*Research Report*

# Subscores for Institutions

Shelby J. Haberman

Sandip Sinharay

Gautam Puhan

Research & Development

June 2006
RR-06-13

# Subscores for Institutions

Shelby J. Haberman, Sandip Sinharay, and Gautam Puhan
ETS, Princeton, NJ

June 2006

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**

Recently, there has been an increasing level of interest in reporting subscores. This paper examines the issue of reporting subscores at an aggregate level, especially at the level of institutions that the examinees belong to. A series of statistical analyses is suggested to determine when subscores at the institutional level have any added value over the total scores. The methods are applied to two operational data sets. For the data under study, the results provide little support in favor of reporting subscores for either examinees or institutions.

Key words: KR-20, proportional reduction in error, reliability

**Acknowledgments**

The authors thank David Wright, Neil Dorans, Paul Holland, Jiahe Qian, Tim Moses, and Dan Eignor for useful advice, and Kim Fryer for help with proofreading.

## 1. Introduction

What are subscores and why are they desirable? Educational and psychological tests often have different subsections based on content categories or blueprints. For example, a test on mathematics knowledge may have subsections on algebra and geometry. Similarly, a test of general ability can have subsections on mathematics, reading, and writing. Scores assigned to these subsections are commonly known as *subscores*. Subscores resulting from the administration of tests with high-stakes outcomes are desirable for at least two important reasons. First, failing candidates want to know their strengths and weaknesses in different content areas to plan for future remedial work. Second, states and academic institutions such as colleges and universities want a profile of performance for their graduates to better evaluate their training and focus on areas that need instructional improvement (Haladyna & Kramer, 2004).

Despite this apparent usefulness of subscores, certain important factors must be considered before making a decision on whether to report subscores at either the individual or institutional level. Although many tests are designed to cover a broad domain, and the total test score is considered to be a composite of different abilities measured by different subsections, it is debatable whether a subsection with fewer items than the total test can be viewed as a mini-test that can precisely measure a unique ability.

Haberman (2005) argued that a subscore may be considered useful only when it provides a more accurate measure of the construct being measured than is provided by the total score. Wainer et al. (2001) suggested that a test used for diagnostic purposes must yield scores that are reliable both for the total test and for the subscores associated with specific subsections or content areas. Furthermore, to be useful for diagnostic purposes, the subscores must focus as closely as possible on the content areas in which the examinee may be having difficulty. Finally, Tate (2004) has emphasized the importance of ensuring reasonable subscore performance in terms of high reliability and validity to minimize incorrect instructional and remediation decisions.

From the above review, it is apparent that the quality of the subscores must be assessed before considering score reporting at the subscore level. It also serves as an important reminder of the following: Just as inaccurate information at the total test score level

can lead to inaccurate pass and fail decisions with damaging consequences to both the testing programs and test takers, inaccurate information at the subscore level can also lead to incorrect remediation decisions resulting in large and needless expense for state or institutions.

The studies cited above have mainly focused on the use of subscores at the examinee level (ignoring any information from institutions or state agencies). However, as mentioned earlier, subscores at the institutional level could also be of interest for planning remedial and training programs. Moreover, subscores may not offer added value at the examinee level but may do so at the institutional level. For example, it is possible that the true subscores underlying subtests A and B are perfectly correlated (in which case subscores do not have any added value) within each institution in a population of institutions, but the institution means may have a lower correlation (in which case subscores may have any added value). Therefore it is important to examine the adequacy of subscores at the institutional level.

Institutional level subscoring can prove to be useful when there is considerable variation in test performance between different institutions. For example, on a typical Praxis™ test, there are several user states and institutions that have examinee populations that may differ considerably in terms of the measured ability. Variation in test scores at the state or institutional level may justify investigating the use of subscores at these levels.

This paper performs a thorough analysis to determine when subscores at the institutional level have any added value over the total score for the tests concerned. First, an individual-level analysis is performed, as in Haberman (2005), that examines whether individual-level subscoring is justified. Then, a similar analysis is developed to determine whether reporting of test subscores is justified at an institutional level. The approach used involves an analysis of proportional reduction in error variance in estimation of true institutional subscore means. The basic criterion applied is that the mean subscore for examinees from an institution is not worth reporting if the true institutional mean is more accurately predicted by the mean total score of examinees from the institution than by the mean total subscore of examinees from the institution. All the computations involved are quite simple and use popular software programs, so that operational implementation is straightforward for the suggested methods. In addition, the methods can be directly

applied if score reporting is considered at a different level of aggregation, say by states rather than by institutions.

At the institutional level, the analysis of appropriate reporting practice depends on the number of examinees from the institution who take the test under study. Although cases certainly can arise in which no evidence exists that reporting of subscores is ever appropriate, it is quite common for analysis to reveal that subscores may be usefully reported if the number of examinees from the institution is sufficiently large, but reporting is inappropriate if the number of examinees in the institution is relatively small. Thus this report considers minimum sample-size requirements for reporting of means of subscores of examinees from a particular institution.

Longford (1990) studied the issue of reporting subscores at the college level, performing a multilevel variance component analysis on data at the pilot stage of development of a test, so that there were issues like voluntary participation of colleges, motivation of students, the lack of many colleges, and a model assumption (of normality) that was admitted to be "contentious" (p. 111). Our study is different from that of Longford (1990) in three basic ways: (a) We perform an analysis using a measure that is very close to the classical reliability measure—hence the method is more intuitive, (b) there are no contentious model assumptions, and (c) we analyze large operational test data sets from a test with high-stakes outcomes that involves a large number of institutions.

Section 2 describes the methodology involved, and Section 3 discusses the results obtained when the methodology is applied to two data sets from a basic skills test with high-stakes outcomes belonging to the Praxis series. Discussions and conclusions are provided in Section 4.

## 2. Methodology and Analysis

This section describes our step-by-step approach for determining whether, when, and how to report institutional-level subscores. We begin with a description of an examinee-level analysis in Section 2 to determine if the examinee-level subscores offer any added value over the total scores. This section closely follows Haberman (2005). However, the question of the usefulness of institutional-level subscores is different from the question of usefulness

of the examinee-level subscores. Sections 2.2 and 2.3 describe analyses required at the institutional level.

### *2.1 Examinee-Level Analysis of Mean-Squared Error*

At the examinee level, analysis involves the observed subscore $s$, the true subscore $s_t$, the observed total score $x$, and the true total score $x_t$. It is assumed that $s_t$, $x_t$, $s - s_t$, and $x - x_t$ all have positive variances. As usual in classical test theory, $s$ and $s_t$ have common mean $E(s)$, $x$ and $x_t$ have common mean $E(x)$, and the true scores $s_t$ and $x_t$ are uncorrelated with the errors $s - s_t$ and $x - x_t$. For random variables $u$ and $v$ with finite means and variances, the expectation of $u$ is $E(u)$, the standard deviation of $u$ is $\sigma(u)$, the variance of $u$ is $\sigma^2(u)$, the covariance of $u$ and $v$ is $c(u, v)$, the correlation of $u$ and $v$ is $\rho(u, v)$, and the squared correlation of $u$ and $v$ is $\rho^2(u, v)$. It is assumed that the true subscore $s_t$ and true total score $x_t$ are not collinear, so that $|\rho(s_t, x_t)|$ is less than 1. This assumption also implies that $|\rho(s, x)| < 1$.

