ETS®

R&D Connections
July 2006

The Facts About Subscores

William Monaghan

Policy makers, college and university admissions officers, school district administrators, educators, and test takers all see the usefulness of assessments that report subscores. These individuals each need to make decisions on various courses of action to take, and many want to use subscores to help determine how they should move forward.

However, educational researchers at assessment organizations are expressing caution when it comes to subscores. While they want to be responsive to the desires of the educational marketplace, assessment organizations are interested in the appropriate use of test scores. In particular, they want to prevent any harm coming from the misuse of test scores. Without enough supporting data and the ability to meet the criteria discussed later, educational researchers are reluctant to endorse the use of subscores for making inferences about admissions or courses of instruction, or for making inferences on the relative strengths and weaknesses of programs of instruction in schools or classrooms. Such reluctance is particularly strong when subscores are drawn from tests not specifically designed to measure what subscores purport to measure.

The focus of this paper is on the issues surrounding the reporting of subscores, including the difficulty in creating subscores from tests not designed to support them. For such tests, the total score is often a more accurate measure of an individual's knowledge or skills in a subdomain of interest than a subscore derived from only those items that purport to measure the subdomain directly.

What Are Subscores and Why Is There Such an Interest in Them?

Educational and psychological assessments are often made up of subsections based on content categories or blueprints (Haberman, Sinharay, & Puhan, 2006a). Some examples of this are an assessment of general ability or an assessment of the knowledge of elementary school teachers, who must know several subjects. Each of these assessments needs to have subsections on mathematics, reading, and writing. Subscores, or trait scores as they are sometimes called in writing assessments (Dorans, 2005), are the scores derived from these subsections, which are typically used to form the total score. (In some writing assessments, trait scores are not used to compute the overall score.)

For students, subscores are desirable because students want to see their strengths and weaknesses in different content areas and use this information to plan future remedial studies. States and academic institutions want a profile of performance for their graduates to evaluate their curriculum effectiveness better and to focus on areas that need remediation (Haladyna & Kramer, 2004). The reporting of subscores is a common proviso in the proposal requests issued by state education departments and sizeable school districts. Because of the demand, assessment organizations recognize the business value in offering subscores.

Practitioners may be tempted to turn to subscores for admissions purposes especially when they have a number of candidates with almost identical total scores on their admissions tests and who are similar in the other factors considered. If the subscores these admissions officials considered provided unique information, this would be a sensible practice. Unfortunately, this uniqueness in the information offered by subscores is not often evident as this paper discusses later.

In the labor market, the demand for subscores is also strong. Employers use criterion-based scores to make decisions and subscores to inform remediation. In addition, some employers want to be able to hire personnel based on individual proficiencies and rely on subscores to support this activity.

Subscores also conceivably fulfill another need. There is great pressure from the public to limit the number of tests students take so there is more time spent on instruction. States and school districts want to curtail the resources and expenses needed to administer assessments. Deriving more information from any one test seemingly would be a solution. The thinking is that assessment organizations obtain a vast amount of data from their tests that they can then compartmentalize to report on the individual skills of a test taker.

The Role of Assessment Design in Supporting Subscores

Assessments are designed to answer specific questions:

- What is the probability that a student will do well in a particular program of study?
- Does a candidate have the required knowledge to be certified in a certain profession?
- Has a test taker mastered the necessary subject matter?
- What concepts does the individual need yet to master?

An assessment's design is the most important factor in determining the value of reporting subscores from the assessment. In constructing a test, an assessment organization convenes committees of subject matter experts to identify the knowledge, skills, and other characteristics to assess. The assessment organization then establishes how best to measure these characteristics given the resources of the client sponsoring the program. The assessment organization works with the client to agree on what scores to report given the purpose of the test, and this may or may not include provisions for reporting subscores.

Having a provision to report subscores gives test design specialists a certain set of requirements to fulfill. They must ensure that there are enough test items assessing the given knowledge, skill, or other characteristic to produce a stand-alone score for each separate area on which results are desired. In addition, these practitioners need to include enough items to differentiate the subscore from the total score meaningfully.

If the intention is for more than one subscore, the subscore's potential utility inversely depends on how highly it correlates with other features measured in the test. Subscores that correlate too highly with the rest of the test will offer little additional information over and above what is provided by the total score.

The real issue of reporting subscores is with established assessment programs, which in most instances have not included the reporting of subscores in their original design specifications. Forces in the educational marketplace are pressuring assessment organizations to report subscores with these programs regardless of the primary purpose of the assessments.

Criteria for Reporting Subscores

Practitioners must consider several key factors before deciding to support the reporting of subscores for an individual or on an institutional level. An article in the journal *Applied Measurement in Education* (Tate, 2004), for example, emphasized the importance of ensuring that subscores be sufficiently reliable and valid in order to minimize incorrect institutional and remediation decisions.

