ETS®

Research Report

When Can Subscores Have Value?

Shelby J. Haberman

Research & Development

May 2005
RR-05-08

When Can Subscores Have Value?

Shelby J. Haberman
ETS, Princeton, NJ

May 2005

Abstract

In educational tests, subscores are often generated from a portion of the items in a larger test. Guidelines based on mean-squared error are proposed to indicate whether subscores are worth reporting. Alternatives considered are direct reports of subscores, estimates of subscores based on total score, combined estimates based on subscores and total scores, and residual analysis of subscore. Applications are made to data from two testing programs.

Key words: Reliability, Kelley's formula, mean-squared error, regression, true score

Acknowledgements

The author thanks Neil Dorans, John Mazzeo, Sandip Sinharay, Alina von Davier, Matthias von Davier, Wendy Yen, and Jinming Zhang for helpful discussions and comments. Results concerning the SAT® I examination are derived from data provided by the College Board® and are used with permission.

A basic criterion for reporting of a subscore based on a portion of the items in a larger test should be whether the subscore provides a more accurate measure of the construct it measures than is provided by the total score from the larger test. This standard for subscore reporting is readily handled by using classical test theory. Arguments are based on least squares and mean-squared error. Section 1 provides the basic theory required for the analysis. Section 2 considers some examples from testing programs at ETS. Section 3 provides some conclusions from results of analysis.

1 Mean-Squared Errors for True Subscores

A true subscore can be estimated by use of the observed subscore, by use of the observed total score, or by a combination of the observed subscore and the observed total score. It is also possible to estimate the residual true subscore from regression of the true subscore on the true total score.

Classical Test Theory Background

To study these estimations, it is helpful to introduce some elementary classical test theory. Let A be an observed subscore, let A_T be the corresponding true score, and let A_E be the error $A - A_T$ of measurement. Let E represent a mean, σ^2 represent a variance, and σ represent a standard deviation, so that $E(A)$ is the mean of A, $\sigma^2(A)$ is the variance of A, and $\sigma(A)$ is the standard deviation of A. Under classical test theory (Lord & Novick, 1968; Holland & Hoskens, 2003), A_T and A_E are uncorrelated and $E(A_E) = 0$, so that $E(A_T) = E(A)$ and

$$\sigma^2(A) = \sigma^2(A_T) + \sigma^2(A_E).$$

Similarly, let B be an observed total score, let B_T be the corresponding true score, and let $B_E = B - B_T$ be the error of measurement, so that $E(B_E) = 0$, $E(B) = E(B_T)$, B_T and B_E are uncorrelated, $E(B_E) = 0$, and

$$\sigma^2(B) = \sigma^2(B_T) + \sigma^2(B_E).$$

It is also the case that the measurement error B_E is uncorrelated with the true score A_T and the measurement error A_E is uncorrelated with the true score B_T. Let $\sigma^2(A)$ and $\sigma^2(B)$

be assumed to be positive. Let Cov denote a covariance, and let ρ denote a correlation, so that $\mathrm{Cov}(A, A_T)$ is the covariance $\sigma^2(A_T)$ of A and A_T, and

$$\rho(A, A_T) = \frac{\mathrm{Cov}(A, A_T)}{\sigma(A)\sigma(A_T)} = \frac{\sigma(A_T)}{\sigma(A)} \tag{1}$$

is the correlation of the observed score A and the true score A_T. Let ρ^2 denote a squared correlation, so that

$$\rho^2(A, A_T) = \frac{\sigma^2(A_T)}{\sigma^2(A)}$$

is the reliability coefficient of A. Similarly,

$$\rho^2(B, B_T) = \frac{\sigma^2(B_T)}{\sigma^2(B)}$$

is the reliability coefficient of B. Analysis of the subscore A and the total score B is affected by the covariances

$$\mathrm{Cov}(A, B) = \sigma(A)\sigma(B)\rho(A, B)$$

and

$$\mathrm{Cov}(A_T, B_T) = \sigma(A_T)\sigma(B_T)\rho(A_T, B_T) = \sigma(A)\sigma(B)\rho(A, A_T)\rho(B, B_T)\rho(A_T, B_T). \tag{2}$$

