Reporting Diagnostic Scores: Temptations, Pitfalls, and Some Solutions

Sandip Sinharay
Gautam Puhan
Shelby J. Haberman

ETS, Princeton, New Jersey
April 2009

Paper presented at the annual meeting of the National Council on Measurement in Education (NCME) held between April 14 and 16, 2009, in San Diego, CA.

Any opinions expressed in this article are those of the authors and not necessarily of Educational Testing Service.

Unpublished Work Copyright © 2009 by Educational Testing Service. All Rights Reserved. These materials are an unpublished, proprietary work of ETS. Any limited distribution shall not constitute publication. This work may not be reproduced or distributed to third parties without ETS's prior written consent. Submit all requests through www.ets.org/legal/index.html.

Educational Testing Service, ETS, the ETS logo, and Listening. Learning. Leading. are registered trademarks of Educational Testing Service (ETS).
Abstract

Diagnostic scores are of increasing interest due to their potential remedial and instructional benefit. Naturally, the number of testing programs that report diagnostic scores is on the rise, as are the number of research works on such scores. This paper starts by showing examples of diagnostic subscores reported by operational testing programs. Then this paper provides a discussion of existing psychometric methods for reporting diagnostic scores, followed by a brief review of a method proposed by Haberman (2008) that examines if subscores (that are the simplest form of diagnostic scores and are reported by several testing programs) have added value over the total score. Using results from several operational and simulated data sets, it is demonstrated that it is not straightforward to have diagnostic scores with added value. Some recommendations are made for those interested to report diagnostic scores.

Key words: Augmented subscore, Item response theory, Mean squared error, Subscore.
Diagnostic scores are of increasing interest due to their potential remedial and instructional benefit. Failing candidates want to know their strengths and weaknesses in different content areas to plan for future remedial work. States and academic institutions such as colleges and universities often want a profile of performance for their graduates to better evaluate their training and focus on areas that need instructional improvement (Haladyna & Kramer, 2004). According to the National Research Council report “Knowing What Students Know” (2001), the target of assessment is to provide particular information about an examinee’s knowledge, skill, and abilities. Diagnostic scores can provide such information. Naturally, there is a substantial pressure on the testing programs to report diagnostic scores, both at the individual examinee level and at aggregate levels (such as at the level of institutions or states).

What Are Diagnostic Scores?

Figures 1 and 2 show the top and bottom parts of the score report of an imaginary examinee (Mary D. Poppins) on two Praxis Series™ assessments—the Mathematics: Content Knowledge assessment and the Principles Learning and Teaching 7-12 assessment. Figure 1 shows the scaled scores on the two assessments obtained by Mary and Figure 2 shows the raw scores earned by Mary in the different categories on the two assessments. The figure also shows the raw points available in the categories and the range of scores obtained by the middle 50% of a group of examinees of appropriate education level. Figure 2 represents a typical diagnostic score report for examinees—the scores on the categories are the diagnostic scores. The intent is that the examinee will work harder on the categories on which she performed poorly (for example, on “algebra
Figure 3 is a typical diagnostic score report at an aggregate level. It shows, for the Praxis Series™ PPST Writing assessment, the average percent correct score of an institution on four categories of the assessment, the corresponding averages for the state the institution belongs to, and the corresponding averages for the whole nation. Here, the intent is that if an institution finds that its examinees performed poorly on a category compared to the state or the nation, a remedial and instructional workshop can be given to the examinees to improve their performance on the category. The categories usually correspond to the content areas in the test (see Figure 2).

**Subscores, Augmented Subscores and Objective Performance Index (OPI)**

The diagnostic score report shown in Figure 2 is based on raw scores on different categories. These are also referred to as *subscores*. The score report in Figure 3 is based on the average of subscores. Subscores are the simplest (e.g., raw number correct) possible diagnostic scores and are used by several testing programs such as SAT®, ACT, and LSAT.

Wainer et al. (2001) suggested an approach to increase the precision of a subscore by borrowing information from the other subscores. Because subscores are almost always found to correlate moderately or highly with each other, it is reasonable to assume that, for example, the listening score of a student has some information about the reading score. In the approach of Wainer et al., an “augmented” reading score, which is a linear combination of the reading subscore and the listening subscore, is reported for each examinee.
The objective performance index (OPI; Yen, 1987) is another approach to enhance a subscore by borrowing information from other part of the test. This approach uses a combination of item response theory (IRT) and Bayesian methodology. OPI is a weighted average of two estimates of performance: (i) the observed subscore, and (ii) an estimate, obtained using a unidimensional item response theory (IRT) model, of the subscore based on the examinee’s overall test performance. If the observed and estimated subscores differ significantly, then the OPI is defined as the observed subscore expressed as a percent. It should be noted that this approach, because of the use of a unidimensional IRT model, may not provide accurate results when the data are truly multidimensional. Ironically, that is when subscores can be expected to have added value.

**Model-based Approaches for Diagnostic Score Reporting**

This section describes a few of the most extensively used models for diagnostic score reporting. These models are also called cognitive diagnosis models or diagnostic classification models (DCM; Rupp & Templin, 2009). These models assume that (i) solving each test item requires one or more skills, (ii) each examinee has a latent ability parameter corresponding to each of the skills, (iii) the probability that an examinee will answer an item correctly is a mathematical function of the skills the item requires and the latent ability parameters of the examinee. Descriptions of the models are given for dichotomous items, but the description often generalizes to polytomous items.

*Rule Space Method (RSM) and Attribute Hierarchy Method (AHM)*

These two are not probability models and do not fit the general description of the DCMs given above, but are included here because the rule space model (RSM; Tatsuoka, 1983) is one of the earliest attempts at diagnostic score reporting (von Davier, Dibello, &
Yamamoto, 2008). The RSM fits a unidimensional item response theory (IRT) model to the data and computes for each examinee a unidimensional ability estimate $\theta$. The method also estimates for each examinee a person fit parameter $\zeta$ that denotes how deviant the response pattern of the examinee is from the ideal (Guttman) pattern. The points $(\theta, \zeta)$ are plotted in a 2-dimensional space. The RSM approach also finds in the same plot the “centroid points” that correspond to ideal responses of examinees with different skill patterns. Then the method classifies each examinee based on his/her responses into one of the clusters defined by these centroid points. The cluster membership of an examinee determines the skills he/she possesses. The attribute hierarchy method (AHM; Leighton, Gierl, & Hunka, 2004) is an extension of the RSM and makes the assumption of dependencies among the skills/attributes. The use of AHM necessitates the specification of a hierarchy outlining the dependencies among the skills.

