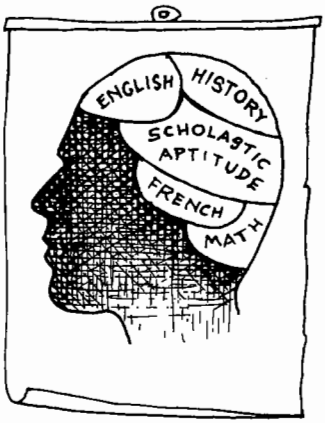

CALIBRATING COLLEGE BOARD SCORES*

William H. Angoff *Educational Testing Service*

MOST DIRECTORS of testing programs are reluctant to use the same form (edition) at different administrations of a mental test and prefer instead to introduce different forms at different times. Their reluctance is understandable. Continued reuse of the same form encourages the collection of files of test questions and makes it possible for some students to acquaint themselves in advance with the questions that will later appear on the test. This practice not only invalidates their performance on the test, but it clearly works to the unfair disadvantage of those students who have not had access to the items. There are other problems too. For example, the measurement of growth, practice, fatigue, and so on, which require two or more administrations of a test, is often

* The present article draws heavily upon "How We Calibrate College Board Scores," which the author published in *The College Board Review*, No. 68, (Summer) 1968. We express our appreciation to the College Entrance Examinations Board for their permission to publish this revision.

rendered infeasible because the second measurement is contaminated by the student's recollection of the questions he was exposed to at the time of the first testing. Although many such problems can be circumvented by the practice of using different test forms at different administrations, this practice brings with it other problems that also require solution. In any testing program which makes use of a number of different forms of the same mental test there will inevitably be variations in difficulty from form to form. Therefore, if the scores of individuals who take the different forms are to be compared with one another for the evaluation of their relative abilities, it is necessary in the interests of equity to calibrate, or "equate," the scores on the different forms.

The process of equating is a statistical one which in our testing programs at Educational Testing Service (ETS) ultimately yields an equation for converting raw scores to scaled scores. Thus, except for random error, one can assume that a student who has earned a scaled score of, say, 563 on a particular test form would have earned that same scaled score whether he had taken a more difficult or less difficult form of the test than the one he actually took. That is, the score of 563 that is reported to him is *his* score and represents *his* level of performance (and, by inference, his level of ability) at the time that he took the test.

More generally speaking, the score of 563 represents a particular level of ability—let us say, verbal ability as measured by the verbal section of the Scholastic Aptitude Test (SAT)—in the same sense that the measurement 62°F represents a measurement (much more precise, of course) of temperature. It is a measurement which is independent of the type of Fahrenheit thermometer used, the time of year in which the measurement was taken, or the temperatures of other places in the world at the time. Similarly, the score of 563 is taken to represent the same level of ability, whoever earns it, in the same sense that 62°F represents the same degree of temperature, whether the object measured is air, water, coffee, or a martini.

THE MEANING OF SCORES

The questions are sometimes asked "What does the score signify? Is it high or is it low? How far is it above the average?" Once again, the temperature analogy is appropriate because the same questions may be asked with respect to the measurement of 62°F: "What does 62° mean? Is it warm or is it cool? How close is it to the average?" It needs little elaboration to say that these questions cannot be answered as they are stated. Sixty-two degrees represents a high temperature when compared with mean temperatures in New York City in January; it represents a low temperature when compared with mean temperatures in New York in July. At any given time it is warm compared with temperatures at the poles, but cool compared with most temperatures at the equator. It is cool for acceptable morning coffee and, most

experts will agree, intolerably warm for a martini. Similarly, 563 is high or low, promising or disappointing, acceptable or unacceptable, depending on the choice of the particular reference group and on the standards set in the educational endeavor to which the student aspires.

The question that *should* be asked in interpreting scores is "How does this student compare with other students who are competing with him?" or, more fundamentally, "Is it possible to compare this student with other students, even though they may not all have been measured with precisely the same form of the test?"

Viewed slightly differently, the purpose of equating is to maintain a constant scale over time in the face of changing test forms and different kinds of students. Only if this purpose is achieved will it be possible to compare students tested today with students tested five years ago, to plot trends, and to draw conclusions regarding, for example, the effects of practice, the effects of growth, the effects of changing curricula, or the effects of changes in the composition of the student group over the course of time.