The covariance and correlation of the true scores A_T and B_T may be determined by use of measurement properties of the remainder test score $C = B - A$. Because the true scores A_T and B_T are not correlated with the measurement errors A_E and B_E, the true score of C is $C_T = B_T - A_T$, the error of measurement $C_E = B_E - A_E$ of C is uncorrelated with the true score C_T, $E(C_T) = E(C)$, $E(C_E) = 0$,

$$\sigma^2(C) = \sigma^2(C_T) + \sigma(C_E),$$

and

$$\rho^2(C, C_T) = \frac{\sigma^2(C_T)}{\sigma^2(C)}.$$

One has

$$\mathrm{Cov}(A, B) = \frac{\sigma^2(A) + \sigma^2(B) - \sigma^2(C)}{2}$$

and

$$\mathrm{Cov}(A_T, B_T) = \frac{\sigma^2(A_T) + \sigma^2(B_T) - \sigma^2(C_T)}{2}.$$

In practice, estimates of the means $E(A)$ and $E(B)$ and the standard deviations $\sigma(A)$, $\sigma(A_E)$, $\sigma(B)$, and $\sigma(B_E)$, and reliability coefficients $\rho^2(A, A_T)$ and $\rho^2(B, B_T)$ are readily obtained from reports on testing programs produced at ETS. Given these estimates, $\sigma^2(A)$, $\sigma^2(B)$, $\sigma^2(A_T)$, $\sigma(A_T)$, $\sigma^2(B_T)$, and $\sigma(B_T)$ are readily estimated. For example,

$$\sigma(A_T) = \sigma(A)\rho(A, A_T).$$

Estimation of $\mathrm{Cov}(A, B)$ and $\mathrm{Cov}(A_T, B_T)$ is slightly more complicated, for it is not necessarily the case that measurement properties of C are directly available from reported data. In typical cases, a test score B is divided into $k \geq 2$ components C_i, $1 \leq i \leq k$, A is C_h for some h from 1 to k, and $B = \sum_{i=1}^{k} C_i$. Corresponding to each C_i is a true score C_{Ti} and an error C_{Ei} such that

$$C_i = C_{Ti} + C_{Ei},$$

$E(C_{Ei}) = 0$, C_{Ei} and C_{Ej} are uncorrelated for $i \neq j$. Data from standard summaries include estimates for $\rho(C_i, C_j)$ and $\rho(C_{Ti}, C_{Tj})$ for $i \neq j$ and for $\sigma(C_{Ti})$, $\sigma(C_i)$, and $\rho^2(C_i, C_{Ti})$. One may exploit the relationships

$$\mathrm{Cov}(A, B) = \sum_{i=1}^{k} \sigma(C_h)\sigma(C_i)\rho(C_h, C_i)$$

and

$$\mathrm{Cov}(A_T, B_T) = \sum_{i=1}^{k} \sigma(C_{Th})\sigma(C_{Ti})\rho(C_{Th}, C_{Ti}).$$

Naturally, if $i = h$, then

$$\sigma(C_h)\sigma(C_i)\rho(C_h, C_i) = \sigma^2(C_i)$$

and

$$\sigma(C_{Th})\sigma(C_{Ti})\rho(C_{Th}, C_{Ti}) = \sigma^2(C_{Ti}).$$

In the analysis in this paper, the reliability estimates produced by testing programs are taken as given. For the cases under study, basic computations involve the KR-20 approach (Kuder & Richardson, 1937; Dressel, 1940) and the Kristof approach (Kristof, 1974).

