Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 14, Number 4, March 2009

Reporting Subscores from College Admission Tests

Per-Erik Lyrén, *Umeå University, Sweden*

The added value of reporting subscores on a college admission test (SweSAT) was examined in this study. Using a CTT-derived objective method for determining the value of reporting subscores, it was concluded that there is added value in reporting section scores (Verbal/Quantitative) as well as subtest scores. These results differ from a study of the SAT I and a study of a basic skills test and thus highlight the need for practitioners and researchers to gather empirical evidence to support the reporting of subscores. The cause of the disparate results seems to be related to differences in the composition of the tests rather than differences in the composition of the examinee groups.

Outcomes of assessments are usually reported in the form of one or several scores. Many testing programs report only a total score, usually in the form of a composite. However, the last few decades have seen an increasing interest in the reporting of subscores, which are derived from subsections of tests. This is also true for college admission tests. For instance, the SAT (College Board, 2008) reports Critical Reading, Mathematics, and Writing scores, the ACT (ACT, 2007) reports English, Mathematics, Reading, and Science scores, and for the Psychometric Entrance Test (PET) a Verbal Reasoning score, a Quantitative Reasoning score, and an English score are reported (Allalouf, 2003). It is not uncommon that tests that were originally designed to produce reliable scores for ranking examinees are then expected to provide information that can be useful for remediation or other purposes as well. For instance, some Israeli institutions use the PET English score for placement (National Institute for Testing and Evaluation [NITE], 2009).

Monaghan (2006) points out that while assessment organizations and test developers want to be responsive to the desire of examinees and other parties with an interest in testing programs to report subscores, they also want to prevent these subscores from being misused. Further, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), states: "When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided" (Standard 1.12, p. 20). This means that before test developers or practitioners decide on which scores to report they should gather validity evidence in support of such a provision. That validity evidence may consist of logical evidence, procedural evidence, and empirical evidence (e.g. Haladyna & Kramer, 2004).

This article focuses on examining empirical evidence to support the reporting of subscores, specifically if the subscores meet certain statistical criteria that can be used to determine their potential added value.

APPROACHES TO SUBSCORE REPORTING

When the Graduate Record Examination (GRE) Program (see e.g. Chalifour & Powers, 1988) started developing subscores they used two subjectively defined criteria: (a) the subscores had to attain a reliability of at least .80, and (b) the disattenuated intercorrelations with other subscores had to be less than .90 (McPeek, Altman, Wallmark, & Wingersky, 1976). In an attempt to find logically meaningful subscores on the GRE Advanced Psychology Test, McPeek et al. used two methods: content analysis and factor analysis. The content analysis was based on the content areas defined

in the test specifications and included item and test analysis with a focus on item discrimination, intercorrelations, and reliability, using the two previously described subjective criteria. The factor analysis was performed to examine if the subscores were essentially unidimensional, and if other potential groupings of items (subscores) existed; however, because it was found that only one predominant factor existed on the entire test it was deemed unnecessary to subject each of the content area subscores to factor analysis. It was concluded that the subscores based on the content analysis had considerable potential to provide information for use in guidance and placement; yet, this only applied to candidates with unusually high or low scores. Further, the GRE Psychology Committee found the three subscores based on the factor analysis to be not useful for guidance and placement.

Longford (1990) used variance components methodology to examine the usefulness of reporting subscores for a college-level general education examination. Two versions of the test were examined: a short version (one form with 1 hour of testing time) and a long version (three forms with 3 hours of testing time). Longford concluded that subscores were worth reporting on the long version only, while on the short version, "any subscore, or a linear combination of subscores, is indistinguishable from a less reliable version of the total score" (p. 109).