**The DINA and NIDA Models**

These two models were discussed by, e.g., Junker and Sijtsma (2001) and are special cases of the constrained latent class models. Let $Q_{jk}$ be the indicator of whether the $j$th item requires the $k$th skill. The collection of all the $Q_{jk}$ is called the $Q$ matrix. Suppose $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ denote the skill vector of an examinee, where $\alpha_k$ denotes whether the examinee has mastered skill $k$. Let $\xi_j = \prod_{k=1}^{K} Q_{jk}^{\alpha_k}$, which is 1 if the examinee has mastered all the skills required to solve item $j$.

Then the deterministic-input noisy “and” gate (DINA) model assumes that

$$P_j(\alpha) = (1 - s_j)^{\xi_j} g_j^{1 - \xi_j}.$$
The quantities $s_j$ and $g_j$ are the slip parameter and the guessing parameter of item $j$, respectively.

The noisy-inputs, deterministic “and” gate (NIDA) model assumes that

$$P_j(\alpha) = \prod_{k=1}^{K} [1 - s_k]^\alpha_k g_k^{1 - \alpha_k},$$

where $s_k$ and $g_k$ are the slip parameter and the guessing parameter of skill $k$, respectively.

*Multiple Classification Latent Class Model (MCLCM)*

Latent class analysis assumes a categorical latent variable that determines the probability that an examinee will answer an item correctly. The latent variable divides the individual into several latent classes. Let $\pi_l$ denote the relative size of the $l$-th latent class, $l=1,2, \ldots L$, and $X_j$ denote the response of an examinee to the $j$-th item, $j=1,2,\ldots,J$. The LCA models the responses to the items of an examinee as

$$P(X_1 = x_1, X_2 = x_2, \ldots X_J = x_J) = \sum_{l=1}^{L} \pi_l \prod_{j=1}^{J} P(X_j = x_j | l).$$

The probability of a correct response to an item depends only on the latent class membership of an examinee in an LCA. Note that both the relative class sizes and the probabilities (also called conditional probabilities) are unknown and should be estimated from the data.

Maris (1999) suggested how to elicit diagnostic information (which is provided by the latent class membership of an examinee) using LCA. The number of latent classes are determined by the number of possible combinations of skills. For example, if a test measures 3 skills, each with 2 levels (that is, an examinee either has it or does not), the
number of latent classes is $2^5 = 32$. In estimating the conditional probabilities, it is assumed that if an examinee has $s$ skills and the number of skills required to solve an item is $S$, then the probability of a correct answer, $P(s \mid S)$, does not depend on the latent class membership and the item. This assumption greatly simplifies the estimation.

**General Diagnostic Model (GDM)**

The general diagnostic model (GDM; von Davier, 2008) is a class of models for diagnostic scoring. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ denote the skill vector of an examinee, where $\alpha_k$ is polytomous and can take one of $L_k$ values, and let $X$ denote the score on an item, where $X$ is polytomous as well. The item specific logits are defined as

$$
\log \left[ \frac{P(X = x)}{P(X = 0)} \right] = \beta_x + \sum_{k=1}^{K} \gamma_{sk} h(q_{ik}, \alpha_k),
$$

where $\beta_x$ is a difficulty parameter for score $x$, $\gamma_{sk}$ is a slope parameter, and $h(q_{ik}, \alpha_k)$ is a function that defines how the $Q$-matrix entries and the skill levels interact. For IRT-type models, $h(q_{ik}, \alpha_k)$ is usually equal to $q_{ik} \alpha_k$. The model is estimated using the maximum likelihood estimation procedure.

**Reparameterized Unified Model (RUM)**

DiBello, Stout, and Roussos (1995) developed the unified model as a probabilistic model to express the stochastic relationship between item responses and status of underlying skills of an examinee. Hartz (2002) reparameterized the model. The model expresses the probability of a correct response on item $j$ as

$$
P_j(\alpha, \eta) = P_j(\eta) \pi_j \prod_{k=1}^{K} (r_{jk}^{*})^{Q_{jk}(1-\alpha_k)},
$$
where $\pi_j^*$ is a conditional difficulty parameter, conditional on having mastered all the

skills required by an item, $P_{j\ell}(\eta) = \frac{e^{1.7(\eta-c_j)}}{1+e^{1.7(\eta-c_j)}}$, the success probability from the Rasch

model with difficulty parameter $c_j$ and ability parameter $\eta$, and $r_{jk}^*$ is an indicator of the

strength of evidence provided by item $j$ about mastery of skill $k$.

**Multidimensional Item Response Theory**

De la Torre and Patz (2005) and Yao and Boughton (2007) examined reporting of
diagnostic scores using multivariate IRT models (MIRT; Reckase, 1997) and the Markov
chain Monte Carlo (MCMC) algorithm for estimating the models. Haberman and
Sinharay (2009) suggested diagnostic reporting using an MIRT model that is fitted with a
stabilized Newton-Raphson approach that is more efficient than the MCMC algorithm. In
a MIRT-based approach, the probability of a correct response on the $j$-th item is
expressed as

$$P_j(\theta) = \frac{\exp(a_{j1}\theta_1 + a_{j2}\theta_2 + \ldots + a_{jk}\theta_k - b_j)}{1 + \exp(a_{j1}\theta_1 + a_{j2}\theta_2 + \ldots + a_{jk}\theta_k - b_j)},$$

where $b_j$ is the difficulty parameter of the item, $a_{jk}$'s are the slope parameters of the item
and $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ is the $K$-dimensional ability parameter of the examinee.

Finally, some of the other DCMs that have been suggested in the literature are:
the Bayesian inference networks (e.g., Almond, Dibello, Moulder, and Zapata-Rivera,
2007), the higher-order latent-trait model (de La Torre, 2005), the DINO and NIDO
models (see, e.g., Rupp & Templin, 2008) and the multicomponent latent trait model
(e.g., Embretson, 1997). See, for example, Dibello, Roussos, and Stout (2007), Fu and Li

**How Good Are the Existing Diagnostic Scores? A Closer Look**

It is not uncommon to observe decent reliabilities of diagnostic scores on personality inventories designed to measure specific personality traits such as anxiety, hostility, trust, etc. For example, Goldberg (1999) reported that the reliabilities of the 30 subscale scores (each subscale consisting of 8 items) in the revised NEO Personality Inventory (NEO PI-R) ranged from 0.61 to 0.85 (mean = 0.75). Considering the relative small number of items in the subscales, these reliabilities seem reasonably high for diagnostic use. The diagnostic scores found in educational measurement literature and operational practice often consist of only a few items. It is not uncommon to find diagnostic scores on, for example, 10 skills based on only about 20-30 items. Almost all of such scores are outcomes of retrofitting (that refers to reporting of diagnostic scores from tests that were designed to measure only one overall skill). This is often done to comply with clients’ requests for more diagnostic information on examinees without an increase in test length. Because these tests have been constructed specifically to measure a single construct, little reason exists to expect useful results from a DCM. In addition, it is often ignored that a diagnostic score in educational measurement refers to a domain area that is usually much broader than those covered by diagnostic scores in personality inventories.