The ETS research report *When Can Subscores Have Value?* (Haberman, 2005) argued that a subscore may be considered useful only when it provides a more accurate measure of the construct it purports to measure than is provided by the total score. Often subscores provide information that is not reliable and or is redundant to the information provided by a total score in assessments not specifically designed to report subscores (Dorans, 2005; Haberman et al., 2006a). ETS research staff members consider three criteria to determine when it is sensible to disaggregate information from the total score on an assessment:

- The information is trustworthy
- The information is not redundant with more trustworthy information
- The information from different sources can be compared and contrasted meaningfully

The Information Is Trustworthy

The scores need to have high reliability and validity, and that reliability and validity should be commensurate with the intended use of the measurement. Reliability refers to how consistently a test measures the intended constructs; validity indicates the degree to which the test assesses what it purports to measure. If a test measures the intended construct poorly or doesn't produce scores that are consistent over multiple administrations, the information it yields is not trustworthy. Using this test data in decision-making wouldn't be wise and could result in the test taker receiving the wrong remediation.

The Information Is Not Redundant With More Trustworthy Information

This criterion addresses the most common problem with the practice of reporting subscores—the fact that subscores often offer only redundant information. The total score is often a better indicator of a person's skill or knowledge in a subarea of the domain being tested than a subscore based on only those items that supposedly measure the subarea. For tests that already provide a reliable total score, producing a subscore—even a reliable one—does not yield any *additional* information. The total score and subscores often share such a high degree of correlation that one could more reasonably predict the subscores a person would receive on different forms of the test from the score on the whole test than from the test taker's subscore.

In *Methods of Examining the Usefulness of Subscores* (Harris & Hanson, 1991), the objective of educational researchers "was to examine the relationship between subscores and test scores and to determine if subscores are providing different and better information for some purposes than the test score" (p. 4). Using the P-ACT+, a practice test for an admissions assessment, the researchers found that "the assumption that the true scores on the subtests are functionally related appears to be reasonably well met for both the English and Mathematics tests" (p. 8). They concluded that for this test battery, "the two English raw subscores do not provide information distinct from the total English raw score; likewise, the two Mathematics raw scores do not appear to be providing information distinct from the total Mathematics raw test score" (p. 8).

ETS researchers examined this issue more recently on a few of the organization's assessments (Haberman, 2005; Haberman, Sinharay, & Puhan, 2006b). One of the assessments they examined was the ParaPro Assessment. Using an administration of the test, the researchers statistically analyzed the data to determine the trustworthiness of subscores estimated from candidate responses. They then estimated what the subscores would be from the total score and determined the trustworthiness of those subscores as well. Table 1 on the next page shows a comparison of the ratings given to the two sets of subscores. In each case, the measure of trustworthiness is from 0 to 100, with 100 representing the highest rating. In the comparison, the subscore estimated from the total score is more trustworthy for all of the subscore categories.

The Information From Different Sources Can Be Compared and Contrasted Meaningfully

This criterion relates to people's perception that if numbers look alike they are comparable. Look-alike numbers may or may not be comparable; to ensure numbers are comparable, assessment organizations must do special data collection and analyses (Dorans, 2005). Because subscores are derived from smaller sets of test questions than is the total score, the special data collection and

Table 1

Comparison of Subscore Trustworthiness

	Trustworthiness of subscores resulting directly from candidate responses	Trustworthiness of subscores estimated from the total score
Reading: Skills	77	84
Reading: Application	71	91
Math: Skills	77	83
Math: Application	73	88
Writing: Skills	75	81
Writing: Application	74	81

analyses—which involves some expense—is necessary to make subscores comparable with those from different tests.

Again, if the client made reporting subscores a part of the specifications during the assessment design phase, the assessment will include an adequate number of items to increase the reliability of what is being measured. Other factors that deserve consideration include the number of test takers and types of test takers. For equating purposes, assessment organizations need enough data to compare different test forms of the same assessment reasonably and to scale scores.

Unfortunately, this isn't always the case in assessment programs that report subscores. Some assessment programs only equate the total score. Each score reported—subscores included—should be subject to equating.

Alternatives to Reporting Subscores

Given the issues provided earlier, educational researchers have been working on alternatives to reporting subscores, particularly to provide test takers with diagnostic information. The alternative is often to produce skill profiles. To do so automatically is to generalize what individuals at preset skill levels are able to do and to see which skill level group the test taker's set of skills most closely match. In most cases, the skill profile will refer teachers and students to text that collectively describes what individuals with relatively the same skill set can do. The intention is for teachers to use these descriptions in a formative way to tailor their lesson plans for their students. Students benefit by being able to see reasonably objective descriptions of their skill set and are able to act on suggestions to improve.

