_A$ and $\mathbf{\Sigma}_B$. The choice of σ^2 has to be such that the estimate of $\mathbf{\Sigma}_B$ be positive definite (if too large a σ^2 has been chosen, the model has to be refitted with a smaller σ^2). For each estimated parameter its maximum likelihood estimate and the associated standard error are obtained, as well as the value of $-2 \times$ log-likelihood for the fitted model (useful for likelihood ratio testing). The conditional expectations of the random effects (i.e., the residuals corresponding to the random terms A_{ik} and B_{ijk}) are also provided. The software VARCL has no requirements for balance of the data.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

TABLE 1
Numbers of examinees from the colleges in the short form administration by form

College no.	Total	Form A	Form B	Form C	Excluded examinees[a]
1	61	21	21	19	2
2	217	72	72	73	5
3	400	134	131	135	4
4	278	105	93	80	80
5	132	45	43	44	2
6	183	69	69	45	47
7	133	44	44	45	1
8	75	27	25	23	7
9	81	27	28	26	0
10	72	47	47	48	1
11	48	14	20	14	25
12	100	34	33	33	0
13	148	49	50	49	5
14	171	57	56	58	2
15	176	76	60	40	137
16	117	39	39	39	0
17	57	19	19	19	1
18	40	13	13	14	2
19	96	33	32	31	4
20	77	25	27	25	1
21	151	52	51	48	6
22	158	54	50	54	5
23	172	60	55	57	5
24	93	39	34	26	21
25	81	22	39	20	13
26	72	24	24	24	1
27	132	47	40	45	3
28	80	30	25	25	11
29	130	66	63	61	7
30	242	88	81	78	0
31	112	37	37	38	2
32	84	28	27	29	2
33	517	172	175	170	4
34	332	116	112	104	66
Total	5,159	1,785	1,639	1,735	972

Note. $N = 5{,}159$.
[a] Examinees who scored less than 10 points (out of 48).

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

Results

The estimates of the college mean subscores are given in Tables 3 and 4. They differ from the corresponding arithmetic averages of the subscores by .01 or less. The standard errors associated with the variance component analysis estimates of the means also differ very little from the standard errors associated with the arithmetic averages. For example, the standard errors of the within-skill and within-form means in Table 3 are in the range .060–.064, whereas the corresponding values for the arithmetic averages are .057–.058.

Examinees in the short form administration appear to score substantially higher than in the long form administration. No college has administered both the short form and the long form, and so we can distinguish between genuine differences between the short form and long form populations on one hand and fatigue and loss of interest on the other hand only by looking at the results of the first session in the long form administration. There is strong evidence of fatigue and loss of interest because the subscores in the first session were much higher than in the later sessions.

TABLE 2
Numbers of students in the long form administration

College no.	Number of students	Freshmen	Upperclassmen
1	332	235	97
2	163	57	106
3	468	0	468
4	61	61	0
5	445	445	0
6	66	0	66
7	263	0	263
8	387	387	0
9	182	182	0
10	209	161	48
11	206	178	28
12	177	136	43
13	194	194	0
14	69	0	69
15	812	812	0
16	131	0	131
17	41	18	23
18	629	432	197
Total	4,835	3,298	1,539

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

TABLE 3
Mean skill- and content-area subscores in the short form administration

Form	Reading (12 items)	Writing (12 items)	Critical thinking (12 items)	Mathematics (12 items)	Humanities (16 items)	Social sciences (16 items)	Natural sciences (16 items)	Form total (48 items)
A	6.69	7.51	5.48	5.86	8.86	8.93	7.75	25.54
B	6.99	6.98	5.41	6.28	8.70	8.24	8.73	25.66
C	6.91	6.26	5.77	6.26	9.04	7.86	8.30	25.20

Note. The standard errors are .06 for the skill subscore means (12 items), .07 for the content-area subscore means (16 items), and .12 for the form totals (48 items).