Further, there are few checks done to make sure that the diagnostic scores have decent psychometric properties before reporting them. Even though Standards 2.1 and 6.5 of the *Standards for Educational and Psychological Testing* (1999) demand proof of
adequate reliability of any reported scores, reliability of diagnostic scores is often not reported, so that examinees and users of test results are not informed of the degree to which confidence can be placed in skill classifications. In addition, the reliability figures in some applications of DCMs are based on simulated data (see, for example, Roussos et al., 2007b, pp. 304-305) rather than on empirical data. It may be unwise to report diagnostic information for a test unless there is clear evidence that reliable skill classifications can be obtained from the test data.

There is a lack of studies demonstrating the validity of diagnostic scores. For example, there is little evidence showing that diagnostic information is related to other external criteria. It is difficult to have much confidence in any diagnostic information whose validity has not been established. Haberman (2008) demonstrated via theoretical derivations that the validity of subscores is limited when the subscores are either not reliable or are highly correlated with total scores.

As Sinharay and Haberman (2008b) explained, the data may not provide information as fine-grained as suggested by the cognitive theory or as hoped by the testing practitioner. A theory of response processes based on cognitive psychology may suggest several skills. But a test includes a limited number of items and may not have enough items to provide enough information about all of these skills. Thissen-Roe, Hunt, and Minstrell (2004) have shown in the area of physics education that out of a large number of hypothesized misconceptions in student learners, only a very few misconceptions could be found empirically. Another example is the iSkills™ test (e.g., Katz, Attali, Rijmen, & Williamson, 2008), for which an expert committee identified seven performance areas that they thought comprised Information and Communications
Technology (ICT) literacy skill, but a factor analysis of the data revealed only one factor and the confirmatory factor models in which the factors corresponded to performance areas or a combination thereof, did not fit the data at all (Katz, Attali, Rijmen, & Williamson, 2008). As a result, only an overall ICT literacy score is reported for the test. The basic issue is that the data may not support the conjectures made by content experts about how examinees behave. Hence, an investigator attempting to report diagnostic information has to make an informed judgment on how much evidence the data can reliably provide and report only that much information.

To demonstrate the problems with short subscores, let us consider a licensure test that is designed for prospective teachers of children in primary through upper elementary school grades. The 120 multiple-choice questions focus on four major subject areas: language arts/reading, mathematics, social studies, and science. There are 30 questions per area and a subscore is reported for each of these areas—the subscore reliabilities are between 0.71 and 0.83. We ranked the questions on mathematics and science separately in the order of difficulty (proportion correct) and then formed a form A that consists of the questions ranked 1, 4, 7, …, 28 in mathematics and the questions ranked 1, 4, 7, …, 28 in science. Similarly, we formed a form B with questions ranked 2, 5, 8, …, 29 in mathematics and in science, and a form C with the remaining questions. Forms A, B, and C can be considered roughly parallel forms and, by construction, all of the several thousand examinees took all three of these forms. The reliabilities of these forms range between 0.46 and 0.60. Now let us consider all the 271 examinees who obtained a subscore of 7 on mathematics and 3 on science on Form A. Such examinees will most likely be given more science lessons. Is that justified? We examined the mathematics and
science subscores of the 271 examinees on Forms B and C. Table 1 gives a cross-tabulation of the subscores on form B of such examinees. It shows that the subscores on Form B often lead to different conclusions. The percent of the 271 examinees who obtained a mathematics subscore of 5 or lower is 34 and 39 for Forms B and C. The percent who obtained a science subscore of 6 or higher is 39 and 32 for Forms B and C. The percent of examinees whose mathematics score is higher than their science score is only 59 and 66 on Form B. This simple example demonstrates that remedial and instructional decisions based on short subscores will often be wrong. Note that if we had identified examinees whose mathematics and science scores were closer (e.g., 6 in mathematics and 4 in science), then the remedial and instructional decisions based on the short subscores would be even more inaccurate.

**A Method to Examine if Subscores Have Added Value Given the Total Score**

This section describes the approach of Haberman (2008) to determine whether and how to report subscores. Let us denote the subscore and the total score of an examinee as $s$ and $x$, respectively. Haberman (2008), taking a classical test theory (CTT) viewpoint, assumed that a reported subscore is intended to be an estimate of the true subscore $s$, and considered the following estimates of the true subscore:

- An estimate $s = \bar{s} + \alpha(s - \bar{s})$ based on the observed subscore, where $\bar{s}$ is the average subscore for the sample of examinees and $\alpha$ is the reliability of the subscore.
- An estimate $s = \bar{x} + c(x - \bar{x})$ based on the observed total score, where $\bar{x}$ is the average total score and $c$ is a constant that depends on the reliabilities and
standard deviations of the subscore and the total score and the correlations between the subscores.

- An estimate \( s_{sx} = \bar{s} + a(s - \bar{s}) + b(x - \bar{x}) \) that is a weighted average of the observed subscore and the observed total score, where \( a \) and \( b \) are constants that depend on the reliabilities and standard deviations of the subscore and the total score and the correlations between the subscores.

It is also possible to consider an augmented subscore \( s_{aug} \) that is an appropriately weighted average of all the subscores of an examinee (Wainer et al., 2001) as an estimate of the true subscore. However, \( s_{aug} \) yields results that are very similar to those for \( s_{sx} \) (e.g., Sinharay, 2009). Note that the estimate \( s_{sx} \) is a special case of the augmented subscore \( s_{aug} \); the former places the same weight on all the subscores other than the one of interest instead of weighing them differently. Unless otherwise stated, \( s_{sx} \) will be referred to as the augmented subscore in the rest of the paper.

To compare the performances of \( s_x, s_x, \) and \( s_{sx} \) as estimates of \( s_t \), Haberman (2008) suggested the use of the proportional reduction in mean squared error (PRMSE). The larger the PRMSE, the more accurate is the estimate. This paper will denote the PRMSE for \( s_x, s_x, \) and \( s_{sx} \) as \( PRMSE_s, PRMSE_s, \) and \( PRMSE_{sx} \) respectively. The quantity \( PRMSE_s \) can be shown to be exactly equal to the reliability of the subscore.

Haberman (2008) recommended the following strategy to decide whether a subscore or an augmented subscore has added value:

---

1 A larger PRMSE is equivalent to a smaller mean squared error in estimating the true subscore and hence is desirable.
• If $PRMSE_s$ is less than $PRMSE_x$, declare that the subscore ```does not provide added value over the total score`` because the observed total score will provide more accurate diagnostic information (in the form of a lower mean squared error in estimating the true subscore) than the observed subscore in that case. Sinharay et al. (2007) discussed why this strategy is reasonable and how it ensures that a subscore satisfies professional standards.