CONSTANCY OF SAT SCALED SCORES

One popular misconception is that SAT scaled scores, which are expressed on a 200–800 scale, are reported "on a curve," separately defined and separately determined at each administration of the test. This type of scaling is intentionally *not* carried out. For if it were, it would be impossible to compare students tested at different administrations, since the average scaled score for all administrations, by definition, would be the same. Moreover, under such a system a student's scaled score would depend in part on the caliber of the group with which he happened to take the test. It would be to his advantage, therefore, to take the test with a generally lower-scoring group, because he would stand relatively high in comparison with that group, and would consequently receive a higher score than if he were to take the test with a higher-scoring group. The scale that is used in the College Board program (among others) is a constant scale, defined once and *only* once, and perpetuated in that form from that time on.

To do this we construct for each form of a test an equation by which raw scores on a test form are converted to scaled scores. The methods that are followed in the equating process are so designed as to produce an equation that is *characteristic of the test form itself* and relatively unaffected by the nature of the group of individuals on whom the data were collected to form the basis of the equating process.

The equating of raw scores on two forms of a test requires an assessment of the relative difficulty of those forms. This requirement implies that ideally the same group of individuals should take both forms. But because the data for equating are drawn from operational administrations at which only one form is administered to each student, two separate groups of students must

be chosen for analysis, one group taking one of the two forms and the other group taking the second. However, these two groups are likely to be different, with respect to both average ability and dispersion (spread) of ability. Therefore, any evaluation of the relative difficulties of the forms on the basis of a direct examination of the data for these two groups could well be misleading and biased. Such an evaluation could easily result in a conversion equation for Form X that is contaminated by the characteristics of the groups rather than one that is based solely on the characteristics of the test forms. For example, if the group taking Form X is brighter, we might erroneously decide that Form X is easier. Some means, therefore, is needed to adjust for differences between the two groups.

The device used for making these adjustments is a short "equating" test administered to both groups, A and B, at the time that they take the regular operational test. Sometimes the equating test is a separately timed test administered during the course of the testing session; sometimes it is not a separate test at all, but instead, a collection of questions interspersed throughout the operational test. This collection of questions, nevertheless, is treated statistically as though it were a separate test. At the time that the equating of the operational forms (Forms X and Y) is carried out, scores are derived for the two groups on *both* the equating test *and* the operational forms that were administered to them. Appropriate formulas are then applied to the statistics observed for the two groups to yield estimates of the behavior of the two forms as though they had been administered to the same group.

If the equating test is to be used as a basis for comparing the two groups and making adjustments for differences between them, then it must represent precisely the same test for both groups. The restrictions are easy to satisfy when the equating test is separately timed; in most instances a separately timed equating test and its directions need only to be reprinted.

The restrictions are not so easily satisfied, however, when the equating test is a collection of individual questions interspersed throughout the test. Here special care must be taken to avoid differential contextual effects. For example, questions such as those in reading comprehension or data interpretation sections (questions that pertain to a single passage or to a single graph or set of numerical data) should ordinarily be used as a block because there is a real possibility that they will be interdependent. If they are, any change in the composition of the block could well disturb the meaning of the individual questions. Matching questions (e.g., questions that call for matching people with events, works, philosophies) must be taken as a group. Care should also be taken to put the equating questions in one form in about the same relative position as in the other form. It is advisable to avoid using equating questions that appear near the end of the test, where failure to answer the questions may be caused as much by insufficient time to respond to them as by their inherent difficulty.

The conversion equation for a test form provides a description of its overall difficulty and discriminating power because it essentially "locates" or "places" the test form on the scaled-score scale. For example, if a test form is relatively easy, then the scaled score corresponding to a given raw score will be lower than the corresponding scaled score on a more difficult form. A score of 57, for example, on an easy test form might correspond to a scaled score of 590. On a more difficult form the very same raw score of 57 might correspond to a scaled score of 640. Obviously, the score of 640 can be earned on the easier form only by getting a raw score higher than 57. This result is intuitively equitable, since the successful completion of 57 difficult questions *merits* a higher scaled score than the successful completion of 57 easy questions.