Harris and Hanson (1991) examined English and Mathematics subscores and total scores on the P-ACT+ (American College Testing, 1989) to determine whether the subscores provided different and better information for examinee-level decisions, specifically placement, than the total score. The methods and measures used included a true score method presented by Lord (1965) and Hanson (1989), disattenuated correlations among subscores, effective weights (Wang & Stanley, 1970), and a procedure for estimating classification consistency indices described by Hanson and Brennan (1990). Harris and Hanson concluded that neither the English subscores nor the Mathematics subscores provided information distinct from the respective total scores. Also, in a simulated placement situation the examined subscore did not provide better information than the total score did, at least not with any practical significance.

Tate (2004) performed simulations to examine whether the additional provision of multiple subscores offers useful diagnostic information for individual students. Specifically, the relationship was examined between the average error variance for subscores and hit rates for detecting important subscore differences on the one hand, and test dimensionality and the correlational level among the subscores on the other hand. It was found that the average error variance of the total score increased with decreasing level of correlation and increasing test dimensionality, while the average error variance of the individual subscores depended only on the number of subtest items and not on the level of correlation. Also, the subscore differences depended only on the level of correlation. Tate concluded that the adequacy of the subscore performance, given a specific combination of dimensionality and correlational level, depends on (a) whether the subscores are to be used for absolute decisions or relative decisions, and (b) the choice of estimation method.

Haladyna and Kramer (2004) used several methods and measures to examine the validity of the interpretation and use of subscores from a test measuring knowledge of basic biomedical science, which is one part of an examination program for dentists (Joint Commission on National Dental Examinations, 2004). The methods and measures included (a) repeated measures one-way analysis of variance to determine if the subscore means suggested differences in the difficulty of the subdomains, (b) intercorrelations between subscores, which provides evidence regarding the dimensionality of the item responses; under conditions of unidimensionality the disattenuated correlations should be at least .90, (c) confirmatory factor analysis, also used to assess dimensionality, (d) an item analysis focused on discrimination indices using both the subscore and total score as criterion scores; discrimination indices based on the subscore should be different from those based on the total score, and (e) the reliability of differences among each pair of subscores, using a procedure recommended by Ryan (2003). Haladyna and Kramer concluded that "validity evidence can be garnered to support the interpretation and use of subscores that may be used both by failing candidates for planning future remedial studies and professional schools for the evaluations of their educational programs"(pp. 364–365). They note, however, that had the study focused on cognitive-process dimensions rather than content dimensions the results might have been different.

The idea of using the cognitive processes involved in item solving when creating subscores, rather than using the content areas that make up the test, was utilized in a study by Wainer, Sheehan, and Wang (2000). They used

the item difficulty modeling approach described in Sheehan and Mislevy (1990) and Embretson (1998) to construct subscales better suited for diagnostic feedback on the Education in the Elementary School Assessment (EES), a part of the Praxis program (Dwyer & Villegas, 1993). When defining the subscales, the items were classified according to (a) the primary skill area addressed and (b) the type of information given and requested in the item stem. It was found that the skill-area classification yielded information better suited for remediation than did the content-area classification.

All previously described criteria and methods for examining the performance of subscores are, more or less, subjective. However, Haberman (2008; 2005) proposed an objective criterion derived from classical test theory (CTT). The criterion is based on the conception that there is value in reporting a certain subscore if the observed subscore is a more reliable predictor of the true subscore than the observed total score is. Therefore, when applying this criterion one makes an objective comparison; the size of the subscore measure relative to the total-score measure determines whether there is value in reporting that subscore. Haberman (2008) applied the criterion to subscores on the SAT I examination, which is used for college admission, and found that "none of the section scores of SAT I math or SAT I verbal provide any appreciable information concerning an examinee that is not already provided by the math or verbal total score" (p. 221). In addition, Sinharay, Haberman, and Puhan (2007) used Haberman's criterion to examine the value in reporting subscores on a basic skills tests primarily administered to prospective or practicing teacher's aides. They concluded that there was no added value in reporting either of the subscores Reading Skills, Reading Application, Mathematics Skills, Mathematics Application, Writing Skills, and Writing Application, or the combined Reading score, Mathematics score, and Writing score. Haberman, Sinharay, and Puhan (2009) also used the criterion to examine subscores in the context of a basic skills test used for teacher certification, and Puhan, Sinharay, Haberman, and Larkin (2008) used the criterion to examine subscores from eight certification tests. Like Haberman (2008) and Sinharay et al., both Haberman et al. and Puhan et al. concluded that the subscores did not provide any useful information other than already provided by the total score.