Direct Approximation

Given the summary measures just described, it is a relatively straightforward matter to consider the approximation of the true subscore A_T. For a baseline to compare approximations, consider the trivial prediction of A_T by the constant $E(A)$. The mean-squared error is then

$$\sigma^2(A_T) = E([A_T - E(A)]^2), \tag{3}$$

so that the root mean-squared error is $\sigma(A_T)$. If A_T is approximated by the observed score A, then the mean-squared error is

$$\sigma^2(A_E) = E([A - A_T]^2). \tag{4}$$

The root mean-squared error is

$$\sigma(A_E) = \sigma(A)[1 - \rho^2(A, A_T)]^{1/2}. \tag{5}$$

Alternatively, Kelley's formula may be applied, so that A_T is approximated by

$$K = E(A) + \rho^2(A, A_T)[A - E(A)], \tag{6}$$

and the mean-squared error is

$$\sigma^2(K - A_T) = \rho^2(A, A_T)\sigma^2(A_E) = [1 - \rho^2(A, A_T)]\sigma^2(A_T) \tag{7}$$

(Kelley, 1947). The root mean-squared error is

$$\sigma(K - A_T) = \rho(A, A_T)\sigma(A_E) = \rho(A, A_T)[1 - \rho^2(A, A_T)]^{1/2}\sigma(A). \tag{8}$$

The proportional reduction of mean-squared error from use of K rather than the constant predictor $E(A)$ to predict A_T is

$$\frac{\sigma^2(A_T) - \sigma^2(K - A_T)}{\sigma^2(A_T)} = \rho^2(A, A_T). \tag{9}$$

Regression Approximation

Regression analysis may be employed to approximate the true subscore A_T by the observed total score B (Wainer et al., 2001; Holland & Hoskens, 2003). The covariance

$$\mathrm{Cov}(A_T, B) = \mathrm{Cov}(A_T, B_T),$$

so that (1) and (2) imply that the prediction is

$$L = E(A) + \frac{\mathrm{Cov}(A_T, B_T)}{\sigma^2(B)}[B - E(B)] = E(A) + \rho(B, B_T)\rho(A_T, B_T)\frac{\sigma(A_T)}{\sigma(B)}[B - E(B)], \quad (10)$$

the mean-squared error is

$$\sigma^2(L - A_T) = \sigma^2(A_T) - [\mathrm{Cov}(A_T, B_T)]^2/\sigma^2(B) = [1 - \rho^2(B, B_T)\rho^2(A_T, B_T)]\sigma^2(A_T), \quad (11)$$

and the root mean-squared error is

$$\sigma(L - A_T) = [1 - \rho^2(B, B_T)\rho^2(A_T, B_T)]^{1/2}\sigma(A_T). \quad (12)$$

The proportional reduction in mean-squared error from use of L rather than $E(A)$ to predict A_T is

$$\rho^2(A_T, B) = \frac{\sigma^2(A_T) - [1 - \rho^2(B, B_T)\rho^2(A_T, B_T)]\sigma^2(A_T)}{\sigma^2(A_T)} = \rho^2(B, B_T)\rho^2(A_T, B_T). \quad (13)$$

If $\sigma(L - A_T)$ is less than $\sigma(K - A_T)$, then use of the subscore A by itself is very difficult to justify for estimation of the true score A_T, for the true score A_T in this instance is better approximated by use of the regression based on the observed total score B than by use of the estimate derived from Kelley's formula from the observed subscore A. The condition that $\sigma(L - A_T)$ is less than $\sigma(K - A_T)$ is equivalent to the condition that $\rho(B, B_T)\rho(A_T, B_T)$ exceeds $\rho(A, A_T)$. Thus use of the total score rather than the subscore is increasingly favored as the reliability of the total score increases, the correlation of true subscore and true total score increases, and the reliability of the subscore decreases.

One may also consider joint use of the observed subscore A and the total score B in approximation of A_T. Use of A and B together is equivalent to use of A and the remainder score $C = A - B$, although some changes in formulas are required. The best linear predictor of A_T based on A and B is

$$M = E(A) + \beta[A - E(A)] + \gamma[B - E(B)], \tag{14}$$

where β and γ satisfy the normal equations

$$\beta\sigma^2(A) + \gamma \operatorname{Cov}(A, B) = \operatorname{Cov}(A_T, A) = \sigma^2(A_T)$$

and

$$\beta \operatorname{Cov}(A, B) + \gamma\sigma^2(B) = \operatorname{Cov}(A_T, B) = \operatorname{Cov}(A_T, B_T).$$