• The quantity $PRMSE_x$ will always be at least as large as $PRMSE_s$ and $PRMSE_x$. However, $s_x$ requires a bit more computation than does either $s_s$ or $s_x$. Hence, declare that an augmented subscore has added value only if $PRMSE_x$ is substantially larger compared to both $PRMSE_s$ and $PRMSE_x$.

The computations for application of the method of Haberman (2008) are simple and involve only the sample variances, correlations, and reliabilities of the total score and the subscores. Haberman (2008) and Sinharay et al. (2007) explained that a subscore is more likely to have added value when it has high reliability and it is distinct from the other subscores. Haberman, Sinharay, and Puhan (2009) extended the method of Haberman (2008) to aggregate-level subscores.

\textbf{Results from a Survey of Operational and Simulated Data}

This section provides results from the application of the method of Haberman (2008) to several operational and simulated data sets.

\textit{Results from Operational Data}

Sinharay (2009) summarized several research papers (Haberman, 2008; Harris & Hanson, 1991; Puhan, Sinharay, Haberman, & Larkin, in press; Sinharay & Haberman, 2008a)
that examined whether subscores have added value for operational tests. The findings are summarized in Table 2.

Each row in the table shows, for a test, the number of subscores, average length of the subscores, average reliability of the subscores, average correlation among the subscores, average disattenuated correlation, the number of subscores that have added value, and the number of augmented subscores that have added value (where the assumption was made that an augmented subscore has added value if the corresponding $PRMSE_{xx}$ is larger than the maximum of $PRMSE_x$ and $PRMSE_x$ by 0.01 or larger).

For SAT Verbal (the first row of numbers in Table 2), the subscores refer to the critical reading, analogies and sentence completion scores, percentile scores for which used to be reported to the examinees. For SAT Math, the subscores refer to the scores on four-choice multiple choice questions, five-choice multiple choice questions, and student-produced responses—these were not operationally reported. For SAT (the third row of numbers in Table 2), the subscores actually refer to the SAT Verbal and SAT Math scores. For test TA, the seven subscores, each corresponding to a skill area the test is supposed to measure, were originally intended to be reported, but actually are not reported now. For all the other tests considered in Table 2, the subscores refer to operationally reported subscores.

---

2 where the disattenuated correlation between two subscores is equal to the simple correlations between them divided by the square root of the product of the reliabilities of the two subscores.

3 Note that although the table reports the averages to summarize a lot of information in a compendious manner, for some of these tests, the lengths, reliabilities, and correlations of the subscores are substantially unequal.

4 Changing 0.01 to other small values such as 0.02 or 0.03 did not affect the conclusions much.
Some tests such as CF had constructed response (CR) items. For a subscore with CR items, the “length” refers to the total number of score categories minus the number of items (for example, for a subscore with 4 items, each with score categories 0, 1, and 2, the length is $4 \times 3 - 4 = 8$).

For the P-ACT+ English and Mathematics tests, the numbers shown in Table 2 are from Harris and Hanson (1991), who used three forms each of these tests. The subscore reliabilities were not provided in Harris and Hanson (1991). However, for each form, the correlation and disattenuated correlation between the subscores were provided—these were used to compute the product of the reliabilities of the two subscores and then the Spearman-Brown prophecy formula was used to estimate the reliabilities (as the length of the subscores is known). For these data, the methods of Haberman (2008) were not applied because of the lack of information. However, Harris and Hanson (1991) concluded that the P-ACT+ subscores do not provide information distinct from the total scores using an approach that involves fitting of beta-binomial models to the observed subscore distributions. Hence, it was assumed that none of the P-ACT+ subscores have added value for any of these forms.

Figures 4 to 6 show, for the operational data sets, the percentage of subscores (Figures 4 and 5) or augmented subscores (Figure 6) that had added value. In each of these figures, the Y-axis corresponds to the average disattenuated correlation among the subscores. In Figure 4, the X-axis denotes the average length of the subscores, while, in Figures 5 and 6, the X-axis denotes average subscore reliability.

The three figures plot, for each row listed in Table 2, a number that is the same as the percentage of subscores (or augmented subscores) that have added value at the point
(x, y), where x is the corresponding average subscore reliability multiplied by 100 (or length in Figure 4) and y is 100 times the average disattenuated correlation. For example, in Table 2, the third row shows that the SAT has average length of 69, average disattenuated correlation of 0.76, and two subscores (that is 100% of all subscores) that had added value. Hence Figure 4 has the number 100 plotted at the point (69,76) at the bottom right corner.

Table 2 and Figures 4 to 6 show that in general, long subscores (which have high reliability) tend to have added value. For example, for the test TC2, subscores consisting of about 67 items had added value. However, not all long subscores have added value. For example, for the test TC1, which has an average subscore length of 68, only one of three subscores has added value. In general, tests with low average disattenuated correlation tended to have subscores with added value. However, there were some exceptions. For example, the average disattenuated correlation is 0.85 for Test CF, and none of the subscores have added value. However, for the test TB1, the average disattenuated correlation is 0.90, but one of the two subscores has added value.

Often, percentage of subscores with a specific average length (or average reliability) that have added value depends on the average disattenuated correlations. Hence, each of the figures shows a bold dotted line roughly dividing the plot into two regions in which the percentage is low (zero) and high (positive). Note that this line is somewhat arbitrary and was drawn after a visual examination of the points in the plot and not using any mathematical formula. In each of these figures, as one goes from the top left corner to the bottom right corner (that is, as the average length/reliability increases

\[ 5 \text{In other words, there is an interaction between the two factors average length (or average reliability) and average disattenuated correlation.} \]
and the average disattenuated correlation decreases), the subscores show more tendency
to have added value. Figure 6 shows that the augmented subscores have added value for
many of the operational data sets and that augmented subscores are much more likely to
have added value compared to the subscores themselves.

A Simulation Study

Table 2 and the Figures 4 to 6 were based on only a few data sets so that they are not
expected to be very stable (for example, if one obtains another collection of data sets, a
figure like Figure 4 for those data sets may look substantially different). Hence Sinharay
(2009) performed a simulation study where it was easier to control different factors and
study the effect of the factors of interest.