METHOD

The potential added value of reporting subscores on the SweSAT is examined by using the approach proposed by Haberman (2008; 2005). In addition to providing an objective criterion for determining the value of reported scores, the use of the method is fairly straightforward and is based on statistics that are readily available for test scores (i.e. means, variances, correlations, and reliability coefficients). Also, Sinharay et al. (2007) noted that "another advantage of the measure suggested is that it is conceptually very close to test reliability – so the measure will be intuitively appealing to the practitioners" (p. 23).

Purpose

The main purpose of this study is to examine if there is added value in reporting subscores on the Swedish Scholastic Assessment Test (SweSAT), which is used for selection to higher education in Sweden. The motivation for performing the study is that to this date, Haberman's (2008) study is the only one that has used an objective criterion to examine the value of reporting subscores on a college admission test. Consequently, there is a need for more studies examining similar tests in order to provide a more general picture of the value of reporting subscores.

The motivation for examining the SweSAT specifically is twofold. First, the SweSAT score reporting procedure is going through some changes at the moment. For example, a few years ago the reporting of subscores (corresponding to the five subtests) to the examinees commenced. The motivation for that decision was that the subscores could provide diagnostic information and thus be useful for remediation; however, there was no empirical evidence to support the decision. Another change is that section scores will be provided and most likely used for admission purposes, which has been a reality for the SAT and ACT for decades. Second, examining the SweSAT allows for comparisons between tests from two different countries, that is, between US tests and a non-US test.

The SweSAT

The Swedish higher education admission system is highly centralized and a student can be admitted on the basis of three measures: (a) the grade-point average (GPA) after the last year of high school, (b) the SweSAT score, and (c) criteria locally determined by each university. What is rather unique is that candidates are admitted on the basis of one of these measures, not a combination of them. According to admission regulations, at least one third of candidates should be admitted on the basis of SweSAT scores, and in practice this number is about 40 percent.

The SweSAT is a norm-referenced, paper-and-pencil, multiple-choice test with five subtests: Vocabulary (WORD; 40 items), Swedish Reading Comprehension (READ; 20 items), English Reading Comprehension (ERC; 20 items), Data Sufficiency (DS; 22 items), and Diagrams, Tables, and Maps (DTM; 20 items). The test is administered twice a year with approximately 30,000–50,000 examinees per administration. As previously stated the SweSAT reports only an overall composite score for use in the admission process, while subtest scores are provided to examinees only. The two section scores proposed for use in the admission procedure are a Verbal score (WORD + READ + ERC) and a Quantitative score (DS + DTM). The motivation for using section scores is that they are assumed to increase the predictive validity of the test, especially in programs where mathematics and quantitative reasoning dominate the curriculum, such as engineering. The five subtests are intended to measure fairly distinct constructs, so there is some logical evidence for reporting the subtest scores. Further, the SweSAT test development procedure is standardized and rigorous, so there is some procedural evidence as well. However, there have been no studies aimed at gathering objective empirical evidence to support the reporting of subscores.

Data

The data consisted of examinees' scores from five consecutive administrations of the SweSAT; the 2006 spring and fall administrations (06A and 06B), the 2007 spring and fall administrations (07A and 07B), and the 2008 spring administration. Scores were available at the item, subtest, and test level. The number of examinees was 41,530, 29,787, 38,469, 26,610, and 37,432 for the five administrations respectively. Scores from all five separate subtests of the SweSAT as well as the two section scores are examined.