With a bit of algebra, one finds that

$$\gamma = \frac{\sigma(A)}{\sigma B}\rho(A, A_T)\tau, \tag{15}$$

where

$$\tau = \frac{\rho(B, B_T)\rho(A_T, B_T) - \rho(A, B)\rho(A, A_T)}{1 - \rho^2(A, B)}, \tag{16}$$

and

$$\beta = \rho(A, A_T)[\rho(A, A_T) - \rho(A, B)\tau]. \tag{17}$$

The mean-squared error

$$\sigma^2(M - A_T) = \rho^2(A, A_T)[1 - \rho^2(A, A_T) - \tau^2]\sigma^2(A). \tag{18}$$

The proportional reduction in mean-squared error from use of M rather than $E(A)$ to predict A_T is then

$$\rho^2(A_T, M) = \rho^2(A, A_T) + \tau^2. \tag{19}$$

Obviously, $\sigma^2(M - A_T)$ is no greater than the minimum ν of $\sigma^2(L - A_T)$ and $\sigma^2(K - A_T)$. If $\sigma^2(M - A_T)$ is substantially smaller than ν, then M is worthy of consideration.

All analysis may be reported in terms of A and the remainder score C. For example,

$$M = E(A) + (\beta + \gamma)[A - E(A)] + \gamma[C - E(C)].$$

Approximation of the True Residual

It is also possible to examine the true residual

$$D_T = [A_T - E(A)] - \zeta[B_T - E(B)]. \tag{20}$$

By (2), the regression coefficient

$$\zeta = \frac{\mathrm{Cov}(A_T, B_T)}{\sigma^2(B_T)} = \frac{\rho(A_T, B_T)\sigma(A_T)}{\sigma(B_T)}.$$

This residual is the difference between the true subscore A_T and its best linear predictor based on the true total score B_T. Thus D_T provides a measure of the information provided by the true subscore that is not provided by the true total score. A positive value of D_T would indicate that expected performance on the subscore is better than expected from the total score, while a negative value of D_T suggests a weaker performance on the subscore than predicted by the total score.

The trivial approximation of D_T is the constant predictor 0 that corresponds to a true subscore that is a linear function of the true total score. The mean-squared error is then

$$\sigma^2(D_T) = [1 - \rho^2(A_T, B_T)]\sigma^2(A_T), \tag{21}$$

so that the root mean-squared error is

$$\sigma(D_T) = [1 - \rho^2(A_T, B_T)]^{1/2}\sigma(A_T). \tag{22}$$

Note that $\sigma(D_T) < \sigma(L - A_T)$.

An alternative approximation is

$$D = [A - E(A)] - \zeta[B - E(B)]. \tag{23}$$

In this case,

$$D_E = D - D_T = A_E - \zeta B_E, \tag{24}$$

so that $E(D) = E(D_T) = E(D_E) = 0$, D_T and D_E are uncorrelated, and the mean-squared error of D is

$$\sigma^2(D_E) = \sigma^2(A_E) - 2\zeta\,\mathrm{Cov}(A_E, B_E) + \zeta^2\sigma^2(B_E). \tag{25}$$

To evaluate the mean-squared error, note that

$$\mathrm{Cov}(A_E, B_E) = \mathrm{Cov}(A, B) - \mathrm{Cov}(A_T, B_T).$$

Kelley's formula can be applied here as well. If

$$F = \rho^2(D, D_T)D, \tag{26}$$

then

$$\sigma(F - D_T) = \rho(D, D_T)\sigma(D_E). \tag{27}$$

In (27),

$$\rho^2(D, D_T) = \frac{\sigma^2(D_T)}{\sigma^2(D)} = \frac{\sigma^2(D_T)}{\sigma^2(D_T) + \sigma^2(D_E)}.$$

Note that $\rho^2(D, D_T)$ is the reliability of D.