Data were simulated from the 2-parameter logistic MIRT model (Reckase, 1997;
Haberman, von Davier, & Lee, 2008) for which the item response function for item \(i\) is
given by

\[
(1 + e^{-a_i \theta_1 + a_i \theta_2 + \cdots + a_i \theta_K - b_i})^{-1}, \theta = (\theta_1, \theta_2, \ldots, \theta_K)' \sim N_K ((0,0,\ldots,0)', \Sigma),
\]

(1)

where \(\theta = (\theta_1, \theta_2, \ldots, \theta_K)\) is the \(K\)-dimensional ability parameter for an examinee, \(b_i\) is
the difficulty parameter for item \(i\), \(a_{i1}, a_{i2}, \ldots, a_{iK}\) are the \(K\) slope parameters for item \(i\)
\((a_{ki} \) denotes the loading of item \(i\) on the \(k\)-th dimension), and \(N_K\) denotes the density of
the normal distribution with \(K\) dimensions. Each component of \(\theta\) corresponds to a
 subscore. The diagonals of \(\Sigma\) are set to 1 to ensure identifiability of the model
parameters. For any item \(i\), only one among the slope parameters \(a_{i1}, a_{i2}, \ldots, a_{iK}\) is
assumed to be non-zero, depending on the subscore the item contributes to (e.g., for an
item belonging to the first subscore, \(a_{i1}\) is nonzero while \(a_{i2} = a_{i3} = \ldots = a_{ik}=0\), so that
the simulations are performed from a simple-structure MIRT model (that is equivalent to assuming that the subscores do not share common items).

Generating Item Parameters. One data set was obtained from each of three tests that operationally report subscores or section scores. The first test, which is a test in English, reports two section scores, each of which is based on 100 multiple choice items. The second test, which is the test TC2 in Table 1, reports three subscores, which consist of 66-67 multiple choice items (SinhaRay & Haberman, 2008). The third test, which is the test CA in Table 1, reports four subscores—language arts/reading, mathematics, social studies, and science—each based on 30 items (Puhan et al., 2008). The model given by Equation 1 was fitted using the stabilized Newton-Raphson algorithm (e.g., Haberman et al., 2008) to each of the three data sets to obtain estimated item parameters. For each data set, each operationally reported subscore or section score is considered to measure a skill area and is assumed to contribute to one dimension of \( \theta \). The estimated item parameter values were later used as generating item parameters in the simulation study. For each test, a bivariate normal distribution \( B_k \) was fitted to the log-slope and difficulty parameter estimates of the items belonging to \( k \)-th subscore, \( k = 1, 2, ..., K \). The generating item parameters for the \( k \)-th subscore in the test were randomly drawn from \( B_k \).

Factors Controlled in the Simulation Study. The following factors were controlled in the simulation studies:

- “Number of subscores”. For each of the three above-mentioned tests (that have two, three, and four subscores, respectively), the estimated item parameters were used to simulate data for which the number of subscores (or the dimension of \( \theta \))
is the same as that reported for the test. For example, the estimated item parameters from the data set from the test TC2 (that reports 3 subscores) was used to simulate data that have 3 subscores. Hence, in the simulations, the “number of subscores” can take one of three values: 2, 3, and 4. However, the “number of subscores” refers to more than simply the number of subscores. Each level of this factor also has its own set of item parameters obtained from an operational test data set as described above. For this reason, quotes are put around “number of subscores.”

- Length of the subscores. This paper used four values for the length—10, 20, 30, and 50. Note that the reliability of a test increases as the test length increases. For simplicity, this paper assumed that the different subscores for a given test have the same length.

- Level of correlation ($\rho$) among the components of $\theta$. This paper used six levels: 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95. If the correlation level for a simulation case is $\rho$, it was assumed, to simulate the data sets, that all the off-diagonal elements of $\Sigma$ (which denote the correlations between the components of $\theta$) in Equation 1 are equal to $\rho$. Note that the correlations among the components of $\theta$ are similar to the disattenuated correlations between the subscores. Hence, from Table 1, the choice of these levels of this correlation (especially, the lowest of them) is reasonable.

- Sample size $N$. This paper used three levels of the sample size: 100, 1,000, and 4,000.

Steps in the Simulation Study. For each simulation condition (determined by a
value each of the “number of subscores”, length of the subscores, level of correlation, and sample size), the generating item parameters were drawn once as described above (from the distributions $B_k s$), and then $R=100$ replications were performed. Each replication involved the following steps:

1) Generate the ability parameter $\theta$ for each of the $N$ examinees from the multivariate normal distribution $N_k((0,0,\ldots,0)', \Sigma)$, where the diagonals of $\Sigma$ are 1 and the off-diagonals are the same as the correlation level for the simulation case.

2) Simulate a data set, that is, simulate scores on each item of the test for each examinee, using Equation 1, the draws of $\theta$ in the above step and the above-mentioned generating item parameters for the test.

3) Calculate, for the simulated data set, several quantities, such as correlations among the subscores and the PRMSEs.

Results. Table 3 shows results for sample size of 1,000. The table shows results for four (out of six) values of the level of correlation. The results were very similar for other sample sizes and hence are not shown. Each of the 18 cells (where a cell corresponds to a simulation case) of the table shows, the following eight quantities:

1) $100 \times$ the average reliability of the total score (denoted as $\alpha_{tot}$ in the table), where the average is taken over the $R$ replications.

2) $100 \times$ the average reliability (remember that reliability=$PRMSE_s$) of the subscores (denoted as $PR_s$), where the average is taken over the appropriate

---

$^6$The standard error of relevant quantities were examined to make sure that the choice of $R=100$ produced results that were sufficiently precise.
number of subscores (for example, two subscores when the “number of subscores”=2) in each replication and then over the \( R \) replications.

3) \( 100 \times \) the average correlation between the subscores (denoted as \( r \)), where the average is taken over the appropriate number of correlations (for example, six correlations when the “number of subscores”=4) in each replication and then over the \( R \) replications.

4) \( 100 \times \) the average disattenuated correlation between the subscores. This is denoted as \( r_d \) in the tables.

5) \( 100 \times \) average \( PRMSE_x \) (denoted \( PR_x \)), where the average is taken over the appropriate number of subscores in each replication and then over the \( R \) replications.

6) \( 100 \times \) average \( PRMSE_{xx} \) (denoted as \( PR_{xx} \)).

7) Overall percent of subscores that have added value (denoted as \( \% \) sub). This is the overall percent of cases (out of a total of \( R \times K \), where \( K \) is the number of subscores) when \( PRMSE_x \) is larger than \( PRMSE_x \).

8) Overall percent of augmented subscores that have added value (denoted as \( \% \) aug). This is the overall percent of cases (out of a total of \( R \times K \)) when \( PRMSE_{xx} \) is larger than the maximum of \( PRMSE_x \) and \( PRMSE_{xx} \) by 0.01 or more.

The first eight lines of numbers of Table 3 show the results for two subscores, the next eight lines for three subscores, and the last eight lines for four subscores.

Figures 7 to 9 show, for simulated data with sample size of 1,000, the overall percentage of subscores (Figures 7 and 8) or augmented subscores (Figure 9) that have added value.
These plots (unlike Table 3) show results for all the six levels of correlation from 0.70 to 0.95.

Figures 7-9 are like Figure 4-6 and show, for each combination of subscore length (or average reliability) and level of correlation, a number showing the percentage. The figures show dashed lines roughly dividing the plot into two regions in which the percentage is low (less than 10%, roughly) and high (more than 10%). These three figures also reproduce the corresponding bold dotted lines from Figures 4 to 6 to assist a comparison of results from the operational and simulated data.