The following quantities for the examinee-level data are used (ignoring the information on the institutions) to determine if the examinee-level subscores have any additional value over their total scores:

1. The reliability $\rho^2(x_t, x)$ of the total test score $x$
2. The reliability $\rho^2(s_t, s)$ of subscore $s$
3. The squared correlation $\rho^2(s_t, x_t)$ of the true score $s_t$ and the true total score $x_t$

The KR-20 approach is typically employed to estimate the reliabilities of $s$ and $x$ (Kuder & Richardson, 1937). The squared correlation $\rho^2(s_t, x)$ of the true subscore $s_t$ and the observed total score $x$ is then given by

$$\rho^2(s_t, x) = \rho^2(s_t, x_t)\rho^2(x_t, x)\cdot \tag{1}$$

For details on computation of $\rho^2(s_t, x_t)$, see Haberman (2005).

In the analysis of Haberman (2005), three basic approaches to prediction of the true score $s_t$ are considered. In the first or trivial approach, $s_t$ is predicted by the constant

$E(s)$, so that the mean-squared error is $\sigma^2(s_t)$. In the second approach, one based on the observed subscore $s$, the linear regression

$$\hat{s} = E(s) + \rho^2(s_t, s)[s - E(s)]$$

of $s_t$ on $s$ predicts $s_t$, and the mean-squared error is $\sigma^2(s_t)[1 - \rho^2(s_t, s)]$. In the third approach, based on the observed total score $x$, the linear regression

$$\hat{s}_x = E(s) + \rho(s_t, x)[\sigma(s_t)/\sigma(x)][x - E(x)]$$

of $s_t$ on $x$ predicts $s_t$, and the mean-squared error is $\sigma^2(s_t)[1 - \rho^2(s_t, x)]$. Relative to use of $E(s)$, $\rho^2(s_t, s)$ is the proportional reduction of mean-squared error from use of the estimate $\hat{s}$ based on the observed subscore, while $\rho^2(s_t, x)$ is the proportional reduction in mean-squared error from use of the estimate $\hat{s}_x$ based on the observed total score. Haberman (2005) argues on the basis of these results that subscores should not be reported if $\rho^2(s_t, s)$ is less than $\rho^2(s_t, x)$, for the true subscore is better approximated by use of the total observed score rather than the observed subscore.

Haberman (2005) also considers an option of reporting an estimate of the true subscore $s_t$ based on the linear regression $\hat{s}_a$ of $s_t$ on both the observed subscore $s$ and the observed total score $x$. The study of proportional reduction of mean-squared error also requires the correlation $\rho(s, x)$ of the subscore $s$ and the total score $x$. The regression is

$$\hat{s}_a = E(s) + \beta[s - E(s)] + \gamma[x - E(x)],$$

where

$$\gamma = \frac{\sigma(s)}{\sigma(x)}\rho(s_t, s)\tau,$$

$$\tau = \frac{\rho(x_t, x)\rho(s_t, x_t) - \rho(s, x)\rho(s_t, s)}{1 - \rho^2(s, x)},$$

and

$$\beta = \rho(s_t, s)[\rho(s_t, s) - \rho(s, x)\tau].$$

The mean-squared error is then $\sigma^2(s_t)\{1 - \rho^2(s_t, s) - \tau^2[1 - \rho^2(s, x)]\}$, so that the proportional reduction in mean-squared error relative to $E(s)$ is

$$\rho^2(s_t, \hat{s}_a) = \rho^2(s_t, s) + \tau^2[1 - \rho^2(s, x)].$$

Wainer et al. (2001) discusses the idea of *augmentation*, which means stabilizing the subscores by augmenting data from any particular subscore with information obtained from the other subscores. One can perform augmentation using the approach of Haberman (2005) by considering a linear regression of $s_t$ on other observed subscores $u_k$, $1 \leq k \leq r$. In the most trivial case, $r = 1$ and $u_1 = x - s$ is the total score minus the subscore. Let the true score for $u_k$ be $u_{kt}$. Assume that $s_t$ is not a linear function of $s$ and $u_k$, $s$ is not a linear function of the $u_k$, and no $u_j$ is a linear function of the remaining $u_k$. Then one computes

$$\hat{s}_u = E(s) + \beta_u[s - E(s)] + \sum_{k=1}^{r} \gamma_k[u_k - E(u_k)],$$

where

$$\gamma_k = [\sigma(s)/\sigma(u_k)]\rho(s_t, s)\tau_k,$$

$$\beta_u = \rho(s_t, s)\left[\rho(s_t, s) - \sum_{k=1}^{r} \tau_k \rho(s, u_k)\right]$$

and

$$\sum_{k=1}^{r}[\rho(u_j, u_k) - \rho(s, u_j)\rho(s, u_k)]\tau_k = \rho(s_t, u_j) - \rho(s_t, s)\rho(s, u_j)$$

for $1 \leq j \leq r$. The mean-squared error is

$$\sigma^2(s_t)\left\{1 - \rho^2(s_t, s) - \sum_{k=1}^{r} \tau_k[\rho(s_t, u_k) - \rho(s_t, s)\rho(s, u_k)]\right\},$$

so that, relative to $E(s)$, the proportional reduction in mean-squared error is

$$\rho^2(s_t, \hat{s}_u) = \rho^2(s_t, s) + \sum_{k=1}^{r} \tau_k[\rho(s_t, u_k) - \rho(s_t, s)\rho(s, u_k)].$$

This generalization appears to offer very little added benefit for the data considered in this paper. In the trivial case in which $r = 1$ and $u_1 = x - s$, $\hat{s}_u$ is the same as $\hat{s}_a$.

### *2.2 Institutional-Level Analysis of Mean-Squared Error*

At the institutional level, the analysis of Section 2.2 must be modified by decomposition of scores and subscores into institutional and individual components. Thus subscore $s$ has the decomposition $s = s_I + s_e$, where $s_I$ (the component for the institution of the examinee) is the same for each examinee in that institution and has mean $E(s)$ and variance

$\sigma^2(s_I) > 0$. The score $x$ has the decomposition $x = x_I + x_e$, where $x_I$ (the component for the institution of the examinee) is the same for each examinee in that institution and has mean $E(x)$ and variance $\sigma^2(x_I) > 0$. The residual examinee subscore $s_e = s - s_I$ within institution has mean 0, variance $\sigma^2(s_e) > 0$, and is uncorrelated with the institutional means $s_I$ and $x_I$. The residual examinee total score $x_e = x - x_I$ within institution has mean 0, variance $\sigma^2(x_e) > 0$, and is uncorrelated with $s_I$ and $x_I$. The analysis is not directly concerned with the true scores and errors of Section 2.2, but it should be noted that, under classical assumptions, $s_e = (s_t - s_I) + (s - s_t)$ and $s_t - s_I$ and $s - s_t$ are uncorrelated, so that $s_t - s_I$ has mean 0 and variance $\sigma^2(s_t) - \sigma^2(s - s_t)$. In like fashion, $x_e = (x_t - x_I) + (x - x_t)$ and $x_t - x_I$ and $x - x_t$ are uncorrelated, so that $x_t - x_I$ has mean 0 and variance $\sigma^2(x_t) - \sigma^2(x - x_t)$. It is assumed that $s_I$ and $x_I$ do not have a correlation of 1 or $-1$.