Subscore predictors

According to Sinharay et al. (2007), whenever a subscore is reported to an examinee the goal from a CTT perspective is to predict the examinee's true subscore S_t from his or her observed subscore S. This implies that for a subscore to have added value it should provide a more accurate measure of the construct it purports to measure than is provided by the total score (Haberman, 2008). In terms of predictions, this means that for a subscore to have added value, "*the true subscore should be predicted better by a predictor based on the observed subscore than by a predictor based on the total score*" (Sinharay et al., 2007, p. 23). If this condition is not satisfied, then any selection decision made or diagnostic information provided on the basis of the subscore will have more errors than the corresponding one based on the total score.

The predictors proposed by Haberman (2008) are based on linear regressions for approximations of the true subscore S_t. The regression formula for S_s, the predictor based on the observed subscore S, is equivalent to Kelley's (1947) formula applied to subscores,

$$S_s = E(S) + \rho^2(S_t, S)[S - E(S)],$$

where $\rho^2(S_t, S)$ is the subscore reliability.[1] Further, the regression formula for S_x, the predictor based on the observed total score X, (see e.g. Equation 45 in Holland & Hoskens, 2003) is

$$S_x = E(S) + \rho(S_t, X)\frac{\sigma(S_t)}{\sigma(X)}[X - E(X)],$$

where $\sigma(X)$ is the standard deviation of the observed total score, $\sigma(S_t) = \sigma(S)\sqrt{\rho^2(S_t, S)}$ is computed using the values of the observed subscore variance and the estimated KR-20 reliability, and $\rho(S_t, X)$ denotes the correlation between the true subscore and the observed total score. Further,

$$\rho(S_t, X) = \sqrt{\rho^2(S_t, X_t)\rho^2(X_t, X)}, \quad (1)$$

where $\rho^2(X_t, X)$ is the total score reliability estimated by the KR-20 approach and $\rho^2(S_t, X_t)$ can be estimated from

$$\rho^2(S_t, X_t) = \frac{\left[\sigma(X)\rho(S, X) - \sigma(S)[1 - \rho^2(S_t, S)]\right]^2}{\sigma^2(X)\rho^2(S_t, S)\rho^2(X_t, X)} \quad (2)$$

where $\sigma(S)$ is the standard deviation of the observed subscore, $\rho(S, X)$ is the correlation between the observed subscore and the observed total score, and $\rho^2(S_t, S)$ and $\rho^2(X_t, X)$ are the subscore and total score reliabilities estimated by the KR-20 approach. It should be noted that (1) and (2) are the only equations that have to be used when examining the subscores. (Complete derivations are available from the author.)

Criterion for comparing predictors of the true subscore

Haberman (2008) suggested the proportional reduction of mean squared error (PRMSE) of the predictors S_s and

S_x compared to the mean squared error (MSE) of the trivial predictor $E(S)$ as a criterion for comparing predictors of true subscores. The trivial predictor $E(S)$ gives the same prediction of the true subscore for all examinees (i.e. the average subscore). Before describing the PRMSE in more detail, the MSEs for the predictors S_s and S_x and the trivial predictor are described.

The MSE of a predictor is reasonable to use in this context because a larger MSE implies more error in the decisions made on the basis of that score. The MSE for the trivial predictor $E(S)$ is $\sigma^2(S_t)$, the MSE for the predictor S_s is $\sigma^2(S_t)[1-\rho^2(S_t,S)]$, and the MSE for the predictor S_x is $\sigma^2(S_t)[1-\rho^2(S_t,X)]$. Now, the PRMSE for any predictor A of S_t with finite mean and variance is

$$\Psi(S_t,A)=1-\frac{E([S_t-A]^2)}{\sigma^2(S_t)},$$

so that $\Psi(S_t,E(S))=0$ (Haberman, 2008). Then it can be shown that the PRMSE for the predictor S_s is $\rho^2(S_t,S)$, which is the subscore reliability. Similarly, the PRMSE for the predictor S_x is $\rho^2(S_t,X)$, which is estimated from (1) and (2) and can be thought of as the reliability of the observed total score as a measure of the subtest construct. Thus, for a subscore to have added value the subscore reliability $\rho^2(S_t,S)$ must be larger than $\rho^2(S_t,X)$, which makes sense in terms of correlations as well (we expect $\rho(S_t,S)$ to be larger than $\rho(S_t,X)$ for a subscore to have added value). Sinharay et al. conclude that "a subscore will be favored as the subscore reliability increases …, the total score reliability decreases, and the correlation between true subscore and true total score decreases (which will happen when the subtests measure very different skills)" (p. 24).