2 Examples

To illustrate application of subscores, consider data from the October 2002, administration of the SAT® I examination (Feigenbaum & Hammond, 2003). Results are summarized in Tables 1, 2, 3, and 4. In these tables, Verbal I, Verbal II, and Verbal III refer to the three separate portions of the SAT verbal examination, which are interleaved with Math I, Math II, and Math III, the three separate portions of the SAT math examination. An alternative breakdown of the SAT verbal uses critical reading (CR), analogies (A), and sentence completion (SC). Similarly, an alternate decomposition of SAT math uses four-choice math multiple choice (Math 4c), five-choice multiple choice (Math 5c), and student-produced math responses (Math S). To examine these tables, recall formulas (5), (8), (9), (12), (13), (14), (17), (15), (19), (22), (25), and (27). In Table 2, proportional reduction in mean-squared error is relative to use of a constant predictor equal to the expected subscore. In Table 4, the proportional reduction calculation is relative to use of the constant 0.

In these tables, for any given line, the subscore is A and the total score is B. For example, let A be the subscore of the first verbal section (10 sentence completions, 13 analogies, and a 13-item reading passage), and let B be the total score for the 78-item verbal test. Then $\sigma(A_E)$ is estimated to be 2.9, while $\sigma(K - A_T)$ is estimated to be 2.7. In this case, the subscore is clearly unsatisfactory relative to the approximation L based on the total score, for $\sigma(L - A_T)$ is estimated to be 2.0, a somewhat smaller figure than is available from A itself. Use of M yields only a slight reduction in root mean-squared error, for $\sigma(M - A_T)$ is also 2.0 if two significant figures are used. The weight β assigned to the subscore A is only 0.13. Both L and M are quite respectable estimates for A_T, for the proportional reductions in mean-squared error are both 0.91 to two significant figures. In the case of the residual estimates, $\sigma(D_T)$ is estimated to be 0.8, $\sigma(D_E)$ has estimate 2.2, and $\sigma(F - D_T)$ has estimate 0.7, so that there is little gain from use of F or D instead of the estimate 0 for D_T. Note that the proportional reduction in mean-squared error from use of F rather than 0 is only 0.11.

Similar results apply to the other sections of the verbal examination, and similar results also apply if A is a section of and B is the total score for the math examination. The variations in Table 1 in the coefficient γ mostly just reflects relative lengths of sections. In summary, none of the reported subscores of SAT I math or SAT I verbal provides any appreciable information concerning an examinee that is not already provided by the total score.

On the other hand, the analysis here would certainly support use of separate math and verbal scores. Let A be the math total, and let B be the sum of the math and verbal total. In this case, $\sigma(A_E) = 3.7$, $\sigma(K - A_T) = 3.6$, and $\sigma(L - A_T) = 5.4$, so that the math true score is much less well-predicted by the combined total score than by the the math score. Similarly, for the verbal score, $\sigma(A_E) = 4.6$, $\sigma(K - A_T) = 4.4$, and $\sigma(L - A_T) = 5.7$. There is little value in use of the joint predictor M. Here $\sigma(M - A_T)$ is 3.4 for math and 4.2 for verbal.

In the case of residual analysis, for math, $\sigma(D_T)$ is 4.7, $\sigma(D_E)$ is 2.9, and $\sigma(F - D_T)$ is 2.4. Because the total score is the sum of the math and verbal subscores, the same results apply for the verbal test. The estimated proportional reduction in mean-squared error from use of F rather than 0 is the estimated reliability coefficient 0.73 in both cases. Thus use of F does provide a substantial gain over the trivial estimate of 0. The root mean-squared

Table 1.