If $n$ examinees are observed from a given institution and if $\bar{s}$ is the average subscore for examinees from that institution, then $\bar{s} = s_I + \bar{s}_e$, where $\bar{s}_e$ is uncorrelated with $s_I$ and $x_I$ and has mean 0 and variance $\sigma^2(s_e)/n$. Thus $\bar{s}$ has variance $\sigma^2(s_I) + \sigma^2(s_e)/n$. For the institution, the squared correlation of the institutional mean $s_I$ and the average $\bar{s}$ is then the reliability

$$\rho^2(s_I, \bar{s}) = \frac{\sigma^2(s_I)}{\sigma^2(s_I) + \sigma^2(s_e)/n}. \tag{2}$$

Similarly, if $\bar{x}$ is the average total score for examinees from that institution, then $\bar{x} = x_I + \bar{x}_e$, where $\bar{x}_e$ is uncorrelated with $s_I$ and $x_I$ and has mean 0 and variance $\sigma^2(x_e)/n$. Thus $\bar{x}$ has variance $\sigma^2(x_I) + \sigma^2(x_e)/n$. The squared correlation of the institutional mean $x_I$ and the average $\bar{x}$ is then the reliability

$$\rho^2(x_I, \bar{x}) = \frac{\sigma^2(x_I)}{\sigma^2(x_I) + \sigma^2(x_e)/n}.$$

Analysis also requires the squared correlation $\rho^2(s_I, x_I)$ of institutional mean subscore $s_I$ and institutional mean score $x_I$. This calculation may be accomplished by multivariate analysis of variance, as is shown in Section 2.3. Given this squared correlation,

$$\rho^2(s_I, \bar{x}) = \rho^2(s_I, x_I)\rho^2(x_I, \bar{x}). \tag{3}$$

Analogous to results in Section 2.2, if $E(s) = E(s_t)$ is used to predict $s_I$, then the mean-squared error is $\sigma^2(s_I)$. If

$$\hat{s}_I = E(s) + \rho^2(s_I, \bar{s})[\bar{s} - E(s)],$$

the linear regression of $s_I$ on $\bar{s}$, is used to predict $s_I$, then the mean-squared error is $\sigma^2(s_I)[1 - \rho^2(s_I, \bar{s})]$. If linear regression of $s_I$ on $\bar{x}$ is used to predict $s_I$ by use of

$$\hat{s}_{Ix} = E(s) + \rho(s_I, \bar{x})[\sigma(s_I)/\sigma(\bar{x})][\bar{x} - E(x)],$$

then the mean-squared error is $\sigma^2(s_I)[1-\rho^2(s_I, \bar{x})]$. Relative to use of $E(s)$, the proportional reduction of mean-squared error from use of the linear regression based on $\bar{s}$ is the reliability $\rho^2(s_I, \bar{s})$, while the proportional reduction in mean-squared error from use of the linear regression based on $\bar{x}$ is $\rho^2(s_I, \bar{x})$. Thus a basic requirement for reporting an institutional subscore is that $\rho^2(s_I, \bar{s})$ be greater than $\rho^2(s_I, \bar{x})$. Otherwise, the average total score for the institution predicts the institutional subscore mean $s_I$ better than does the average subscore for the institution. For any nontrivial estimation of the institutional mean $s_I$, it is clearly best if the number of examinees $n$ for the institution is relatively large. In addition, for $n$ sufficiently large, $\bar{s}$ is a better predictor of $s_I$ than is $\bar{x}$. The problem in practice is to ascertain when the sample size for an institution is large enough for the subscores to be worth reporting.

Combined use of $\bar{s}$ and $\bar{x}$ to predict $s_I$ is also possible by use of the same argument as in Haberman (2005). In this case, the regression of $s_I$ on $\bar{s}$ and $\bar{x}$ is

$$\hat{s}_{Ia} = E(s) + \beta_I[\bar{s} - E(s)] + \gamma_I[\bar{x} - E(x)],$$

where

$$\gamma_I = \frac{\sigma(s_I)}{\sigma x_I}\rho(s_I, \bar{s})\tau_I,$$

$$\tau_I = \frac{\rho(x_I, \bar{x})\rho(s_I, x_I) - \rho(\bar{s}, \bar{x})\rho(s_I, \bar{s})}{1 - \rho^2(\bar{s}, \bar{x})},$$

and

$$\beta_I = \rho(s_I, \bar{s})[\rho(s_I, \bar{s}) - \rho(\bar{s}, \bar{x})\tau_I].$$

The mean-squared error is then $\sigma^2(s_I)[1-\rho^2(s_I,\bar{s})-\tau_I^2[1-\rho^2(\bar{s},\bar{x})]$, so that the proportional reduction in mean-squared error relative to $E(s)$ is

$$\rho^2(s_I,\hat{s}_{Ia})=\rho^2(s_I,\bar{s})+\tau_I^2[1-\rho^2(\bar{s},\bar{x})]\cdot \tag{4}$$

As in the augmentation approach of Wainer et al. (2001), one may consider the decomposition $u_k = u_{Ik} + u_{ek}$, where $s_I$ and $u_{Ik}$ are uncorrelated with $s_e$ and $u_{ek}$. Given standard assumptions to prevent collinearity of predictors, $s_I$ is predicted by the institutional means $\bar{s}$ for $s$ and $\bar{u}_k$ for $u_k$. The predictor $\hat{s}_{Iu}$ is then

$$\hat{s}_{Iu}=E(s)+\beta_{Iu}[\bar{s}-E(s)]+\sum_{k=1}^{r}\gamma_{Ik}[\bar{u}_k-E(u_k)],$$

where

$$\gamma_{Ik}=[\sigma s_I/\sigma(u_{Ik})]\tau_{Ik},$$

$$\beta_{Iu}=\rho(s_I,\bar{s})\left[\rho(s_I,\bar{s})-\sum_{k=1}^{r}\tau_{Ik}\rho(\bar{s},\bar{u}_k)\right]$$

and

$$\sum_{k=1}^{r}[\rho(\bar{u}_j,\bar{u}_k)-\rho(\bar{s},\bar{u}_j)\rho(\bar{s},\bar{u}_k)]\tau_{Ik}=\rho(s_I,\bar{u}_j)-\rho(s_I,\bar{s})\rho(\bar{s},\bar{u}_j)$$

for $1 \le j \le r$. The mean-squared error is

$$\sigma^2(s_I)\left\{1-\rho^2(s_I,\bar{s})-\sum_{k=1}^{r}\tau_{Ik}[\rho(s_I,\bar{u}_k)-\rho(s_I,\bar{s})\rho(\bar{s},\bar{u}_k)]\right\},$$

so that, relative to $E(s)$, the proportional reduction in mean-squared error is

$$\rho^2(s_I,\hat{s}_{Iu})=\rho^2(s_I,\bar{s})+\sum_{k=1}^{r}\tau_{Ik}[\rho(s_I,\bar{u}_k)-\rho(s_I,\bar{s})\rho(\bar{s},\bar{u}_k)].$$