Corroborating the results. One way of corroborating the results for the subtest scores is by studying the effective weights of the different subtests. An effective weight (Wang & Stanley, 1970; see also Petersen, Kolen, & Hoover, 1989; Kolen & Brennan, 2004) is an index of the contribution of the subtest to the total test. The more related a score is to the other scores, the greater its effective weight (Petersen et al., 1989). Also, the more related a score is to the other scores, the more related that score is to the total score and thus the less the value of reporting that score. The standard definition of an effective weight is the covariance between the subtest score and the total score. However, because of the large differences in subtest lengths on the SweSAT there will be large differences in subtest variances and covariances. Consequently, the effective weights will be more dependent on test length than on the actual relationships between the subtests. It therefore makes more sense to calculate the effective weights based on scores that are placed on the same scale, for example by standardizing scores to have unit variance. When doing so, the effective weight of a subtest score is $\varepsilon_i=1+\sum_{j\neq 1}\rho_{ij}$, where ρ_{ij} is the correlation between tests *i* and *j*.

RESULTS

Table 1 shows the PRMSEs for subscores from the individual subtests and two sections for five forms of the SweSAT. The PRMSEs are larger for S_s, the predictor based on the observed subscores, than for S_x, the predictor based on the observed total score, for three subtest scores (WORD, DS and DTM) and both section scores across all test forms. This implies that there is added value in reporting these scores. It seems clear that the high subscore reliability estimate contributes to the added value for the WORD subtest, because the average PRMSE for S_x is not particularly low when compared to the other subtest scores. A high subscore reliability estimate, as well as a relatively low PRMSE for S_x, also contributes to the added value for the DS subscore. For the READ subscore the PRMSEs is smaller for S_s than for S_x across all five test forms, indicating that there is no added value in reporting the READ scores. The reason is not that READ scores have low reliability, but rather that the observed total score correlates well with the true READ score and thus is a good measure of reading comprehension.

Compared to the other subtests the DTM subtest has relatively low PRMSEs for both predictors; it has the lowest average reliability estimate and the second lowest correlation between observed total score and true subscore. In spite of the relatively low reliability, there is still added value in reporting DTM scores due to their relatively weak relationship with the total score. The PRMSEs for the predictors of the ERC true score are quite similar, and differ on average by a mere 0.02 units. Yet, the PRMSEs do indicate that there may be value in reporting ERC scores.

Table 1: Estimated Proportional Reduction of Mean Squared Error (expressed as percentages) for Five Subtest Scores and Two Section Scores from Five Test Forms

Test form	PRMSE	Subtest scores					Section scores	
		WORD	DS	READ	DTM	ERC	V	Q
06A	$\rho^2(S_t,S)$	**87**	**80**	74	**72**	**78**	**92**	**86**
	$\rho^2(S_t,X)$	75	61	**81**	66	75	86	67
06B	$\rho^2(S_t,S)$	**85**	**79**	69	**72**	**77**	**90**	**86**
	$\rho^2(S_t,X)$	68	63	**78**	64	73	84	67
07A	$\rho^2(S_t,S)$	**86**	**78**	77	**71**	74	**91**	**85**
	$\rho^2(S_t,X)$	69	61	**81**	62	74	85	66
07B	$\rho^2(S_t,S)$	**87**	**75**	66	**72**	**76**	**91**	**84**
	$\rho^2(S_t,X)$	74	61	**77**	67	75	86	67
08A	$\rho^2(S_t,S)$	**83**	**77**	74	**68**	**76**	**90**	**83**
	$\rho^2(S_t,X)$	70	62	**78**	65	75	85	68
Average	$\rho^2(S_t,S)$	**86**	**78**	72	**71**	**76**	**91**	**85**
	$\rho^2(S_t,X)$	71	61	**79**	65	74	85	67

Note. For each test form and subscore, the larger of the two PRMSEs, indicating the relatively better predictor, is boldfaced.