Root Mean-Squared Errors for True Score Estimation for SAT Subscores

	Items in	Total						
Subscore	subscore	score	$\sigma(A_E)$	$\sigma(K - A_T)$	$\sigma(L - A_T)$	$\sigma(M - A_T)$	β	γ
Verbal I	36	Verbal	2.9	2.7	2.0	2.0	0.13	0.35
Verbal II	30	Verbal	2.8	2.5	1.6	1.6	-0.02	0.35
Verbal III	12	Verbal	1.8	1.5	1.1	1.0	0.19	0.13
CR	40	Verbal	3.4	3.2	2.6	2.5	0.16	0.39
A	19	Verbal	2.1	1.8	1.3	1.2	0.17	0.17
SC	19	Verbal	2.1	1.8	1.3	1.2	0.20	0.18
Math I	25	Math	2.3	2.1	1.7	1.6	0.16	0.35
Math II	25	Math	2.3	2.1	1.5	1.5	0.07	0.33
Math III	10	Math	1.5	1.2	0.7	0.7	0.06	0.13
Math 4c	15	Math	1.9	1.6	0.9	0.9	0.03	0.21
Math 5c	35	Math	2.7	2.6	2.1	2.1	0.10	0.50
Math S	10	Math	1.2	1.1	0.7	0.7	0.16	0.12
Verbal	78	Total	4.6	4.4	5.7	4.2	0.70	0.13
Math	60	Total	3.7	3.6	5.4	3.4	0.76	0.09

Table 2.

Proportional Reduction of Mean-Squared Error Achieved by True Score Estimation for SAT Subscores

Subscore	Total score	K	L	M
Verbal I	Verbal	0.84	0.91	0.91
Verbal II	Verbal	0.80	0.92	0.92
Verbal III	Verbal	0.72	0.86	0.87
CR	Verbal	0.84	0.89	0.90
A	Verbal	0.74	0.87	0.88
SC	Verbal	0.78	0.88	0.89
Math I	Math	0.87	0.92	0.92
Math II	Math	0.83	0.91	0.91
Math III	Math	0.64	0.89	0.89
Math 4c	Math	0.72	0.91	0.91
Math 5c	Math	0.89	0.93	0.93
Math S	Math	0.73	0.89	0.90
Verbal	Total	0.91	0.85	0.92
Math	Total	0.92	0.82	0.93

Table 3.
Root Mean-Squared Error for Residual Estimation for SAT Subscores

Subscore	Total score	$\sigma(D_T)$	$\sigma(D_E)$	$\sigma(F - D_T)$
Verbal I	Verbal	0.8	2.2	0.7
Verbal II	Verbal	0.3	2.2	0.3
Verbal III	Verbal	0.8	1.7	0.7
CR	Verbal	1.3	2.3	1.2
A	Verbal	0.8	1.9	0.7
SC	Verbal	0.8	1.8	0.7
Math I	Math	0.5	1.8	0.5
Math II	Math	0.6	1.8	0.6
Math III	Math	0.4	1.4	0.4
Math 4c	Math	0.5	1.6	0.5
Math 5c	Math	0.5	1.8	0.5
Math S	Math	0.4	1.2	0.4
Verbal	Total	4.7	2.9	2.4
Math	Total	4.7	2.9	2.4

Table 4.
Proportional Reduction of Mean-Squared Error Achieved by Residual Estimation for SAT Subscores

Subscore	Total score	F
Verbal I	Verbal	0.11
Verbal II	Verbal	0.02
Verbal III	Verbal	0.18
CR	Verbal	0.24
A	Verbal	0.16
SC	Verbal	0.15
Math I	Math	0.09
Math II	Math	0.11
Math III	Math	0.08
Math 4c	Math	0.08
Math 5c	Math	0.07
Math S	Math	0.12
Verbal	Total	0.73
Math	Total	0.73

error from use of F is about half the corresponding root mean-squared error from use of 0. On the other hand, the proportional reduction of mean-squared error of 0.73 from use of F to assess deviation of SAT I math from the value expected by SAT I total is somewhat smaller than the proportional reduction in mean-squared error of 0.92 associated with use of K for estimation of SAT I math.

One may argue that it is unreasonable to expect very much information from subscores in the SAT I math and verbal examinations. The SAT I math and SAT I verbal examinations measure relatively limited content areas. On the other hand, some Praxis™ examinations contain parts that test very distinct content areas. For instance, consider the test titled Fundamental Subjects: Content Knowledge with code 0511 (Grant, 2003). This test measures English language arts (E), mathematics (M), citizenship and social science (C), and science (S). Each area is measured with 25 multiple-choice items, and the total raw score is the sum of the scores for each area. Results are summarized in Tables 5, 6, 7, and 8.