### *2.3 Institutional-Level Estimation Procedure*

To estimate the means, variances, and correlations required for an institutional analysis requires mean squares and mean cross products customarily associated with a one-way multivariate analysis of variance (MANOVA) with dependent variables for the observed total score and observed subscore. Let a sample be available with $n_j$ scores from institution $j$, $1 \le j \le J$. Let $N$ be the total number of examinees from all institutions. Assume that $N > J$. For examinee $i$ of institution $j$, let the total score be $x_{ij}$, and let the subscore be

$s_{ij}$. Let $\bar{x}_j$ be the average total score from institution $j$, and let $\bar{s}_j$ be the average subscore from institution $j$. Let $\bar{x}_.$ be the mean total score for all examinees, and let $\bar{s}_.$ be the mean subscore for all examinees. Let the within-institution mean square for total score be

$$M_{xxe} = (N - J)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2,$$

let the within-institution mean square for subscore be

$$M_{sse} = (N - J)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (s_{ij} - \bar{s}_j)^2,$$

and let the within-institution mean cross product for subscore and total score be

$$M_{sxe} = (N - J)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (s_{ij} - \bar{s}_j)(x_{ij} - \bar{x}_j).$$

Let $\bar{x}$, the mean total score for all examinees, be used to estimate $E(x)$, and let $\bar{s}$, the mean subscore for all examinees, be used to estimate $E(s)$. Then the between-institution mean square for the total score is

$$M_{xxI} = (J - 1)^{-1} \sum_{j=1}^{J} n_j (\bar{x}_j - \bar{x}_.)^2,$$

the between-institution mean square for the subscore is

$$M_{ssI} = (J - 1)^{-1} \sum_{j=1}^{J} n_j (\bar{s}_j - \bar{s}_.)^2,$$

and the between-institution mean cross product for subscore and total score is

$$M_{sxI} = (J - 1)^{-1} \sum_{j=1}^{J} n_j (\bar{s}_j - \bar{s}_.)(\bar{x}_j - \bar{x}_.).$$

As in Snedecor and Cochran (1989), $\sigma^2(s_e)$ is normally estimated by $\hat{\sigma}^2(s_e) = M_{sse}$, and $\sigma^2(x_e)$ is normally estimated by $\hat{\sigma}^2(x_e) = M_{xxe}$. For the remaining required estimates, let $n$ denote the number of examinees from an institution, let

$$C = 1 - \sum_{j=1}^{J} (n_j/N)^2$$

measure dispersion of examinees across institutions (Gini, 1912), and let

$$K = NC/(J-1).$$

Note that $C \geq (J-1)/J$, with equality only if all $n_j$ are equal, and $K \geq N/J$, with equality only if all $n_j$ are equal. Then $\sigma^2(s_I)$ has estimate

$$\hat{\sigma}^2(s_I) = K^{-1}(M_{ssI} - M_{sse}),$$

$\sigma^2(x_I)$ has estimate

$$\hat{\sigma}^2(x_I) = K^{-1}(M_{xxI} - M_{xxe}),$$

$\sigma^2(\bar{s})$ has estimate

$$\hat{\sigma}^2(\bar{s}) = \hat{\sigma}^2(s_I) + \hat{\sigma}^2(s_e)/n,$$

$\sigma^2(\bar{x})$ has estimate

$$\hat{\sigma}^2(\bar{x}) = \hat{\sigma}^2(x_I) + \hat{\sigma}^2(x_e)/n,$$

the covariance $c(s_e, x_e)$ of $s_e$ and $x_e$ has estimate

$$\hat{c}(s_e, x_e) = M_{sxe},$$

the covariance $c(s_I, x_I)$ of $s_I$ and $x_I$ has estimate

$$\hat{c}(s_I, x_I) = K^{-1}(M_{sxI} - M_{sxe}),$$

the covariance $c(\bar{s}, \bar{x})$ of $\bar{s}$ and $\bar{x}$ has estimate

$$\hat{c}(\bar{s}, \bar{x}) = \hat{c}(s_I, x_I) + \hat{c}(s_e, x_e)/n,$$

the covariance $c(s_I, \bar{x})$ of $s_I$ and $\bar{x}$ has estimate

$$\hat{c}(s_I, \bar{x}) = \hat{c}(s_I, x_I),$$

$\rho^2(s_I, \bar{s})$ has estimate

$$\hat{\rho}^2(s_I, \bar{s}) = \hat{\sigma}^2(s_I)/\hat{\sigma}^2(\bar{s}),$$

$\rho^2(x_I, \bar{x})$ has estimate

$$\hat{\rho}^2(x_I, \bar{x}) = \hat{\sigma}^2(x_I)/\hat{\sigma}^2(\bar{x}),$$

$\rho(s_I, x_I)$ has estimate

$$\hat{\rho}(s_I, x_I) = \hat{c}(s_I, x_I)/[\hat{\sigma}(s_I)\hat{\sigma}(x_I)],$$

$\rho(\bar{s}, \bar{x})$ has estimate

$$\hat{\rho}(\bar{s}, \bar{x}) = \hat{c}(\bar{s}, \bar{x})/[\hat{\sigma}(\bar{s})\hat{\sigma}(\bar{x})],$$

$\rho(s_I, \bar{x})$ has estimate

$$\hat{\rho}(s_I, \bar{x}) = \hat{c}(s_I, \bar{x})/[\hat{\sigma}(s_I)\hat{\sigma}(\bar{x})],$$

and $\tau_I$ has estimate

$$\hat{\tau}_I = \frac{\hat{\rho}(x_I, \bar{x})\hat{\rho}(s_I, x_I) - \hat{\rho}(\bar{s}, \bar{x})\hat{\rho}(s_I, \bar{s})}{1 - \hat{\rho}^2(\bar{s}, \bar{x})}.$$

Results for augmentation are derived by very similar arguments, so that details are omitted.

Some changes in procedure are necessary in special cases. The following simple rules appear adequate in practice, although the approach of Bock and Peterson (1975) is worth consideration even if the ideal condition that all $n_j$ are equal does not hold. If $M_{ssI} \leq M_{sse}$, then no evidence exists that $s_I$ has a positive variance, so that $\hat{s}_I$, $\hat{s}_{Ix}$, $\hat{s}_{Ia}$, and $\hat{s}_{Iu}$ are all approximated by $\bar{s}_\cdot$, and all proportional reductions in mean-squared error may be approximated by 0. If $M_{ssI} > M_{sse}$ but $M_{xxI} \leq M_{xxe}$, then no evidence exists that $x_I$ has a positive variance, so that $\hat{s}_{Ix}$ is approximated by $\bar{s}_\cdot$, and $\hat{s}_{Ia}$ and $\hat{s}_I$ have the same approximation. Thus the estimated proportional reduction in mean-squared error for $\hat{s}_{Ix}$ is estimated by 0, and the proportional reduction in mean-squared error for $\hat{s}_I$ and $\hat{s}_{Ia}$ are estimated to be the same. If $M_{ssI} > M_{sse}$ and $M_{xxI} > M_{xxe}$ but

$$(M_{sxI} - M_{sxe})^2 \geq (M_{ssI} - M_{sse})(M_{xxI} - M_{xxe}),$$

so that the normal estimate of $\rho^2(s_I, x_I)$ is greater than or equal to 1, then no evidence exists that $s_I$ is not a linear function of $x_I$. In this instance, $\hat{s}_{Ia}$ and $\hat{s}_{Ix}$ are estimated to be the same, $\hat{c}(s_I, x_I)$ is set to $\hat{\sigma}(s_I)\hat{\sigma}(x_I)$, and $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ and $\hat{\rho}^2(s_I, \bar{x})$ are set equal to $\hat{\rho}^2(x_I, \bar{x})$.