In Figures 7 to 9, there are up to three points (corresponding to the three values of “number of subscores”) for each \((X,Y)\)-coordinate. To avoid overlapping points in the figures, one of the three points was moved slightly up and another slightly down.

Table 3 and Figures 7 to 9 lead to the following conclusions:

- Overall, the percent of times when the subscores have added value increases with an increase in their lengths (or reliability) and with a decrease in the correlations among them (that is, as they become more distinct). This is expected from the discussions in Haberman (2008).

- If the average length of the subscores is 10, subscores are rarely of added value. Of sixteen such cases in Table 3, the percent of times when the subscores have added value is less than 1 in nine cases and has a significant non-zero value only when the level of correlation is only 0.70, which, according to Table 2, is rare in practice. This conclusion supports the findings of Table 2 in which none of the short subscores had any added value, but is stronger because the tests considered in Table 2 had very few subscores with length 10 or less. If the length of the
subscores is 10, the augmented subscores have added value

- always for level of correlation 0.7,
- often for level of correlation between 0.75 and 0.85,
- sometimes for level of correlation 0.9, and
- rarely for level of correlation 0.95.

- If the level of correlation is 0.9 or higher, subscores rarely have added value. Augmented subscores often have added value if the level of correlation is 0.9, but even they do not have any added value if the level is 0.95. This finding mostly agrees with the findings from Table 2, but seems to be more general (for example, because of a gap at the top right corner of Figure 4).

- If the average length of the subscores is 20 or larger, whether subscores have added value depends on the level of correlation. For example, for length 20, subscores have added value more than 50% of the time if the level of correlation is less than or equal to 0.75, while, for length 50, they have added value more than 50% of the time if the level of correlation is less than or equal to 0.85. Thus, there is an interaction between the length of the subscores and the level of correlation.

- The dotted and dashed lines in Figures 7-9 agree quite closely, which indicates that the conclusions are roughly similar from operational data and simulated data regarding when a subscore has added value. While the results from operational data have the advantage that they correspond to real data, the results from the simulated data have the advantage that they are based on several data sets and are more stable than those for the real data.

- The table and the figures show that it is not straightforward to have subscores that
have added value. The subscores have to be long (consisting of at least 20 items) and sufficiently distinct from each other (with disattenuated correlations less than 0.85) to have any hope of having added value. On the other hand, it is much easier to have augmented subscores that have added value. In Figure 9, most of the percentages are higher than 50.

- The “number of subscores” does not affect the percentage of cases when the subscores have added value, but the values of reliability etc. in Table 3 often change as this number changes.

- The PRMSE of the augmented subscores suggested by Wainer et al. (2001) is almost always very close to those suggested by Haberman (2008) (the difference between them was almost always less than 0.01); hence the results for the augmentation of Wainer et al. (2001) are not shown.

A question that naturally arises here is: Could the results be different if a DCM was applied? The answer is most likely a no. Haberman and Sinharay (2009) found that PRMSE of subscores based on MIRT are very close to PRMSE of subscores based on CTT. They also found in a study of several operational data sets (the same ones used in Puhan et al., in press) that MIRT-based subscores had added value if and only if the CTT-based subscores had added value. In addition, Henson, Templin and Douglas (2007) showed that the use of subscores resulted in only a modest reduction in correct classification rates in comparison with a DCM.
Recommendations and Conclusions

This paper demonstrated that even though there is an increasing interest in reporting diagnostic score, few diagnostic scores have adequate psychometric quality. Our recommendations on diagnostic score reporting are given below:

• As implied above, evidence of reliability and validity of the information should be reported whenever diagnostic scores are provided.

• If a DCM is employed to report diagnostic scores, the burden of proof lies on the person applying a DCM to demonstrate that the model parameters can be reliably estimated, that the model approximates the observed pattern of responses better than a simpler model (for example, a univariate item-response theory model), that the computations required to fit the model are not excessively time-consuming, and, that the diagnostic scores reported by the model have added value over a simple subscore or over the score(s) reported by a simple model. The simplest of the DCMs passing these ordeals may be operationally used for diagnostic scoring.

• Any reported diagnostic information should be based on a sufficient number of items. Sinharay (2009) provides some ideas regarding this issue. It is also important to ensure that the skills of interest are as distinct as possible from each other (though this is quite a difficult task before seeing the data).

• It is important to remember the advice of Luecht, Gierl, Tan & Huff (2006) that “inherently unidimensional item and test information cannot be decomposed to produce useful multidimensional score profiles—no matter how well intentioned or which psychometric model is used to extract the information” and that we should not “...try to extract something that is not there” (p. 6). Thus, for some
tests, it will not be justified to report diagnostic scores. Changing the structure of such tests, for example, using sound assessment engineering practices for item and test design (Luecht et al., 2006) may be the only option in order to be able to report diagnostic scores. If restructuring the test is not a reasonable option, then, instead of diagnostic score reporting, one can consider alternatives such as scale anchoring (e.g., Beaton & Allen, 1992), which makes claims about what students at different scale points know and can do, and item mapping (Zwick, Senturk, Wang & Loomis, 2001).

- Augmented subscores are often found to have added value. See, for example, the above-mentioned results from Sinharay (2009). This finding should come as good news to testing companies. Augmented subscores may be difficult to explain to the general public, who may not like the idea that, for example, a reported reading subscore is based not only on the observed reading subscore, but also on the observed writing subscores. However, this difficulty is more than compensated by the higher PRMSE (that is, more precision) of the augmented subscore. Note that if a test has only a few short subscores, an augmented subscore may have added value, but, should not be reported because its PRMSE, although substantially larger than $PRMSE_s$ and $PRMSE_{eq}$, will still not be adequately high.

- Diagnostic scores must be reported on some established scale. A temptation exists to make this scale comparable to the scale for the total score or to the fraction of the scale that corresponds to the relative importance of the diagnostic score, but these choices are not without difficulties given that diagnostic scores and total scores typically differ in reliability. In addition, if the diagnostic score is worth
reporting at all, then it presumably does not measure the same construct as the total score. Another important issue with reporting diagnostic scores is that linking of them would seem to be essential for most practical applications. In typical cases, equating is feasible for the total score but not for diagnostic scores (for example, if an anchor test is used to equate the total test, only a few of the items will correspond to a particular diagnostic score so that an anchor test equating of the diagnostic score is not feasible).
References


S Sinharay, S. (2009). When can subscores be expected to have added value? Results from operational and simulated data. (ETS Research Memorandum No. RR-09-xx). Princeton, NJ: ETS.