Here the direct estimate K of true subscore is roughly comparable to the estimate L of true subscore derived from the total score. Use of M provides a modest but appreciable improvement in all cases. Results are best for mathematics, and in all cases a relatively substantial weight is given to the direct estimate. With M, proportional reductions of mean-squared error are around 0.8, so that estimation of the true subscores by use of M can be regarded as relatively successful; however, the proportional reductions in error achieved from M are somewhat smaller than those achieved with M for subscores of SAT I math or verbal. The essential issue would appear to be that the subscores are less accurately predicted by total score in the Praxis case.

For residual analysis, appreciable gains over estimation of D_T by 0 are only seen for English language arts and mathematics, and even here the gains require use of F. The proportional reductions in mean-squared error reported in Table 8 are all relatively modest.

Table 5.
Root Mean-Squared Error for True Score Estimation for Praxis Subscores

Subscore	$\sigma(A_E)$	$\sigma(K - A_T)$	$\sigma(L - A_T)$	$\sigma(M - A_T)$	β	γ
E	1.7	1.5	1.5	1.3	0.44	0.10
M	1.9	1.7	1.9	1.5	0.51	0.12
C	1.8	1.5	1.3	1.2	0.60	0.03
S	2.0	1.7	1.3	1.3	0.57	0.05

Table 6.
Proportional Reduction of Mean-Squared Error Achieved by True Score Estimation for Praxis Subscores

Subscore	K	L	M
E	0.73	0.70	0.80
M	0.79	0.73	0.83
C	0.68	0.77	0.81
S	0.69	0.80	0.82

Table 7.
Root Mean-Squared Error for Residual Estimation for Praxis Subscores

Subscore	$\sigma(D_T)$	$\sigma(D_E)$	$\sigma(F - D_T)$
E	1.3	1.5	1.0
M	1.6	1.6	1.1
C	1.0	1.6	0.8
S	1.0	1.7	0.9

Table 8.
Proportional Reduction of Mean-Squared Error Achieved by Residual Estimation for Praxis Subscores

Subscore	F
E	0.43
M	0.48
C	0.29
S	0.25

3 Conclusions

The methods of subscore analysis proposed are very easily implemented and provide a rational criterion for assessing the value of subscores. Results suggest that a good deal of caution is needed. Subscores are most likely to have value if they have relatively high reliability by themselves and if the true subscore and true total score have only a moderate correlation. Both conditions are important. The SAT subscores are relatively unsuccessful due to the very high correlations of their true scores with the true total score; however, many of the subscores are rather reliable. Appropriate approximations of the true subscore give very high weight to the total score. The Praxis subscores are often less reliable than are many of the SAT subscores, but the correlation of true subscores to true total score is somewhat more modest than for the SAT subscores. Nonetheless, even for the Praxis subscores, which are all based on 25 items and measure very different content areas, the subscores are best used when combined with the total score, and the reliability of the resulting combination M is somewhat less than for the total score. Although the results here do not prove that subscores cannot be useful, they do suggest that claims for the value of subscores should be treated skeptically and should be verified by use of procedures similar to those in this report.

This report emphasizes simple approaches to subscores. It is possible that alternatives can be constructed that are quite attractive in particular applications. For example, subscore predictions from total scores may be based on use of log-linear models or use of item-response theory. Thus additional work can be considered to aid in subscore assessment.

References

Dressel, P. L. (1940). Some remarks on the Kuder-Richardson 20 (KR-20) reliability statistic for formula scored tests. *Psychometrika*, *5*, 305–310.

Feigenbaum, M., & Hammond, S. (2003). *Test analysis, College Board, SAT® I: Reasoning Test, Fall 2002 administrations, 3YSA03-3YSA05* (Report No. SR-2003-37). Princeton, NJ: ETS.

Grant, M. (2003). *Fundamental subjects: Content knowledge (0511), test analysis, form 3ypx1* (Report No. SR-2003-62). Princeton, NJ: ETS.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, *68*, 123–149.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, *39*, 491–499.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Swygert, K. A., & Thissen, D. (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). Mahwah, NJ: Erlbaum.