Table 2: Intercorrelations and Effective Weights for the Five Subtests

	Total test	WORD	DS	READ	DTM	ERC	Effective weights
WORD	.84	1.0	.40	.62	.43	.60	3.05
DS	.75	.40	1.0	.51	.66	.51	3.07[a]
READ	.81	.62	.51	1.0	.50	.63	3.26
DTM	.74	.43	.66	.50	1.0	.49	3.08
ERC	.81	.60	.51	.63	.49	1.0	3.23

Note. The correlations are averages from the five administrations of the SweSAT examined in this study. The effective weights are calculated assuming the subtest scores have been standardized.

[a] The disagreement between the value of the effective weight and the sum of the correlations is due to rounding.

Both section scores have rather high reliability, but the total score seems to be more related to the Verbal score than to the Quantitative score. This is not surprising as the Verbal section has almost twice as many items as the Quantitative section. Still, there is added value in reporting both section scores.

Table 2 displays intercorrelations and effective weights of the subtests. The table shows that the effective weights of the READ and ERC scores are similar and higher than those for the WORD, DS and DTM scores. This indicates that there are less of the WORD, DS and DTM constructs in the total test than there are of the other subtest constructs.

DISCUSSION

The general conclusion is that there is added value in reporting all the examined subtest and section scores, with the exception of the READ score. The main reason for this divergence is that the observed total score is a better predictor of the READ true score than the READ observed score is. This implies that the total score is a better (i.e. more reliable) measure of reading comprehension than the reading comprehension score

itself. The most probable explanation is that all the other subtests, possibly with the exception of the vocabulary test, require some degree of reading comprehension. This is confirmed by a study of the latent structure of the SweSAT (Carlstedt & Gustafsson, 2005), which indicated that the Crystallized intelligence (*Gc*) factor has the strongest association with the SweSAT scores and that the two reading comprehension scores (READ and ERC) are the ones most strongly associated with this factor. Further, the effective weights indicate that it is more likely that there is added value in reporting the WORD, DS, and DTM subscores than it is in reporting READ and ERC scores, which is consistent with the main findings in this study.

The prospective use of Verbal and Quantitative scores in the admission process is supported by the results in this study. The use of section scores in the admission process will most likely lead to increased predictive validity, which hopefully will lead to an increase in the total amount of validity evidence gathered around the SweSAT. In the process, it is important to bear in mind the need for supporting documentation of any major changes made to the test or in the proposed use or interpretations of the test scores. If section scores are used for admission purposes then the original purpose of the test, which is to be a test of general developed ability used equally across all educational programs, will have changed. This is a major change in the proposed use and interpretation of the test scores, and therefore all users of the admission system, especially the colleges/universities and the examinees, need to be informed about such a change.

Another important issue to consider when introducing additional scores is the issue of equating. The five SweSAT subtest scores are reported to individual examinees for remedial purposes only and thus need not be equated with scores from previous administrations. However, scores being used for admission purposes need to be equated, and the critical question is how the score equating procedure should be executed during the transition from reporting only the total score to reporting the total score and the two section scores. Issues regarding equating designs and methods need attention and further research.