Required computations may be performed with the help of the SAS NESTED and GLM procedures (SAS Institute, 1996).

## 3. Results

The subscore analyses from Section 2 were applied to two administrations (forms) of a basic skills test belonging to the PRAXIS series. This test is designed for prospective and practicing paraprofessionals (i.e., teacher's aides) and measures skills and knowledge in reading, mathematics, and writing, as well as the ability to apply those skills and knowledge to aid in classroom instruction. Results were initially considered for six different subscores, namely theory and application of mathematics, theory and application of reading, and theory and application of writing. Although the results for six subscores were of primary interest, we further examined the results for the case with the three subscores for writing, mathematics, and reading that were obtained by pooling the theory and application portions of each of the three content areas. A final analysis pooled the reading and writing parts into one verbal subscore and retained the mathematics subscore. For each of the data sets, about a fourth of the examinees did not report their institutions. As a consequence, these examinees were removed from the analysis. The precise effect of this omission cannot be readily determined. Even after removing these examinees, the number of examinees for the two test forms were 3,240 and 2,497, respectively. The respective number of institutions were 712 and 654. The number of students $n_j$ in an institution $j$ ranged from 1 to 160 in these data, with the median size being 2 for both test forms, the 75th percentile being 4 for both test forms, the 95th percentile being 16 and 14 for the respective test forms, and the 99th percentiles being 41 and 28 for the respective test forms. Given these numbers, the number of institutions for which any score reports are possible is clearly quite limited unless reports combine more than one administration

### *3.1 Examinee-Level Analysis of Mean-Squared Error*

For the total score, the reliability of both the test forms was 0.94. The first two rows of Tables 1, 2, and 3 show the estimates of subscore reliability $\rho^2(s_t, s)$ and the proportional reduction $\rho^2(s_t, x)$ given by (1), both from individual-level analysis and expressed as percentages. The values indicate that the correlation of the true subscores are substantially higher with the observed total score than with the observed subscores, so that individual-level subscores should not be reported. In addition, the eigenvalues were

computed from the $6 \times 6$ estimated correlation matrix of the individual subscores. Figure 1 shows the corresponding scree plots (Cattell, 1956) for the two test forms. The figure strongly suggests that a single composite score exists such that each subscore can be very well approximated by use of a linear transformation of the composite score.

The results should not come as a big surprise as other studies also found subscores to have little added value. For instance, Harris and Hanson (1991) found subscores to have little added value for the English and mathematics tests from the P-ACT+ examination, and Haberman (2005) found subscores to have little added value for the SAT® I verbal and mathematics examinations.

### *3.2 Institutional Analysis*

At the institutional level, results were obtained for numbers $n$ of examinees per institution of 30, 100, and 150. Because the maximum number of students in an institution is 160, the upper bound of 150 appeared reasonable for the application. Tables 1, 2, and 3 show the proportional reductions in mean-squared error for these values of $n$ for six subscores, three subscores, and two subscores, respectively.

The last nine rows of the table show, for $n = 30, 100$, and 150, the values of the institutional level proportional reductions (expressed as percentages) discussed earlier and given by (3), (2), and (4), respectively.

Figure 2 compares the proportional reductions of mean-squared error at the institutional level for observed means of total scores and for observed means of subscores for six subscores, three subscores, and two subscores for each of the two test forms.

The tables and the figure reveal the following:

- On several occasions, $M_{ssI} > M_{sse}$, $M_{xxI} > M_{xxe}$, and

$$(M_{sxI} - M_{sxe})^2 \geq (M_{ssI} - M_{sse})(M_{xxI} - M_{xxe}),$$

  so that $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ and $\hat{\rho}^2(s_I, \bar{x})$ are set equal to $\hat{\rho}^2(x_I, \bar{x})$. This indicates rather small between-institution variation.

- The criterion of mean-squared error consistently favors prediction of institutional subscore means by observed institutional total score means rather than by observed in-