SCALING ACHIEVEMENT TEST SCORES

The foregoing discussion dealt with the problem of adjusting the scores on alternate and interchangeable forms of a test, so that a person of given ability will earn the same scaled score regardless of the form of the test he happens to have taken. The type of solution that is given here is appropriate to the problem that we face when, for example, we wish to compare or merge data for individuals or groups of individuals who have taken different forms of the same test. From a theoretical point of view, at least, this is a relatively simple problem. The problem is different and far more complex, however, when we wish to compare two individuals (or groups) who have taken entirely different tests, for example, in chemistry and in French.

Of course, from a logical or educational point of view it makes no sense to compare the scores of two individuals who have taken entirely different tests. But like it or not, such comparisons are inevitable, and logical or not, they are made. Just as the grade-point average is a composite of marks earned by different students in different combinations of courses and just as it is used for comparing and evaluating the relative accomplishments of different individuals (e.g., for determining pass, fail, and honor status), scores on different Achievement Tests are also compared, merged, averaged, and ranked in many college admissions offices as they must be. Recognizing as a fact of life that incomparable things *will* be compared, it behooves us to construct a system that, while it cannot be wholly satisfactory, will represent an improvement in the status quo and avoid some of its obvious imperfections. The problem and its solution are as follows.

In the College Board Admissions Testing Program, all candidates typically take the Scholastic Aptitude Test. No option is given. In the Achievement Test series, however, there are options. Candidates may take one, two, or three of the 16 Achievement Tests in the series; moreover, they may take any one, two, or three they feel inclined or prepared to take. Therefore, while virtually all candidates take the SAT, the group of candidates taking any

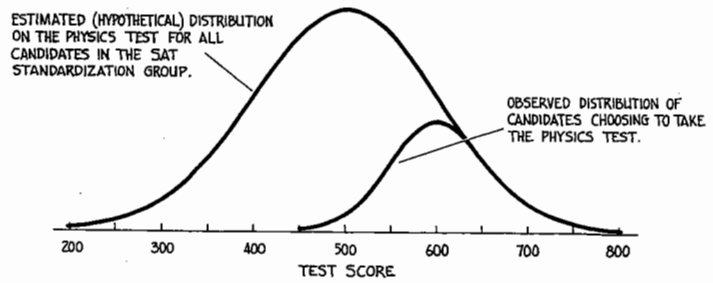

FIGURE 1

Hypothetical distribution of physics test scores for everyone and for students electing to take the physics test. Source: Angoff (1968)

one of the Achievement Tests is a self-selected group, each one different from the next in overall level and dispersion of ability. The problem arises when we wish to compare the scores of the groups of students who take different Achievement Tests.

In order to make these comparisons more equitably, the system of scores reported for the various Achievement Tests (which all use the same 200–800 score scale used for the SAT) is constructed to reflect the level and dispersion of abilities (as measured by the SAT) of the students who characteristically take each of the tests. The statistical process that results in this type of scale construction involves making an estimate of the performance on each Achievement Test of the entire original College Board standardization group for the SAT, assuming that they all had the appropriate instruction in the subject and all took the test. As a result, the scores for a test such as the physics test, which is typically taken by a relatively high-scoring segment of all the candidates who take the SAT, are automatically "placed" relatively high on the scale. Correspondingly, the scores for a test which is typically taken by a low-scoring segment of all the candidates who take the SAT are "placed" relatively low on the scale. A schematic diagram describing the distribution of scores on the physics test for *all* candidates taking the SAT and also for those choosing to take the physics test appears in Figure 1.

PURPOSE OF SCALING

The purpose of this type of scaling is to ensure that a candidate who chooses to compete with more able candidates is not put at a disadvantage, that is, that a candidate who is average in a highly selected group of candidates will earn a higher scaled score than a candidate who is average in a less able group. The

intent is to make it impossible for a candidate to "beat the game" by taking advantage of the machinery of the testing program and making a *strategic* choice of the particular Achievement Test (or tests) that will yield the highest score(s) for him. Moreover, since the scores on the various tests are scaled in accordance with the abilities of the candidates who typically choose to take them, it is possible for college admissions officers to take an average of the scores offered by each candidate with relative confidence that the average represents an equitable basis for comparing students who have taken different combinations of tests.