The results of this study are contradictory to the results found by Sinharay et al. (2007) on the basic skills test and by Haberman (2008) on the SAT I examination, who concluded that there was no added value in reporting section scores. The cause of the disparate results on two such seemingly similar tests as the SAT I and the SweSAT is not apparent. However, there are two potential explanations: (a) differences in the examinee groups, and (b) differences in the composition of the tests. While both tests are taken mainly by high school seniors, about half of the SweSAT examinees are 21 years or older. The analysis was consequently rerun on one administration of the SweSAT with the examinees who were at most 20 years old; yet, there was no practical difference in results. Regarding the composition of the tests, there are differences. The most notable difference is that the SweSAT subtest WORD makes up almost a third of the total score (40 of 122) and half of the Verbal score (40 of 80), while the examined version of the SAT I from 2002 had no similar items measuring vocabulary out of context but 19 analogy items. The analysis was rerun on one administration of the SweSAT, with the vocabulary test excluded. A comparison of the PRMSEs showed that both section scores would be worth reporting in this case as well; however, only two of the four subtest scores (DS and ERC) would be worth reporting. In this case the most defensible decision would be not to report the subtest scores, which is consistent with the findings on the SAT I. The disparate results from examinations of the SAT I scores and the SweSAT scores as well as the manipulated SweSAT data emphasize the need for practitioners and test developers to examine empirical evidence before deciding on which scores should be reported, irrespective of whether it concerns a test under development, an already existing test, or a test that is undergoing compositional changes.

There are some operational diagnostic programs that seem promising. For the SAT there is an online diagnostic score reporting system called SAT Skills Insight®, which is based on behavioral anchors. When a test-taker enters the score band that he or she scored within, the system provides information including (a) academic skills, listed by skill group, that are typical of students who score within the same score band, (b) suggestions on how test takers can move beyond their score bands in each content sections, and (c) selected sample questions with answers (College Board, 2008). The Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT; College Board & National Merit Scholarship Corporation, 2008a) has a section in its score report called "Improve Your Skills", where test-takers are given a personalized analysis of their areas of weakness as well as specific suggestions for how to improve (College Board & National Merit Scholarship Corporation, 2008b). For the Graduate Management Admission Test (GMAT) there is a system

for providing diagnostic information called GMAT Focus, which is based on a decision theory model (Rudner & Talento-Miller, 2007), that anchors the subscale results to the total scale performance. Under this model one can report, for example, "Your responses to items involving algebra were typical of an examinee that scored in the top 20 percentile of all test takers on the Quantitative scale" (Rudner & Talento-Miller, 2007).

Rudner and Talento-Miller compared the scores from decision theory to scores derived from item response theory (IRT) models, and made an interesting observation about the logical inconsistency in the use of overall scale IRT parameters for estimating subscale abilities:

> *By the logic of IRT, when the item parameters are on the same scale, it should not matter which set of items are used to estimate ability. … Thus all subsets, all subskill estimates, should be identical. If they are not, then either there is error in the estimate, the item parameters are wrong, or the model does not hold. Given the increased correlations with augmentation, the factor structures, the very high intercorrelations, and the relatively low number of items used to estimate subskill ability, it appears that the variance in subscale scores is due to error. Providing such subscale scores to test-takers who want to improve their overall score is tantamount to telling them to chase error.* (p. 15, emphasis added)

Reporting subskills for an IRT calibrated and scored test can be problematic. Linear tests such as the SweSAT do not have this problem. Subscore reporting that meets appropriate statistical criteria can clearly provide meaningful feedback to test takers.

While this study examined statistical criteria for reporting subscores, it is also important to examine other aspects of the score reporting procedure. For instance, the question of whether the examinees find the subscores useful for remediation is most relevant. Related questions are whether the examinees understand their scores, and whether the scores are reported in a way that facilitates appropriate interpretations and prevent misinterpretations. Theses issues have recently gained interest in the context of state and national assessments (see e.g. Goodman & Hambleton, 2004; Wainer, Hambleton, & Meara, 1999; Ryan, 2003). Yet, little or no research has been devoted to the issue of how to properly communicate scores on college admission tests to the examinees and other interested parties. Hence, there is a gap that needs to be filled.

References

ACT (2007). *ACT technical manual*. Iowa City, IA: Author.

Allalouf, A. (2003). *Scoring and equating at the National Institute for Testing and Evaluation* (Based on Allalouf, A. (1999) Scoring and equating at the National Institute for Testing and Evaluation. Research Report No. 269). Jerusalem: National Institute for Testing and Evaluation.