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

TABLE 4
Mean skill- and content-area subscores in the long form administration

Reading	Writing	Critical thinking	Mathematics	Humanities	Social sciences	Natural sciences	Total
18.69	19.02	15.36	16.35	24.31	22.86	22.27	69.42

Note. The standard errors are .10 for the skill subscore means (36 items), .12 for the content-area subscore means (48 items), and .20 for the form totals (144 items).

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

The critical thinking items appear to be the most difficult, followed by mathematics. Humanities items are more difficult than the items in the other two subject areas. Difficulties across forms *A*, *B*, and *C* are substantially different for writing, social sciences, and natural sciences items.

Estimates of the variance component matrices are discussed separately for the analyses carried out.

Skill Subscores in the Short Form Administration

The estimate of the examinee-level variance matrix of the skill subscores is given in Table 5. According to the model (2) this variance matrix has two components, $\hat{\mathbf{\Omega}}_B^* = \sigma^2\mathbf{I} + \hat{\mathbf{\Omega}}_B$. The smallest eigenvalue of $\hat{\mathbf{\Omega}}_B^*$ is 2.43, and it is an estimate of the upper bound for σ^2. The estimates of $\mathbf{\Omega}_B$ and $\mathbf{\Sigma}_B$ depend on the imputed value of σ^2. For the choice of $\sigma^2 = 2$, we have

$$\hat{\mathbf{\Sigma}}_B = \mathbf{D}_4^T(\hat{\mathbf{\Omega}}_B^* - \sigma^2\mathbf{I})\mathbf{D}_4 = \mathbf{D}_4\hat{\mathbf{\Omega}}_B^*\mathbf{D}_4 - \sigma^2\mathbf{D}_4^T\mathbf{D}_4 = \begin{bmatrix} 2.484 & & & \\ .106 & 1.136 & & \\ -.007 & .325 & .928 & \\ -.175 & .562 & .570 & 2.003 \end{bmatrix}.$$

The asymptotic standard errors for the square roots of the variances (sigmas) (1.58, 1.07, .96, 1.42) are (.02, .08, .09, .05), and the standard errors for the covariances are in the range .04–.08. Although all the sigmas have high t ratios, it is clear that the within-college differences of the subscores critical thinking–reading and writing–reading are much smaller than for the contrasts involving the mathematics subscore. For the extreme choice of $\sigma^2 = 2.43$, the student level variance $\hat{\mathbf{\Sigma}}_B$ becomes singular,

$$\hat{\mathbf{\Sigma}}_B = \begin{bmatrix} 2.376 & & & \\ .106 & .276 & & \\ -.007 & -.106 & .068 & \\ -.175 & .132 & .139 & 1.129 \end{bmatrix},$$

and the variance of the contrast critical thinking–reading is very small.

The estimated college-level variance matrix for the centered parameterization is

$$\hat{\mathbf{\Sigma}}_A = \begin{bmatrix} .900 & & & \\ .014 & .012 & & \\ .045 & .006 & .009 & \\ .122 & -.002 & .016 & .056 \end{bmatrix}.$$

The standard errors for the sigmas (.95, .11, .09, .24) are (.11, .04, .03, .04). Even the variance for the critical thinking–reading contrast (.009) is significantly greater than 0; the data appear to contain abundant information about the variation of the skill subscores. The college-level variance matrix for the subscores, $\hat{\mathbf{\Omega}}_A = \mathbf{C}_4\hat{\mathbf{\Sigma}}_A\mathbf{C}_4$, is given in Table 5. Its eigenvalue decomposition confirms what can be observed from $\hat{\mathbf{\Sigma}}_A$: The total score, or a linear

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

TABLE 5
Eigenvalue decompositions of the estimated variance matrices for the short form data, administration, and skill subscores

