Listening. Learning. Leading.®

R&D Connections
No. 6 • August 2008

The Practice of Comparing Scores on Different Tests

Appropriate comparisons require careful data collection, analysis, and interpretation

By Neil J. Dorans

To what extent can we use scores to compare performance on one test with performance on a different test?

Scores are the end products of assessment processes. Scores are used for admissions, placement, diagnosis and a variety of other purposes. Their properties are sometimes misunderstood, taken for granted, ignored or presumed to be something they are not. Yet they have an impact on decisions that affect individuals and institutions.

Scores from different assessments are often treated as if they were interchangeable, even when they are not. The urge to make comparisons compels people—even those who should know better—to forget that each assessment is a tool crafted for a specific purpose.

To emphasize this point, there is an example that can be taken from professional sports.

Michael Jordan was thought by many to be the consummate professional athlete in the United States in the 1990s. His achievements on the basketball court led his team to three consecutive championships in 1991-93. Jordan then tried a second professional sport—baseball—at the height of his career in 1994.

Jordan, however, soon discovered that hitting a baseball at the professional level is hard to do and rejoined his former basketball team leading them to three additional championships.

As a contrast, Tony Gwynn of the San Diego Padres baseball team (1982-2001) was a hitter par excellence. Gwynn also was a standout college basketball player being drafted into the National Basketball Association before he chose a career in baseball.

Meaningless comparisons are made nearly every day by test takers, admissions officers, policy makers and the press.

On the surface, Jordan's and Gwynn's careers would allow observers to make claims about the athleticism required to compete in both sports. Should Jordan be considered any less of an athlete than Gwynn due to his subpar performance in baseball?

While it may be fun to debate this issue, making these kinds of comparisons is not something that should be done in a casual manner, if at all. Jordan's foray into professional minor league baseball is only evidence that dunking with a magisterial flourish and spraying hits all over the field are not exchangeable skills. The skill sets needed to compete at the highest levels in both basketball and baseball are different.

In a somewhat similar way, many college seniors compare their GRE® Verbal and Quantitative scores to their SAT® Reading and Math scores and draw inferences about their

08-29-2006, 10:14 AM #**6**

emengee [sic]
Junior Member
Join Date: Jun 2005
Threads: 5

Although I have not actually looked at the statistics in the recent past, I seem to remember that if you compare certain percentile scores in both the SAT I Math and the GRE General Math, you will find that scores are much higher on the GRE. So one must look at their scores bearing this in mind: if you got a 620 on the SAT and a 700 on the GRE, for example, your percentile score might actually be LOWER on the GRE. It seems to be the other way on the verbal sections. On the SAT I got a 660 or so, and on the GRE I got something like a 560, but the percentiles were nearly identical. I was actually more pleased with my GRE score than my SAT score because I was expecting a much lower score percentile-wise. So I guess the moral of the story is to look at percentiles along with your scores because they'll really tell you how well you're doing.

Figure 1 — Test takers often try to compare their performance on the SAT with their later performance on the GRE, as in this excerpt from an actual Internet discussion group (emengee, 2006). Making valid comparisons between scores on these tests is not as simple as this Internet user believes.

academic growth because they think that the scores are comparable.

Actually these scores are not comparable. The scales of the two assessments are not linked. To link them would require administering both tests to a common group of people or administering common questions on both tests.

While both of these data collection strategies are possible, they still might not yield meaningful results, especially if there were no incentive for the test takers to take the exercise seriously. Even so, meaningless comparisons—such as the comparison of GRE scores to SAT scores seen in Figure 1—are made nearly every day by test takers, admissions officers, policy makers and the press.

These comparisons might satisfy a craving for comparison, but should not be taken seriously. Such comparisons confuse more than they enlighten.

When can scores from different assessments be viewed as the same, essentially the same, pretty much the same, sort of the same, hardly the same or the same in name only? *Linking and Aligning Scores and Scales* (2007), which I co-edited with Paul Holland of ETS and Mary Pommerich of the Defense Manpower Data Center, addresses this question.

The book provides guidance about how to answer the question in practice. It examines linking issues ranging from the relatively easy task of producing interchangeable scores on alternative versions of a test to the daunting challenge of aligning state standards with the National Assessment of Educational Progress (NAEP) scales (http://nces.ed.gov/nationsreportcard/).

There are different types of linkage between scores. There are a variety of data collection designs and data analysis procedures that can be used to achieve the linkages. Features of testing situations that affect the type of linking can be divided into the following categories: test content, conditions of measurement, and examinee population. As described in Dorans et al. (2007, part 1), different types of linking scenarios that vary along these and other dimensions include equating, linking tests in

transition, concordance, vertical scaling and linking state scales to a national scale.

Equating

Equating is at the pinnacle of linking. The goal of equating is to produce interchangeable scores (Dorans et al., 2007, part 2).