American College Testing (1989). *P-ACT+ Program technical manual*. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Carlstedt, B., & Gustafsson, J.-E. (2005). Construct validation of the Swedish Scholastic Aptitude Test by means of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology, 46*(1), 31–42.

Chalifour, C., & Powers, D. E. (1988). *Content characteristics of GRE Analytical Reasoning items* (ETS Research Report No. RR-88-07). Princeton, NJ: Educational Testing Service.

College Board (2008). *The SAT Program handbook*. New York: Author.

College Board, & National Merit Scholarship Corporation (2008a). *Official educator guide to the PSAT/NMSQT*. New York: College Board. Retrieved March 4, 2009, from http://professionals.collegeboard.com/profdownload/official-educator-guide-to-the-psat-nmsqt.pdf

College Board, & National Merit Scholarship Corporation (2008b). *Understanding 2008 PSAT/NMSQT scores*. New York: College Board. Retrieved March 4, 2009, from http://professionals.collegeboard.com/profdownload/understanding-psat-nmsqt-scores.pdf

Dwyer, C. A., & Villegas, A. M. (1993). *Guiding conceptions and assessment principles for the Praxis Series: Professional assessments for beginning teachers*. (ETS Research Report No. RR-93-17). Princeton, NJ: Educational Testing Service.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241–259.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380–396.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220.

Haberman, S. J. (2005). *When can subscores have value?* (ETS Research Report No. RR-05-08). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79–95.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & The Health Professions, 27*(4), 349–368.

Hanson, B. A. (1989). Scaling the P-ACT+. In R. L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+*. Iowa City, IA: American College Testing.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*(4), 345–359.

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

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*(1), 123–149.

Joint Commission on National Dental Examinations (2004). *National board dental examinations technical report*. Chicago: American Dental Association.

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking* (2nd ed.). New York: Springer.

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

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika, 30*, 239–270.

McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). *An investigation of the feasibility of obtaining additional subscores on the GRE Advanced Psychology Test* (GRE Board Professional Report No. 74-4P). Princeton, NJ: Educational Testing Service. (ERIC Document No. ED163090)

Monaghan, W. (2006). *The facts about subscores* (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. Retrieved January 29, 2009, from http://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

National Institute for Testing and Evaluation (2009). How your scores are used? Retrieved February 3, 2009, from http://www.nite.org.il/scripts/english/txt.asp?pc=145617455&selFol=776228773&flag=0

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Educational Measurement/Macmillan.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2008). *Comparison of subscores based on classical test theory methods* (ETS Research Report No. RR-08-54). Princeton, NJ: Educational Testing Service.

Rudner, L. M., & Talento-Miller, E. (2007, April). *Diagnostic testing using decision theory.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies.* Greensboro, NC: South Carolina Department of Education.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sheehan, K. M., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*(3), 255–272.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28.

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

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*(4), 301–335.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement, 37*(2), 113–140.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 4*, 663–704.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83–105.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance.* Paper presented at the annual meeting of the Psychometric Society, Montreal, Québec, Canada.

Notes

[1] Initially, Sinharay et al. (2007) also considered the observed subscore S as a predictor. However, they showed that S will always be inferior to S_s as a predictor of S_t and therefore the predictor S is not discussed in this study.

Acknowledgements

The author thanks Mats Hamrén for his data collection work, and Marie Wiberg, Christina Stage, Lawrence Rudner, and three anonymous reviewers for their valuable comments on the manuscript.

Citation

Lyrén, Per-Erik (2009). Reporting Subscores from College Admission Tests. *Practical Assessment, Research & Evaluation*, 14(4). Available online: http://pareonline.net/getvn.asp?v=14&n=4.

Corresponding Author

Per-Erik Lyrén
Department of Educational Measurement (BVM)
Umeå University
901 87 Umeå, Sweden

E-mail: per-erik.lyren [at] edmeas.umu.se