Level		Eigenvalues	Eigenvectors
Examinee			
$\hat{\mathbf{\Omega}}_B =$	$\begin{bmatrix} 4.96 & 2.56 & 2.49 & 2.00 \\ & 5.30 & 2.42 & 2.16 \\ & & 4.96 & 2.11 \\ & & & 5.04 \end{bmatrix}$	11.95 3.15 2.72 2.43	(.50 .53 .50 .46) (.36 .27 .17 −.88) (.21 −.76 .62 −.03) (.76 −.27 −.58 −.12)
College			
$\hat{\mathbf{\Omega}}_A =$	$\begin{bmatrix} .80 & .82 & .85 & .91 \\ & .84 & .87 & .92 \\ & & .90 & .97 \\ & & & 1.07 \end{bmatrix}$	3.61 .032 .005 .001	(.47 .48 .50 .54) (.25 .55 .10 −.79) (.84 −.49 −.22 .10) (−.10 −.48 .83 −.26)

combination of the subscores very close to $(1, 1, 1, 1)$, represents the dominant dimension of the trait. The second dimension has a high loading on mathematics.

The critical administration size for the total score is $n_1^* = c \cdot tr(\hat{\mathbf{\Omega}}_B^*)/tr(\hat{\mathbf{\Omega}}_A) = 3.32c$. If the choice of $\alpha_1 = .10$ and $\alpha_2 = .90$ is adopted, then $c = 6.57$ and $n_1^* = 22$. Thus, a college has to administer the short form to at least 22 students for the total score differences among colleges to be observable. For representation of the second dimension, the linear contrast $(-1, -2, 0, 3)$ is used. The critical administration for the contrast is $n_2^* = 505$. Among the colleges taking part, only one (No. 33 in Table 1) had more examinees than n_2^*. Therefore, this linear contrast does not represent an observable trait. The critical administration sizes n_2^* for other contrasts close to the second eigenvalue of $\hat{\mathbf{\Sigma}}_A$ are also around 500; their exact choice has little influence on n_2^*. They correspond to unrealistically large administrations. I conclude that in the short form administration college mean subscores for the skills do not contain information useful for comparisons among colleges.

Subject-Area Subscores in the Short Form Administration

The fitted variance matrices for the examinee- and the college-level variation are given in Table 6. The eigenvalue decomposition of $\hat{\mathbf{\Omega}}_A$ suggests that the total score represents a dominant dimension of the trait underlying the subject-area subscores. The same critical administration size for the total score, n_1^*, is obtained as for the skill subscores, because the fitted variances for the total score in the two analyses are identical. The second

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

TABLE 6
Eigenvalue decompositions of the estimated variance matrices for the short form administration; subject-area subscores

Level		Eigenvalues	Eigenvectors		
Examinee					
$\hat{\mathbf{\Omega}}_B =$	$\begin{bmatrix} 7.95 & 4.13 & 3.97 \\ & 7.76 & 4.04 \\ & & 7.75 \end{bmatrix}$	15.92	(.58	.58	.57)
		3.88	(.70	.01	−.71)
		3.67	(.41	−.82	.41)
College					
$\hat{\mathbf{\Omega}}_A =$	$\begin{bmatrix} 1.41 & 1.51 & 1.56 \\ & 1.64 & 1.71 \\ & & 1.80 \end{bmatrix}$	4.82	(.54	.58	.61)
		.034	(.71	.08	−.70)
		.008	(.46	−.81	.37)

dimension of the trait corresponds to the contrast of natural sciences and humanities. The corresponding critical administration size is 552, larger than any administration in the pilot year. Therefore, only the total score means are worth reporting in the short form administration, and any comparison of subscores would only reflect the comparison of the true total score means.