Table 1. Cross-tabulations of examinees

<table>
<thead>
<tr>
<th>Math subscore</th>
<th>Science subscore</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>14</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>11</td>
<td>15</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td></td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>11</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12</td>
<td>13</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td></td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>27</td>
<td>56</td>
<td>66</td>
<td>57</td>
<td>31</td>
<td>15</td>
<td>3</td>
<td></td>
<td>271</td>
</tr>
</tbody>
</table>
Table 2. Results from Analysis of Operational Data Sets

<table>
<thead>
<tr>
<th>Name/Nature of the test</th>
<th>No. of subscores</th>
<th>Av. length</th>
<th>Av. α</th>
<th>Av. corr. (disatt.)</th>
<th>Average corr.</th>
<th>How many subscores have added value?</th>
<th>How many aug. subs have added value?</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT Verbal</td>
<td>3</td>
<td>26</td>
<td>0.79</td>
<td>0.74</td>
<td>0.95</td>
<td>None</td>
<td>One</td>
</tr>
<tr>
<td>SAT Math</td>
<td>3</td>
<td>20</td>
<td>0.78</td>
<td>0.75</td>
<td>0.97</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>SAT</td>
<td>2</td>
<td>69</td>
<td>0.92</td>
<td>0.70</td>
<td>0.76</td>
<td>Both</td>
<td>Two</td>
</tr>
<tr>
<td>Praxis</td>
<td>4</td>
<td>25</td>
<td>0.72</td>
<td>0.56</td>
<td>0.78</td>
<td>Two</td>
<td>Four</td>
</tr>
<tr>
<td>P-ACT+ Eng</td>
<td>2</td>
<td>25</td>
<td>0.80</td>
<td>0.76</td>
<td>0.96</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>P-ACT+ Math</td>
<td>2</td>
<td>20</td>
<td>0.80</td>
<td>0.71</td>
<td>0.95</td>
<td>None</td>
<td>NA</td>
</tr>
<tr>
<td>CA (for teachers in elementary schools)</td>
<td>4</td>
<td>30</td>
<td>0.74</td>
<td>0.59</td>
<td>0.79</td>
<td>One</td>
<td>Four</td>
</tr>
<tr>
<td>CB (for teachers of special ed. programs)</td>
<td>3</td>
<td>19</td>
<td>0.46</td>
<td>0.42</td>
<td>0.96</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>CC (for beginning teachers)</td>
<td>4</td>
<td>19</td>
<td>0.38</td>
<td>0.44</td>
<td>1.00</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>CD (for teachers of social studies)</td>
<td>6</td>
<td>22</td>
<td>0.63</td>
<td>0.54</td>
<td>0.87</td>
<td>None</td>
<td>Six</td>
</tr>
<tr>
<td>CE (for teachers of Spanish)</td>
<td>4</td>
<td>29</td>
<td>0.80</td>
<td>0.65</td>
<td>0.80</td>
<td>One</td>
<td>Two</td>
</tr>
<tr>
<td>CF (for principals and school leaders)</td>
<td>4</td>
<td>25</td>
<td>0.48</td>
<td>0.41</td>
<td>0.85</td>
<td>None</td>
<td>Four</td>
</tr>
<tr>
<td>CG (for teachers of mathematics)</td>
<td>3</td>
<td>16</td>
<td>0.62</td>
<td>0.59</td>
<td>0.95</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>CH (for paraprofessionals)</td>
<td>3</td>
<td>24</td>
<td>0.85</td>
<td>0.76</td>
<td>0.89</td>
<td>None</td>
<td>Three</td>
</tr>
<tr>
<td>TA (measures cognitive and technical skills)</td>
<td>7</td>
<td>11</td>
<td>0.42</td>
<td>0.51</td>
<td>1.00</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>TB1 (tests mastery of a language)</td>
<td>2</td>
<td>44</td>
<td>0.85</td>
<td>0.77</td>
<td>0.90</td>
<td>One</td>
<td>Two</td>
</tr>
<tr>
<td>TB2 (tests mastery of a language)</td>
<td>2</td>
<td>43</td>
<td>0.90</td>
<td>0.68</td>
<td>0.75</td>
<td>Two</td>
<td>Two</td>
</tr>
<tr>
<td>TC1 (measures achievement in a discipline)</td>
<td>3</td>
<td>68</td>
<td>0.85</td>
<td>0.76</td>
<td>0.90</td>
<td>One</td>
<td>Three</td>
</tr>
<tr>
<td>TC2 (measures achievement in a discipline)</td>
<td>3</td>
<td>67</td>
<td>0.87</td>
<td>0.72</td>
<td>0.82</td>
<td>Two</td>
<td>Three</td>
</tr>
<tr>
<td>TD1 (measures school and individual student progress)</td>
<td>4</td>
<td>15</td>
<td>0.70</td>
<td>0.73</td>
<td>0.98</td>
<td>None</td>
<td>No</td>
</tr>
<tr>
<td>TD2 (measures school and individual student progress)</td>
<td>6</td>
<td>13</td>
<td>0.70</td>
<td>0.75</td>
<td>1.00</td>
<td>None</td>
<td>No</td>
</tr>
</tbody>
</table>

Note. The reliability is denoted as \( \alpha \). Augmented subscores are denoted as “aug. subs”. The first four tests were discussed by Haberman (2008). The next two tests were discussed in Harris & Hanson (1991). The next eight, denoted CA-CH, are certification tests discussed in Puhan et al. (in press). The next seven, denoted TA, TB1, …TD2, were discussed in Sinharay & Haberman (2008a).
Table 3. Results from Analysis of Simulated Data Sets