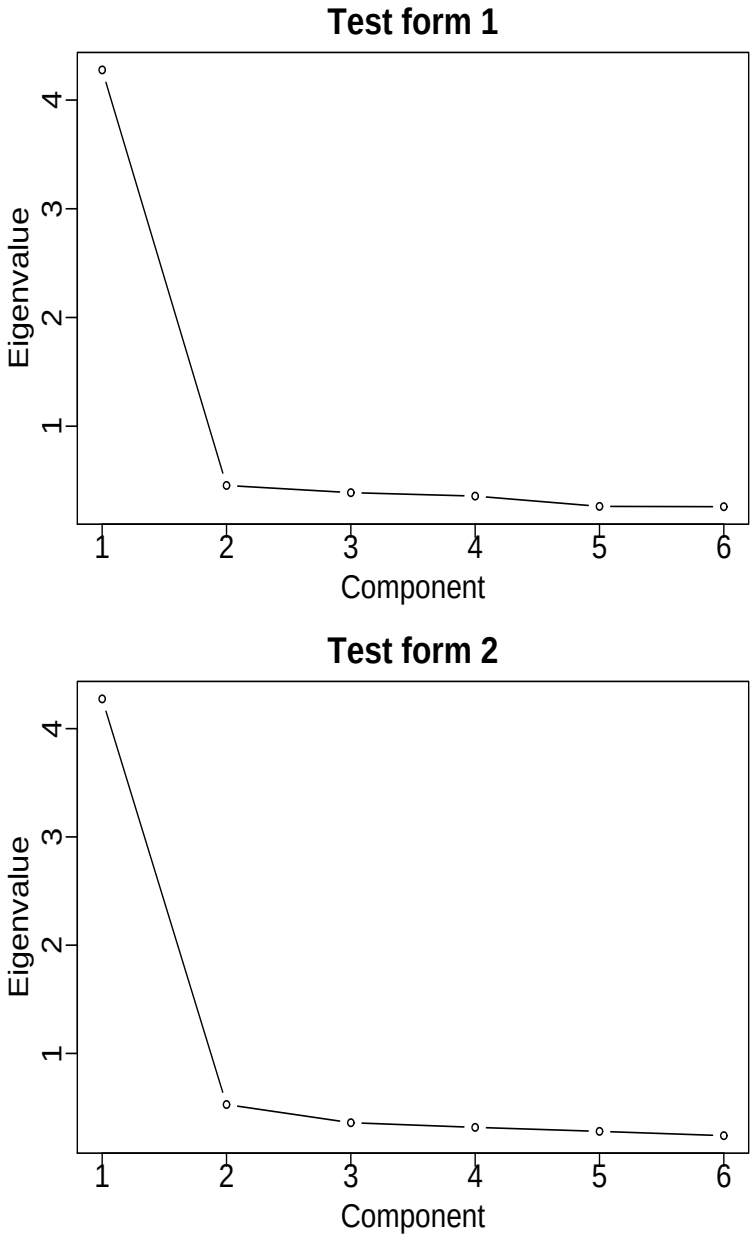


***Figure* 1. Scree plots for the two test forms.**

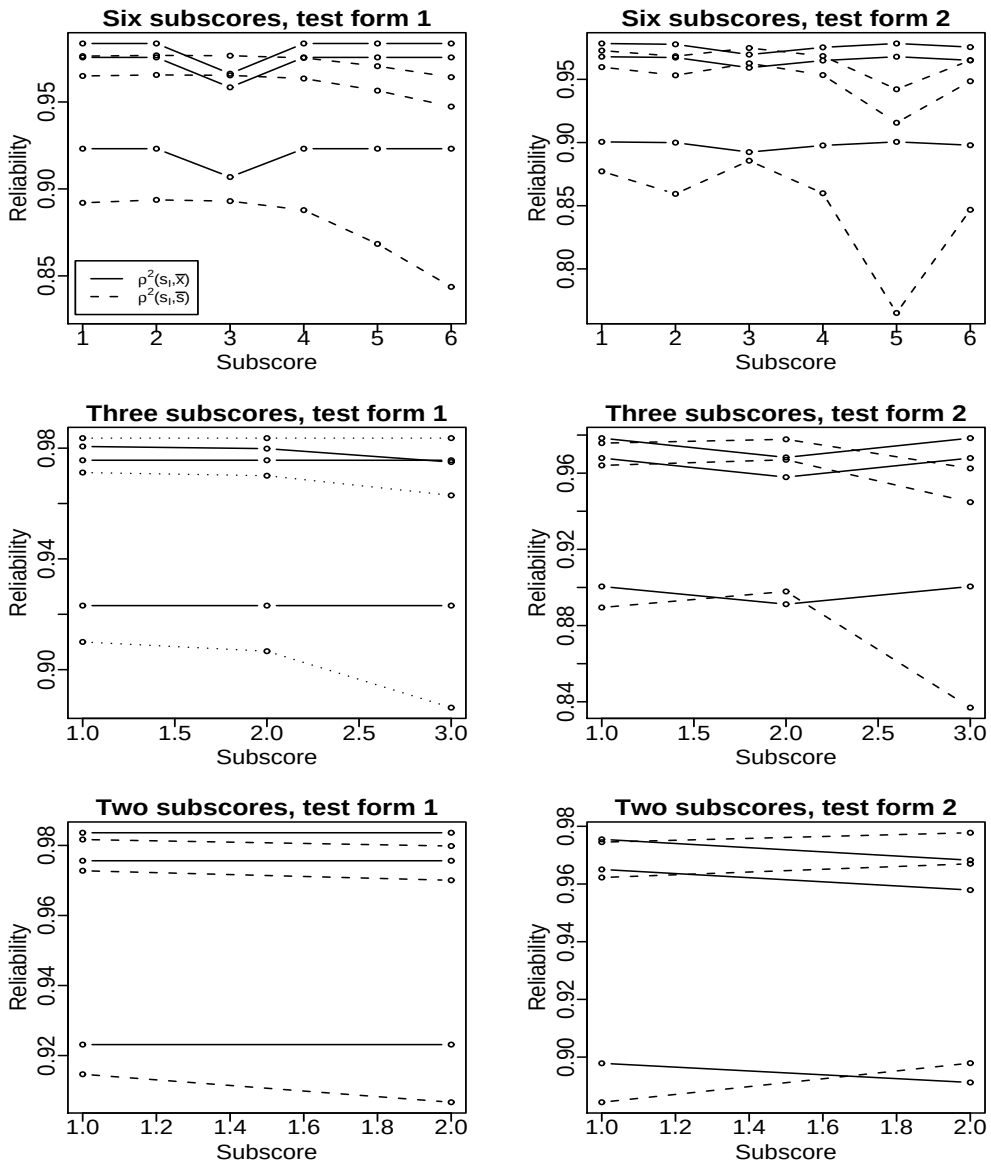


***Figure* 2. Comparison of proportional reduction of mean-squared error of institutional-level observed total scores and institutional-level observed subscores.**

*Note.* In any plot, the three solid lines show the proportional reduction in mean squared error of institutional-level observed total scores, a lower line indicating smaller institution size, and the three dashed lines show the proportional reduction of mean-squared error of institutional-level observed subscores, a lower line indicating smaller institution size.

**Table 1.**
***Percent Reduction (*** $100 \times$ ***Proportional Reduction)***
***in Mean-Squared Error With Six Subscores***

| | $n$ | Test form 1 | | | | | | Test form 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Subscore | | | | | | Subscore | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| $\hat{\rho}^2(s_t, s)$ | | 77 | 71 | 77 | 73 | 75 | 74 | 78 | 75 | 79 | 58 | 76 | 75 |
| $\hat{\rho}^2(s_t, x)$ | | 84 | 91 | 83 | 88 | 81 | 81 | 88 | 91 | 86 | 83 | 83 | 83 |
| $\hat{\rho}^2(s_I, \bar{x})$ | 30 | 92[a] | 92[a] | 91 | 92[a] | 92[a] | 92[a] | 90[a] | 90 | 89 | 90 | 90[a] | 90 |
| | 100 | 98[a] | 98[a] | 96 | 98[a] | 98[a] | 98[a] | 97[a] | 97 | 96 | 97 | 97[a] | 97 |
| | 150 | 98[a] | 98[a] | 97 | 98[a] | 98[a] | 98[a] | 98[a] | 98 | 97 | 98 | 98[a] | 98 |
| $\hat{\rho}^2(s_I, \bar{s})$ | 30 | 89 | 89 | 89 | 89 | 87 | 84 | 88 | 86 | 89 | 86 | 77 | 85 |
| | 100 | 97 | 97 | 97 | 96 | 96 | 95 | 96 | 95 | 96 | 95 | 93 | 95 |
| | 150 | 98 | 98 | 98 | 98 | 97 | 96 | 97 | 97 | 98 | 97 | 94 | 97 |
| $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ | 30 | 92[a] | 92[a] | 91 | 92[a] | 92[a] | 92[a] | 90[a] | 90 | 90 | 90 | 90[a] | 90 |
| | 100 | 98[a] | 98[a] | 97 | 98[a] | 98[a] | 98[a] | 97[a] | 97 | 97 | 97 | 97[a] | 97 |
| | 150 | 98[a] | 98[a] | 98 | 98[a] | 98[a] | 98[a] | 98[a] | 98 | 98 | 98 | 98[a] | 99 |

[a]For the corresponding subscore $M_{ssI} > M_{sse}$, $M_{xxI} > M_{xxe}$ and $(M_{sxI} - M_{sxe})^2 \geq (M_{ssI} - M_{sse})(M_{xxI} - M_{xxe})$ so that $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ and $\hat{\rho}^2(s_I, \bar{x})$ are set equal to $\hat{\rho}^2(x_I, \bar{x})$.

stitutional subscore means. The observed institutional subscore means come close to be favored only for two subscores and at least 100 examinees, as can be observed from Table 3. Again, this result is not a big surprise as Longford (1990) also found subscores to have of little added value for one of the tests considered.