Long Form Administration

The results from the long form data replicate our findings from the short form analysis. For example, the estimated college-level variance matrix in the centered parameterization is

$$\hat{\mathbf{\Sigma}}_A = \begin{bmatrix} 9.62 & & & \\ -.26 & .16 & & \\ -.15 & -.03 & .06 & \\ -.66 & .34 & .20 & 2.17 \end{bmatrix},$$

and the total score and the mathematics–reading contrast are reasonable choices for the first two dimensions of the trait for the subscores. The critical administration size for the total score is 7, and for the mathematics–reading contrast, 71. The third dimension of the trait can be represented by the contrast of critical thinking and writing, but its critical administration size is greater than 1,000. Therefore, in the long form administration it is useful to report two scores related to skills; the first three skills (reading, writing, and critical thinking) could be collapsed into a single category, thus simplifying the test design.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

Conditional Expectations of the Random Effects

Diagnosis of the model fit in ordinary regression is usually based on inspection of the residuals. In variance component models we have "residuals" associated with units at each level of the hierarchy—colleges (grades) and examinees. The separation of the total residual $y - \hat{M}$ into its components $\hat{A}$ and $\widehat{B + \epsilon}$ is obtained by considering the conditional expectations of the random effects given the parameter estimates. The formulae and their connection to the EM algorithm are discussed in Longford (1987). The college-level conditional expectations $\{\hat{\mathbf{A}}_i\}$ are also useful for comparison of the colleges that participated in the pilot year. Colleges at either extreme of the scale defined by the fitted variance for the component of the vector of residuals are of interest both for diagnostic and substantive reasons. Table 7 lists the residuals for the skill subscores in the short form administration for the average of the subscores and the mathematics–reading contrast.

Equality of the covariance structure of the subscores across the forms is the most important aspect of parallelness. Comparison of the columns in Table 7 (in particular for colleges with large numbers of examinees) provides evidence of parallelness of the three forms. Parallelness of the forms can be more formally explored by separate analyses for each form.

Concluding Remarks

The analyses indicate that the two-way matrix design of the test is not justified. For colleges using the short form administration, only the total score is worth reporting; any subscore, or a linear combination of subscores, is indistinguishable from a less reliable version of the total score. In the long form administration, there are two observable dimensions of the latent trait for the skill subscores; it is meaningful to consider two subtests—one containing all the mathematics items and the other all the other items of the test (reading, writing, and critical thinking).

We cannot, on the basis of the analyses, determine why the complex factor structure is not observable in the college-mean test subscores. On one hand, the items may be an imperfect representation of the domain indicated by the label, or they may be heavily contaminated by a "general" domain. This problem may be alleviated by careful review and content analysis of the test items. On the other hand, the true factorial structure underlying the knowledge/skill domains may be genuinely unidimensional, in which case the test cannot serve the intended purpose. We can only speculate that there may be a multivariate factorial structure underlying academic skills (and subject areas), but the traits of the skills defined in GENED are all subsumed in a single dimension.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

TABLE 7
Conditional expectations of the random effects A_{i1}^* and A_{i4}^*; average of subscores and mathematics–reading contrast