Given, then, that there are wide variations in the ability levels (and dispersions) of the groups of candidates taking the various tests, and given also that the scales for the tests reflect these variations, one would expect that the highest possible scaled scores on the various tests would vary substantially and systematically from test to test. As a result, it would be possible for an able student of physics to earn a higher scaled score than an equally able student of a subject that is generally chosen by less able students, simply because the scale for the physics test permitted it. In order, then, to equalize the opportunities for high scores among the different subgroups of candidates, the maximum score of 800 is imposed across the board for all tests, and the test specifications for each of the Achievement Tests are so written as to ensure, as nearly as possible, that a score of 800 may be achieved on every test and on every form. Similarly, the minimum score of 200 is imposed across the board for all tests; any raw score that would ordinarily "scale out" below 200 is reported as 200. However, because precise discriminations in the vicinity of 200 are not as often necessary on the Achievement Tests, there is no corresponding effort to ensure that a score of 200 is possible on every test. The 200 and 800 limits simply mean that scores are not reported *beyond* those limits. The principal reasons for having them are: (1) as already indicated, to minimize *gross inequities* across the test offerings, and (2) to make it clear that the tests can discriminate adequately only within a limited range of scaled scores.

NECESSITY FOR REFINED SCORES

For the most part, the processes of equating and scaling the scores for College Board tests produce effects of relatively small magnitude, and thus they could be regarded as only a pedantic refinement. But that is not so. They are a refinement of a basic sort, introduced to ensure that no student will be put either to an advantage or a disadvantage simply because he happened to have taken an easier or more difficult form of the test or because he happened to have taken the test with a less able or more able group of students.

Fundamentally, what is desired is a measurement of ability that will serve the best interests of the students, a measurement that is not only accurate and relevant for them, but also, and just as importantly, equitable. It is to this goal of equity that the "refinement" of equating and scaling is directed.

ments. If we can do a properly designed experiment, we are in a much better position to draw valid conclusions about what causes what, but the possibility of a designed experiment is not always open to us. When we can't experiment, we must do what we can with available data, but this doesn't mean that we shouldn't keep our eyes open to the faults that such data have.

CONCLUSIONS

So what have we learned from our look at sports statistics? We have learned these do's and don'ts:

(1) Don't waste time arguing about the merits or demerits of something if you can gather some statistics that will answer the question realistically.
(2) If you're trying to establish cause-and-effect relationships, do try to do so with a properly designed experiment.
(3) If you can't have an experiment, do the best you can with whatever data you can gather, but do be very skeptical of historical data and subject them to all the logical tests you can think of.
(4) Do remember that your experience is merely a hodgepodge of statistics, consisting of those cases that you happen to remember. Because these are necessarily small in number and because your memory may be biased toward one result or another, your experience may be far less dependable than a good set of statistics. (The bias mentioned here can come, for instance, from the fact that people who believe in the bunt tend to remember the cases when it works, and vice versa.)
(5) Do keep in mind, though, that the statistics of the kind discussed here are averages, and special cases may demand special action. This is not an excuse for following your hunches at all times, but it does mean that 100% application of what is best on the average may not be a productive strategy. The good manager has a policy, perhaps based on statistics, that takes care of most decisions. The excellent manager has learned to recognize occasional situations in which the policy needs to be varied for maximum effectiveness.

REFERENCES

E. Cook. 1966. *Percentage Baseball.* Cambridge, Mass.: MIT Press.

R. Hooke. 1967. Review of Cook (1966). *Journal of the American Statistical Association* 62:688–690.

G. R. Lindsey. 1963. "An Investigation of Strategies in Baseball." *Operations Research* 11:477–501.

VARIETIES OF MILITARY LEADERSHIP

Hanan C. Selvin *State University of New York, Stony Brook*

WORKERS ON an assembly line, students in a third-grade classroom, and soldiers in an army training camp do different kinds of work in radically different settings, but they have in common one important social relationship. They all spend a good part of their day in close contact with lower-level leaders, such as foremen, teachers, and company-level officers, both commissioned and noncommissioned. From both individual experience and empirical research we know that the behavior of workers, students, soldiers, and others in subordinate positions, at work and afterward, is significantly affected by the actions of their leaders.

The empirical study reported here shows how the actions of company leaders in twelve U.S. Army training companies affected the *nonduty* behavior of several hundred soldiers undergoing basic training. The unraveling of these effects of leadership was unusually complex. Unlike the student and the worker, who usually are subject to only one leader in the course of a