- Use of both observed subscore mean and observed total score mean generally provides only relatively small gains over use of observed subscores and hardly any gain over use of observed total scores.

- Results vary appreciably from form to form.

- Although reporting institutional subscore means has little justification in the preponderance of cases, reporting such means does not necessarily lead to poor estimates, for the reliability at the institutional level is generally high.

The results for augmented subscores $\hat{s}_{Iu}$'s are not provided, primarily because $\hat{s}_{Iu}$'s result

**Table 2.**
***Percent Reduction (***$100 \times$ ***Proportional Reduction) in Mean-Squared Error With Three Subscores***

| | $n$ | Test form 1 | | | Test form 2 | | |
|---|---|---|---|---|---|---|---|
| | | Subscore | | | Subscore | | |
| | | 1 | 2 | 3 | 1 | 2 | 3 |
| $\hat{\rho}^2(s_t, s)$ | | 85.3 | 85.5 | 84.1 | 86.5 | 83.7 | 85.2 |
| $\hat{\rho}^2(s_t, x)$ | | 87.3 | 85.9 | 86.9 | 89.5 | 85.4 | 86.8 |
| $\hat{\rho}^2(s_I, \bar{x})$ | 30 | 92.3[a] | 92.3[a] | 92.3[a] | 90.1[a] | 89.1 | 90.1[a] |
| | 100 | 97.6[a] | 97.6[a] | 97.6[a] | 96.8[a] | 95.8 | 96.8[a] |
| | 150 | 98.4[a] | 98.4[a] | 98.4[a] | 97.8[a] | 96.8 | 97.8[a] |
| $\hat{\rho}^2(s_I, \bar{s})$ | 30 | 91.0 | 90.7 | 88.6 | 88.9 | 89.8 | 83.7 |
| | 100 | 97.1 | 97.0 | 96.3 | 96.4 | 96.7 | 94.5 |
| | 150 | 98.1 | 98.0 | 97.5 | 97.6 | 97.8 | 96.2 |
| $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ | 30 | 92.3[a] | 92.3[a] | 92.3[a] | 90.1[a] | 90.3 | 90.1[a] |
| | 100 | 97.6[a] | 97.6[a] | 97.6[a] | 96.8[a] | 96.8 | 96.8[a] |
| | 150 | 98.4[a] | 98.4[a] | 98.4[a] | 97.8[a] | 97.8 | 97.8[a] |

[a]For the corresponding subscore $M_{ssI} > M_{sse}$, $M_{xxI} > M_{xxe}$ and $(M_{sxI} - M_{sxe})^2 \geq (M_{ssI} - M_{sse})(M_{xxI} - M_{xxe})$ so that $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ and $\hat{\rho}^2(s_I, \bar{x})$ are set equal to $\hat{\rho}^2(x_I, \bar{x})$.

in hardly any added benefit for the data.

### *3.3 Multivariate Analysis of Variance*

A basic difficulty that is encountered with these data can be explored by canonical analysis for a one-way multivariate analysis of variance (MANOVA) on the six subscores (Bock, 1975, chapter 6). As in the discussions of augmentation, let the subscores be denoted by $u_k$ for $1 \leq k \leq r = 6$. Let $\mathbf{C}_I$ be the institutional covariance matrix.[1] with row $k$ and column $k'$ equal to $c(u_{Ik}, u_{Ik'})$, and let $\mathbf{C}_e$ be the error covariance matrix with row $k$ and column $k'$ equal to $c(u_k - u_{Ik}, u_{k'} - u_{Ik'})$. Consider the system of relative eigenvalues and normalized relative eigenvectors such that

$$\mathbf{C}_I \mathbf{v}_k = \lambda_k \mathbf{C}_e \mathbf{v}_k$$

**Table 3.**
***Percent Reduction (***$100 \times$ ***Proportional Reduction)***
***in Mean-Squared Error With Two Subscores***

| | $n$ | Test form 1 | | Test form 2 | |
|---|---|---|---|---|---|
| | | Subscore | | Subscore | |
| | | 1 | 2 | 1 | 2 |
| $\hat{\rho}^2(s_t, s)$ | | 91.2 | 85.5 | 92.0 | 83.7 |
| $\hat{\rho}^2(s_t, x)$ | | 91.4 | 85.9 | 92.1 | 85.4 |
| $\hat{\rho}^2(s_I, \bar{x})$ | 30 | 92.3[a] | 92.3[a] | 89.8 | 89.1 |
| | 100 | 97.6[a] | 97.6[a] | 96.5 | 95.8 |
| | 150 | 98.4[a] | 98.4[a] | 97.5 | 96.8 |
| $\hat{\rho}^2(s_I, \bar{s})$ | 30 | 91.5 | 90.7 | 88.4 | 89.8 |
| | 100 | 97.3 | 97.0 | 96.2 | 96.7 |
| | 150 | 98.2 | 98.0 | 97.5 | 97.8 |
| $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ | 30 | 92.3[a] | 92.3[a] | 89.9 | 90.3 |
| | 100 | 97.6[a] | 97.6[a] | 96.5 | 96.8 |
| | 150 | 98.4[a] | 98.4[a] | 97.6 | 97.8 |

[a] For the corresponding subscore, $M_{ssI} > M_{sse}$, $M_{xxI} > M_{xxe}$ and $(M_{sxI} - M_{sxe})^2 \geq (M_{ssI} - M_{sse})(M_{xxI} - M_{xxe})$ so that $\hat{\rho}^2(s_I, \hat{s}_{Ia})$ and $\hat{\rho}^2(s_I, \bar{x})$ are set equal to $\hat{\rho}^2(x_I, \bar{x})$.

for $1 \leq k \leq r$, $\lambda_k \geq \lambda_{k+1}$ for $k < r$, $\mathbf{v}_k' \mathbf{C}_e \mathbf{v}_k = 1$ for $1 \leq k \leq r$, and $\mathbf{v}_k' \mathbf{C}_e \mathbf{v}_{k'} = 0$ for $k \neq k'$. For $n$ examinees from an institution, the maximum possible value of $\rho^2(d_I, \bar{d})$ for a linear combination $d$ of the $u_k$ with institutional mean $d_I$ is $\lambda_1/(\lambda_1 + 1/n)$. This maximum is achieved if $d = \mathbf{v}_1' \mathbf{u}$ for $\mathbf{u}$ with coordinates $u_k$ for $1 \leq k \leq r$. Thus, in terms of institutional reliability, $d$ can be regarded as the optimal linear combination of subscores. For a linear combination $f$ of the $u_k$ with institutional mean $f_I$ such that $f_e = f - f_I$ and $d_e = d - d_I$ are uncorrelated, $\rho^2(f_I, \bar{f})$ cannot exceed $\lambda_2/(\lambda_2 + 1/n)$. The upper bound is achieved for $f = \mathbf{v}_2' \mathbf{u}$. Thus, in terms of institutional reliability, $f$ may be termed the second optimal linear combination of subscores because $f$ is the optimal linear combination of subscores subject to the constraint that $f_e$ and $d_e$ are uncorrelated.