College No.	Short form							
	Form A		Form B		Form C		All forms	
	A_{i1}^*	A_{i4}^*	A_{i1}^*	A_{i4}^*	A_{i1}^*	A_{i4}^*	A_{i1}^*	A_{i4}^*
1[a]	2.33	.19	1.90	.16	2.70	.14	2.37	.21
2	.53	−.14	.74	−.04	.58	.01	.64	−.06
3	.21	.12	−.19	−.16	.04	−.01	.02	.01
4	−.92	−.04	−.84	−.04	−.77	−.04	−.87	−.01
5	−.48	.00	−.59	.10	−.76	−.06	−.65	.02
6	−1.05	−.11	−1.13	−.32	−.99	−.23	−1.09	.24
7	−.31	−.22	−.40	.22	.00	−.04	−.25	−.04
8[a]	.98	.00	.90	.10	.33	−.02	.80	.02
9[a]	1.40	.22	1.24	.29	1.58	.13	1.62	.22
10	.47	.10	−.15	−.03	.45	−.11	.26	−.07
11[a]	−.80	−.09	−1.21	−.11	−1.21	−.14	−1.25	−.17
12	.28	.18	.78	.14	.37	.12	.51	.23
13	−1.55	−.09	−1.38	.10	−1.22	−.04	−1.45	−.02
14	−.02	−.06	.53	−.50	.01	.01	.18	−.22
15	−1.23	−.33	−1.34	−.26	−1.58	.03	−1.44	−.26
16	1.25	−.02	1.05	.05	.90	.16	1.13	.09
17[a]	.22	−.05	−.50	−.12	−.01	−.15	−.12	−.17
18[a]	1.22	.09	−.36	−.09	.33	.06	.41	−.03
19[a]	−.48	−.06	.19	.02	−.09	−.02	−.14	−.04
20[a]	−.78	−.11	−.23	−.04	−.25	−.09	−.45	−.04
21	.20	−.08	−.15	.34	.05	.04	.04	.10
22	−.30	−.05	−.32	−.12	−.53	−.11	−.40	−.10
23	−.68	−.05	−.11	.12	−.73	−.13	−.54	−.07
24[a]	−.71	−.12	−.81	−.35	−.33	−.05	−.66	−.18
25[a]	−.58	.01	.16	.27	−.45	−.14	−.35	−.02
26[a]	−1.13	−.17	.38	−.01	−.42	.12	−.45	−.04
27	−.09	.02	−.48	−.07	.19	−.10	−.15	−.07
28[a]	−.27	.05	.11	−.28	−.35	.01	−.19	−.02
29	−1.20	−.22	−1.36	−.47	−1.50	−.31	−1.41	−.43
30	1.28	.38	1.60	.45	1.68	.33	1.55	.50
31	.30	−.09	−.04	−.14	.38	−.03	.22	−.12
32	.71	.16	.81	.18	.94	.01	.84	.12
33	.49	.42	.39	.48	.27	.46	.38	.54
34	−.27	.15	−.18	.10	.01	.18	−.15	.21

[a] Colleges with fewer than 100 examinees.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

The assumption of normality in our analysis is, of course, contentious. A form of sensitivity analysis can be carried out by applying transformations to the response variable and then fitting the appropriate variance component models. I have considered the arcsine, the logarithm, and the z transform ($(y - a)/b$, for form and subscore specific a and b). The eigenvalue decomposition of the estimated variance matrices and the critical administration sizes are remarkably resistant with respect to these transformations. For example, for the long form administration the total score is confirmed as the dominant dimension of the trait with each transformed response variable, and the critical administration sizes all round up to 7. The skill contrast mathematics–reading represents an observable dimension of the trait underlying the trait subscores with critical administration sizes in the range 71–73.

Another form of sensitivity analysis can be carried out by including in the analysis scores of all the examinees who participated in the short form administration (see Table 1). The results about the observable traits are replicated, although the critical administration sizes are increased by about 10%.

References

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73,* 43–56.

Goldstein, H. (1987). *Multilevel models in educational and social research.* New York and London: Oxford University Press and Charles Griffin.

Goldstein, H., & James, A. N. (1983). Efficient estimates for a multiple matrix sample design. *British Journal of Mathematical and Statistical Psychology, 36,* 167–174.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455–467.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72,* 320–340.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74,* 817–827.

Longford, N. T. (1988). *VARCL. Software for variance component analysis of data with hierarchically nested random effects (maximum likelihood). Manual.* Princeton, NJ: Educational Testing Service.

Longford, N. T. (1989). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R. D. Bock (Ed.), *Multilevel analysis of educational data.* San Diego: Academic Press.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22,* 259–267.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models (computer program).* Mooresville, IN: Scientific Software, Inc.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016

Searle, S. R. (1971). Topics in variance component estimation (a *Biometrics* invited paper). *Biometrics, 27,* 1–76.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling. An integrated theory for matrix sampling. *Journal of Educational Measurement, 14,* 343–399.

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

Author

NICHOLAS T. LONGFORD is Visiting Research Scientist, Educational Testing Service, 21-T, Rosedale Road, Princeton, NJ 08541, and Visiting Associate Professor, Department of Mathematics, UCLA, Los Angeles, CA 90024. He specializes in variance component analysis.

Downloaded from http://jebs.aera.net at PENNSYLVANIA STATE UNIV on February 20, 2016