Large-scale testing programs develop different editions of the same test using a common blueprint. Large representative samples of examinees, sound data collection practices, and appropriate methods are used to equate scores on these test editions. Equating ensures that examinees are treated fairly by adjusting for differences in the test difficulty of different test editions to produce interchangeable scores.

It is a common mistake to assume that all linkings are equatings. The interchangeability of scores associated with equating, however, is not achieved simply because certain numerical operations have been performed to relate scores on two different tests. Certain requirements must be met.

For example, only tests that measure the same construct can be equated. A math test cannot be equated to a reading test. Likewise, height cannot be equated to weight. A short test that produces poorly replicated scores cannot be equated to a long test that produces very stable scores. Even reading tests from different test publishers cannot be equated to each other despite what some score users might think, unless they are built to essentially the same specifications and targeted for basically the same populations.

Transitional Linkages

A different linking scenario arises when a testing program changes a test and wants to link scores across new and old versions (e.g., the SAT change from a few years ago). A change, which can be small or large, might occur in the blueprint or test assembly specifications, in the test administration conditions, or in the mode of administration.

Some changes may be easily accommodated. Some changes in mode of administration, however, such as changing the language of administration, present serious challenges to score linking. Testing programs that are in transition must ask: Are scores from the previous version of the test interchangeable with scores from the new version of the test? In our book (Dorans et al., 2007, part 3), we examine how to test for this interchangeability and discuss things to consider when comparing scores across transition lines.

Concordance

Another linking scenario occurs when there is an interest in linking scores across related but distinct tests. Typically, the tests measure similar constructs, are administered to similar kinds of examinees, and are used for the same purpose, but differ in terms of specifications and perspective.

The term *concordance* is used to describe this type of linkage. There are good, bad, and ugly concordances as described in the book. The ACT and SAT concordance (Dorans, Lyu, Pommerich, & Houston, 1997) developed in jointly by ACT, College Board® and ETS is an example of a good concordance (Dorans et al., 2007, part 4).

Without this linkage, which provides a crosswalk across the score scales of the two tests, users would have to resort to flawed forms of comparisons, such as the norms for each test. If there were no concordances, students, parents, and admissions staff would use percentiles, a practice that presumes that the ACT and SAT groups were equivalent, which they are not.

The average ACT composite score in 2004 was 20.9, while the average SAT V+M (Verbal and Math combined) score was 1026. Would it be correct to presume that a 20.9 was equivalent to a 1026? No, because the ACT and SAT norms groups are not equivalent.

The concordance developed by ACT and College Board in 1997 reports that 21 corresponds to 1000, while 1020 corresponds to

22. This disparity shows that in 2004, the group that took the SAT scored more than a point higher on the ACT scale and a more than 25 points on the SAT V+M scale: **1026**>1020(22)>1000(21)>(**20.9**) than did the group that took the ACT. A concordance contradicts a common sense misconception of equivalence between groups and leads to fairer treatment of test takers.

Vertical Scaling

Another linking scenario arises when there is an interest in making comparisons of performance across different levels of difficulty for a given construct.

Vertical scaling, a term used to describe linkages in this scenario, used to be practiced by a small group of psychometricians responsible for a few nationally-standardized primary- and secondary-school achievement test batteries (Dorans et al., 2007, part 5).

Vertical scaling has received a great deal of attention since states have begun creating their own assessments. In the realm of K–12 testing, test scores are often compared across grades even though test content and test populations differ. Linkages of this sort must ensure that the comparisons are meaningful despite the changes in content and examinees, and sometimes assessment conditions.

Relating Group-Level Scores to Individual-Level Scores

Linking group-level scores to individual-level scores presents a particular set of challenges. The accountability movement has triggered an interest in making meaningful quantitative comparisons across scores on NAEP and assessments designed to measure whether individuals meet state standards (Dorans et al., 2007, part 6).

This is the final linking scenario described in the book, in which there is an examination of the procedures used for these linkages and a discussion about the validity of the results.

Uncommon Measures: Equivalence and Linkage Among Educational Tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999) was produced for the National Research Council by the Committee on Equivalency and Linkage of Educational Tests. That report stated that it was not feasible to link the scales of high-stakes state assessment tests to the NAEP scales.

The descent of linking is steep from the relatively lofty state of equating two editions of a well-crafted test to settings in which group-based scores, estimated from the responses obtained from potentially unmotivated individuals, are linked to the scales of high-stakes tests.

Still, there remains the pervasive tendency to compare the incomparable, despite the fact that the capacity to do it properly is limited.

Linking and Aligning Scores and Scales represents our attempt to clarify when it does and does not make sense to compare scores from different tests.

References

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT Assessment and recentered SAT I sum scores. *College and University*, *73*, 24-34.

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

emengee. (2006, August 29). Relationship between SAT I Math and GREs [Msg 6]. Message posted to http://talk.collegeconfidential.com/graduate-school/232488-relationship-between-sat-i-math-gres.html

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

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