To estimate $\lambda_k$ and $\mathbf{v}_k$ for the required values of $k$, the canonical analysis from a one-way MANOVA may be used. For examinee $i$ of institution $j$, let the subscore value for

$u_k$ be $u_{ijk}$, and let $\bar{u}_{jk}$ be the average subscore $u_k$ from institution $j$. Let $\bar{u}_{.k}$ be the mean subscore $u_k$ for all examinees, let $\mathbf{M}_e$ be the $r$ by $r$ within-institution matrix of mean cross products with row $k$ and element $k'$ equal to

$$M_{kk'e} = (N - J)^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (u_{ijk} - \bar{u}_{jk})(u_{ijk'} - \bar{u}_{jk'}),$$

and let $\mathbf{M}_I$ be the between-institution matrix of mean cross products with row $k$ and column $k'$ equal to

$$M_{kk'I} = (J - 1)^{-1} \sum_{j=1}^{J} n_j (\bar{u}_{jk} - \bar{u}_{.k})(\bar{u}_{jk'} - \bar{u}_{.k'}).$$

Let the $k$th largest relative eigenvalue of $\mathbf{M}_I$ relative to $\mathbf{M}_e$ be $\hat{\nu}_k$, and let the corresponding relative eigenvector be $\hat{\mathbf{v}}_k$. Let $\hat{\lambda}_k = K^{-1}(\hat{\mu}_k - 1)$. Then the estimate of the maximum possible $\rho^2(d_I, \bar{d})$ is $\hat{\lambda}_1/(\hat{\lambda}_1 + 1/n)$, and the corresponding estimate of the maximum possible value of $\rho^2(f_I, \bar{f})$ is $\hat{\lambda}_2/(\hat{\lambda}_2 + 1/n)$.

Results for the two test forms are summarized in Table 4.

**Table 4.**
***Percent Reduction (*** $100 \times$ ***Proportional Reduction) in Mean-Squared Error With Total Score, Optimal Linear Combination of Subscores, and Optimal Second Linear Combination***

| $n$ | Test form 1 | | | Test form 2 | | |
|---|---|---|---|---|---|---|
| | $\hat{\rho}^2(x_I, \bar{x})$ | $\hat{\rho}^2(d_I, \bar{d})$ | $\hat{\rho}^2(f_I, \bar{f})$ | $\hat{\rho}^2(x_I, \bar{x})$ | $\hat{\rho}^2(d_I, \bar{d})$ | $\hat{\rho}^2(f_I, \bar{f})$ |
| 30 | 0.923 | 0.925 | 0.427 | 0.901 | 0.910 | 0.505 |
| 100 | 0.976 | 0.976 | 0.713 | 0.968 | 0.971 | 0.773 |
| 150 | 0.984 | 0.984 | 0.788 | 0.978 | 0.981 | 0.834 |

Results for the optimal linear combination are virtually the same as results for the total score $x$. For the case of the linear combination $f$, reliability is not very satisfactory without an $n$ greater than 100, and the reliability is much lower than for $\bar{x}$. Thus a fundamental problem is that very little information appears available that is not provided by the total score.

## 4. Discussion and Conclusion

This paper demonstrates that reporting subscores can be quite different at an institutional level than at an individual level even though the basic arguments are quite similar. Few studies explore this issue for operational tests, with the exception of Longford (1990), who analyzed data from the pilot stage of development of a test. Our suggested analyses can be performed with output from standard statistical software and does not involve difficult computations, so that routine use of the proposed methodology is quite straightforward.

In the example under study, reporting examinee means on subscores does not appear to be justified for any realistic institution size, although reporting mean total scores for an institution does not appear to be problematic even for the smallest sample-size condition (30) examined. The results suggest that any possible use of subscores is most likely to succeed with more aggregated subscores and large institutions.

The methods used in this report can be directly applied to score reporting at a different type of aggregation, say states rather than institutions. It is also a straightforward matter to extend the approach to a hierarchy of aggregations, say institutions within states.

Another issue with reporting subscores for institutions is that equating and/or scaling for subscores is essential if information from more than a single form is to be used to characterize results for an institution. Although such information can be available in survey assessments such as NAEP, in typical cases that involve tests designed for assessment of individuals rather than groups, equating is available for the total score but not for subscores (for example, if an anchor test is used to equate the total test, only a few of the items will correspond to a particular subscore so that an anchor test equating of the subscore is not feasible). No proof exists that scaling is feasible in a particular application, and the possibility exists that scaling that may be rather adequate for an individual is far from satisfactory if applied to an institution, for the correlation structure for institutional means may be quite different than the correlation structure for individual results. It should also be emphasized that an application of scaling of subscores to the total score conceptually requires the subscore to measure the same construct as the total score, in which case there is no point reporting a subscore. Further, a subscore will typically involve too little data

for accurate and precise scaling.

In the case of large institutions, it is prudent to perform outlier analysis to detect unusual distributions of subscores or total scores. Such analysis is quite distinct from any outlier analysis performed at an individual level. This is a possible area for future research.

The combined estimate based on both the subscore mean and the total score mean is a reasonable candidate for some applications. However, the estimate did not help much, at least in the example under study. Further the estimate is not easy to explain to institutional users.

Analysis has been based on linear methods, so it is possible that other methods of analysis might yield different results.

## References


Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Bock, R. D., & Petersen, A. C. (1975). A multivariate correction for attenuation. *Biometrika, 62*, 673–678.

Cattell, R. B. (1956). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276.

Gini, C. (1912). *Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statische.* Bologna, Italy: Cuppini.

Haberman, S. (2005). *When can subscores have value?* (ETS RR-05-08). Princeton, NJ: ETS.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions, 24*(7), 349-368.

Harris, D. J., & Hanson, B. A. (1991, April). *Methods of examining the usefulness of subscores.* Paper presented in the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151-160.

Longford, N. T. (1990). Multivariate variance component analysis: An application in test development. *Journal of Educational Statistics, 15*(2), 91–112.

SAS Institute. (1996). *SAS/STAT software: Changes and enhancements through release 6.11.* Cary, NC: Author.

Snedecor, G., & Cochran, G. (1989). *Statistical methods.* Ames: Iowa State University Press.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17(2)*, 89-112.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds), Test scoring (pp. 343-387). Hillsdale, NJ: Lawrence Erlbaum Associates.

## Notes

[1] Note that the between-institution variance matrix obtained from the one-way MANOVA, which is an estimate of $\mathbf{C}_I$, has one large eigenvalue and a few negative eigenvalues for the six-subscore case and three-subscore case for both the test forms, which is some proof that most of the between-institution variance lies in the total score and not in the subscores.