Among the approaches that have shown some promise in set situations are general diagnostic models, scale anchoring, and the practice of using a weighted combination of subscores with the total score.

The general diagnostic model (GDM) approach (von Davier, 2005) utilizes a variety of psychological measurement and statistical models that range from item response theory and latent class analysis to skill profile models. In practice, the GDM approach enables analysts to gather evidence about the necessary number of subscores or subskill components through the comparison of models of different complexity. The approach allows the system to automatically develop skill profiles based on information on which skills are required by the tasks in the assessment.

In assessments designed to measure multiple skills or domains, modeling multiple subskills with the GDM can be relatively straightforward unless the scores on measures from the domains are highly correlated. Researchers do caution against using the GDM approach or other skill profile models on assessments designed specifically to measure a single skill or domain. In such cases, the assessment design eliminates sources of variation other than the one stemming from the intended target of inference.

Scale anchoring is a process that uses a selection of test items to describe students at different proficiency levels. Content experts select the individual test items that describe the skills and knowledge best at given proficiency levels. They usually choose test items that students at a certain proficiency level have a high probability to get correct. The content experts and assessment specialists then set percentages for the minimum number of correct responses needed for certain levels of proficiency. From the test items, the process attempts to establish what students at the different proficiency levels typically can do. However, researchers caution that this process does rely on a small set of test items and can be subjective as the views of content experts will often conflict.

The weighted combination of subscores with the total score is a statistical method that tries to produce a more reliable measure of student subscale performance. Since the total score is based on many more items than the subscore, it is often much more reliable. This method "borrows" the total score information by combining it with the subscore—giving each its appropriate weight—to arrive at a better estimate of the subscale performance of the students.

Although research to date in the approaches described above and others has been promising, even these alternative approaches are subject to meeting the conditions listed earlier in the section Criteria for Reporting Subscores as demonstrated in the following excerpt dealing with a study that investigated using the GDMs approach with ETS's Test of English as a Foreign Language™ (TOEFL®) program:

> If skill classifications and skill profile reports to clients are required for TOEFL iBT Reading and Listening, these reports should be accompanied by a note pointing out the high correlations among the skills and the effects these high correlations will have on the reports. The majority, or seven out of the eight Reading and Listening skills, are strongly correlated to overall ability, and the eighth skill is found not to correlate across test forms in the pilot data (von Davier, 2005, p. 27).

Conclusion

So in the end, are subscores worth reporting? The answer depends on a number of factors, as discussed earlier. However, the ultimate responsibility either to report subscores or not lies with the individual assessment programs and the assessment organizations responsible for them. Assessment organizations need to be responsive and responsible. They need to serve the needs of the educational marketplace and to do so in an ethical manner. They need to ensure that all of the assessment results reported "be accurate and communicate meaningfully with the intended recipients" (ETS, 2002, p. 53).

One argument about reporting subscores centers on the harm resulting from the practice. The harm comes from the use of the information. What if an admissions committee considers only one subscore and disregards a number of other factors in deciding whether to admit a candidate to a program? What is a test taker to believe if this person receives one subscore indicating strong writing skills and another subscore calling for improvement in writing skills after taking a different version of the same test a month later?

Given these situations and others, assessment organizations must try to put as much effort as is practical into ensuring the value of any subscores it reports in terms of their reliability, lack of redundancy, and equatability. Additionally, the organizations should provide those who use the assessment results with guidelines on the proper use of these scores.

References

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS RR-05-16). Princeton, NJ: ETS.

Dorans, N. J. (2005, March). *Why trait scores can be problematic.* Presentation to the ETS Visiting Panel on Research, Princeton, NJ: ETS.

ETS. (2002). *ETS standards for quality and fairness.* Princeton, NJ: ETS.

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

Haberman, S. J., Sinharay, S., & Puhan, G. (2006a). *Subscores for institutions* (ETS RR-06-13). Princeton, NJ: ETS.

Haberman, S. J., Sinharay, S., & Puhan, G. (2006b, June). *Under what conditions should we report subscores? Some research results, emerging principles, and recommendations.* Presentation to ETS Strategic Business Unit senior staff, Princeton, NJ: ETS.

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

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

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

Acknowledgements

Thanks to Tim Davey, Neil Dorans, Matthias von Davier, Kim Fryer, Shelby Haberman, Gautam Puhan, John Mazzeo, and Sandip Sinharay for contributions and comments on this article.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001

Send comments about this publication to the above address or via the Web at:

http://www.ets.org/research/contact.html