| No. of sub- | Length of the sub- | 10 Correlation | 20 Correlation | 30 Correlation | 50 Correlation |
| sub- | scores | | | | |
| scores | | 70 80 90 95 | 70 80 90 95 | 70 80 90 95 | 70 80 90 95 |
| 2 | $\alpha_{tot}$ | 73 75 76 77 | 85 86 86 87 | 89 90 90 91 | 93 94 94 94 |
| | $P_{RS}$ | 62 62 63 63 | 77 77 77 77 | 83 83 83 83 | 89 89 89 89 |
| | $r$ | 44 51 58 61 | 54 62 69 74 | 57 66 75 79 | 62 71 80 85 |
| | $r_{d}$ | 71 82 92 97 | 70 80 90 96 | 69 79 90 95 | 69 79 90 95 |
| | $P_{Rx}$ | 63 68 73 76 | 72 77 82 85 | 75 80 86 88 | 79 84 89 92 |
| | $P_{Rx}$ | 68 70 74 76 | 80 81 84 85 | 85 86 87 89 | 90 91 92 93 |
| | % sub | 36 00 00 00 | 100 46 00 00 | 100 100 01 00 | 100 100 57 00 |
| | % aug | 100 94 08 01 | 100 100 65 01 | 100 100 98 03 | 95 96 100 16 |
| 3 | $\alpha_{tot}$ | 75 77 78 79 | 86 87 88 88 | 90 91 92 92 | 94 94 95 95 |
| | $P_{RS}$ | 56 56 56 56 | 72 72 72 72 | 80 80 80 80 | 87 87 87 87 |
| | $r$ | 39 45 51 54 | 50 58 65 69 | 56 64 72 76 | 61 69 78 82 |
| | $r_{d}$ | 70 80 91 96 | 70 80 90 95 | 70 80 90 95 | 70 80 90 95 |
| | $P_{Rx}$ | 61 67 74 77 | 69 75 82 85 | 72 79 86 89 | 75 82 88 92 |
| | $P_{Rx}$ | 66 70 74 77 | 78 80 84 86 | 83 85 87 90 | 89 89 91 93 |
| | % sub | 19 00 00 00 | 85 19 00 00 | 100 66 00 00 | 100 100 20 00 |
| | % aug | 100 92 38 10 | 100 100 74 13 | 100 100 91 16 | 97 100 100 39 |
| 4 | $\alpha_{tot}$ | 80 82 83 84 | 89 90 91 91 | 92 93 94 94 | 95 96 96 96 |
| | $P_{RS}$ | 57 57 57 57 | 72 72 72 72 | 80 80 80 80 | 87 86 87 87 |
| | $r$ | 40 46 52 55 | 50 58 65 69 | 55 63 72 76 | 60 69 78 82 |
| | $r_{d}$ | 70 81 91 96 | 69 80 90 95 | 69 80 90 95 | 69 79 90 95 |
| | $P_{Rx}$ | 62 70 78 82 | 69 76 84 88 | 72 79 87 90 | 73 81 89 93 |
| | $P_{Rx}$ | 69 73 79 82 | 79 81 85 88 | 83 85 89 91 | 89 90 92 94 |
| | % sub | 22 01 00 00 | 81 23 00 00 | 99 67 01 00 | 100 100 26 00 |
| | % aug | 100 92 41 08 | 100 100 74 13 | 100 100 91 24 | 83 100 100 58 |
Figure 1. Top part of an operational score report for an examinee.
**Figure 2. Bottom part of an operational score report for an examinee.**

<table>
<thead>
<tr>
<th>TEST CATEGORY*</th>
<th>Raw Points Earned</th>
<th>Raw Points Available</th>
<th>Average Performance Range **</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MATHEMATICS: CONTENT KNOWLEDGE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I. ALGEBRA AND NUMBER THEORY</td>
<td>1</td>
<td>8</td>
<td>2-5</td>
</tr>
<tr>
<td>II. MEASUREMENT, GEOMETRY, AND TRIGONOMETRY</td>
<td>9</td>
<td>12</td>
<td>5-8</td>
</tr>
<tr>
<td>III. FUNCTIONS AND CALCULUS</td>
<td>8</td>
<td>14</td>
<td>5-9</td>
</tr>
<tr>
<td>IV. DATA ANALYSIS, STATISTICS, AND PROBABILITY</td>
<td>7</td>
<td>8</td>
<td>3-6</td>
</tr>
<tr>
<td>V. MATRIX ALGEBRA AND DISCRETE MATHEMATICS</td>
<td>4</td>
<td>8</td>
<td>3-5</td>
</tr>
<tr>
<td><strong>PRINCIPLES OF LEARNING AND TEACHING: GRADES 7-12</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I. STUDENTS AS LEARNERS: DEVELOPMENT, DIVERSE LEARNERS, MOTIVATION, ENVIRONMENT</td>
<td>8</td>
<td>8</td>
<td>4-6</td>
</tr>
<tr>
<td>II. INSTRUCTION AND ASSESSMENT: INSTRUCTION/ASSESSMENT STRATEGIES, PLANNING</td>
<td>3</td>
<td>8</td>
<td>4-6</td>
</tr>
<tr>
<td>III. TEACHER PROFESSIONISM: REFLECTIVE PRACTITION, LARGER COMMUNITY</td>
<td>8</td>
<td>8</td>
<td>5-7</td>
</tr>
<tr>
<td>IV. STUDENTS AS LEARNERS: CASE HISTORIES/SHORT-ANSWER QUESTIONS</td>
<td>12</td>
<td>16</td>
<td>8-12</td>
</tr>
<tr>
<td>V. INSTRUCTION AND ASSESSMENT: CASE HISTORIES/SHORT-ANSWER QUESTIONS</td>
<td>7</td>
<td>12</td>
<td>8-12</td>
</tr>
<tr>
<td>VI. COMMUNICATION TECHNIQUES: CASE HISTORIES/SHORT-ANSWER QUESTIONS</td>
<td>6</td>
<td>8</td>
<td>4-8</td>
</tr>
<tr>
<td>VII. TEACHER PROFESSIONISM: CASE HISTORIES/SHORT-ANSWER QUESTIONS</td>
<td>6</td>
<td>8</td>
<td>4-6</td>
</tr>
</tbody>
</table>

ETS will retain your score for ten years for reporting purposes.

Message Codes:  A = Score automatically reported to state licensing agency.
                V = Score reported to recipient listed.
                N = Test not required by DI. Score not reported.
                E = Your score on this test qualifies for ETS recognition of excellence!

* Category-level information indicates the number of test questions answered correctly for relatively small subsets of the questions. Because they are based on small numbers of questions, category scores are less reliable than the official scaled scores, which are based on the full set of questions. Furthermore, the questions in a category may vary in difficulty from one test form to another. Therefore, the category scores of individuals who have taken different forms of the test are not necessarily comparable. For these reasons, category scores should not be considered a precise reflection of a candidate's level of knowledge in that category and ETS recommends that category information not be used to inform any decisions affecting candidates without careful consideration of such inherent lack of precision.

** The range of scores earned by the middle 50% of a group of examinees of appropriate education level (see Interpret your Scores section in this website) taking this test during the most recent three academic years. N/A means that this range was not computed because the test was taken by fewer than 30 examinees within the most recent three academic years. N/A indicates that this test section was not taken and, therefore, the information is not applicable.

Copyright © 2008 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service.

The PRAXIS SERIES is a trademark of ETS.
Figure 3. An operational score report for an institution with made up numbers
Figure 4. The percent of subscores that have added value for different subscore length and average disattenuated correlation for the operational data.
Figure 5. The percent of subscores that have added value for different average subscore reliability and average disattenuated correlation for the operational data.
Figure 6. The percent of augmented subscores that have added value for different average subscore reliability and average disattenuated correlation for the operational data.
Figure 7. The overall percent of subscores that have added value for different subscore length and level of correlation for simulated data with sample size 1000.
Figure 8. The overall percent of subscores that have added value for different average subscore reliability and level of correlation for simulated data with sample size 1000.
Figure 9. The overall percent of augmented subscores that have added value for different average subscore reliability and level of correlation for simulated data with sample size 1000.