CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

1

Introduction

Standardized tests, whether to evaluate student performance in coursework or to choose among applicants for college admission or to license candidates for various professions, are often marathons. Tests designed to evaluate knowledge of coursework typically use the canonical hour; admissions tests are usually two to three hours; and licensing exams can take days. Why are they as long as they are? To answer this question, we must consider the purposes of the test. Most serious tests have serious purposes – admission to a college or not, getting a job or not, being allowed to practice your profession or not. The extent to which a test score can serve these purposes is its validity, which is usually defined as "the degree to which evidence and theory support the intended interpretations of those test scores."[1] But the validity of a test's scores is bounded by the test's reliability.[2] Reliability is merely a standardized measure of the score's stability, ranging from a low of 0 (essentially a random number) to a high of 1 (the score does not fluctuate at all). A test score that has low reliability must perforce have an even lower validity, and its usefulness diminishes apace.

Thus, the first answer to the question "Why are tests so long?" that jumps immediately to mind is derived from the inexorable relationship between a test's length and its reliability. However, even though a test score always gets more reliable as the test generating it gets longer, *ceteris paribus*, the law of diminishing returns sets in very quickly. In Figure 1.1, we show the reliability of a typical professionally prepared test as a function of its length. The figure shows that the marginal gain of moving from a 30-item test to a 60- or even 90-item one is not worth the trouble unless such small additional

[1] Linda Steinberg, October 26, 2020, personal communication.
[2] More accurately, the validity is bounded by the *square of its reliability*. So, for example, if a test's reliability is 0.90, its validity can be no higher than 0.81.

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

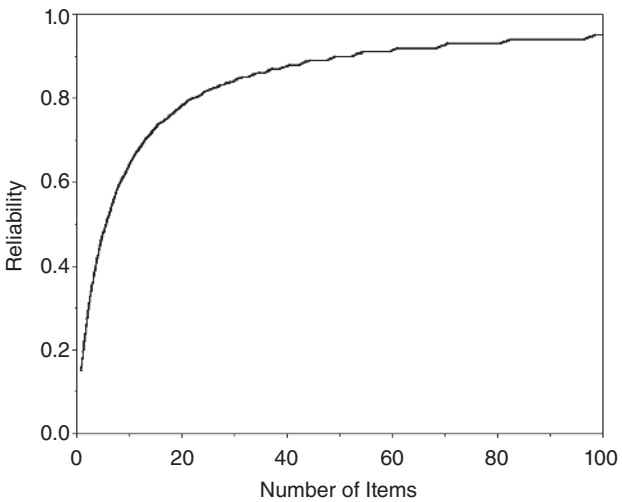

Figure 1.1 Spearman–Brown function showing the reliability of a test as a function of its length, if a one-item test has a reliability of 0.15.

increments in reliability are of practical importance. We must also note that reliability does drop precipitously as test length shrinks from 30 toward zero. But the puzzle still remains: Why are typical tests so long if, after a test's length surpasses 30, it is about as reliable as we are likely to need. So, if such extra precision is rarely necessary, why are tests as long as they are?

1.1 A Clarifying Example: The US Census

Our intuitions can be clarified with an example, the decennial US census. According to the 2020 decennial census, the United States had 331,449,281 residents; however, the Bureau of the Census estimates that this number has a nontrivial error of uncertain size. The estimate is that 782,000 more people resided in the United States than were reported, but it is also estimated that about one third of the time the estimate would be that the decennial census overcounted the population by at least 30,000 or undercounted the population by no less than 1,620,000 people. The budget of the 2020 census was $14.2 billion, or approximately $42.84 per person counted. Is it worth this amount of money to just get this single number? Before answering this question, consider the function shown in Figure 1.2, which provides the results of all decennial censuses for the past 150 years. The curve shown is a fitted quadratic function

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

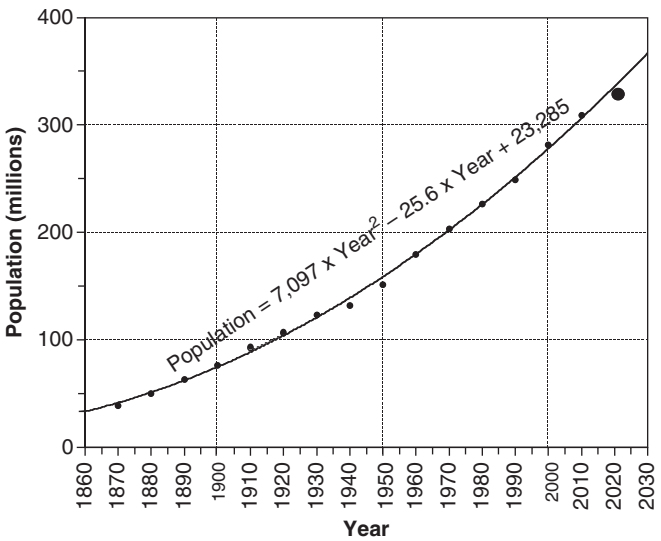

Figure 1.2 US population from 1870 to 2020.

to the data from 1870 through 2010. The large dot associated with 2020 is the actual population estimate from the 2020 census; the value of the curve, which passes slightly above it, is the estimate obtained from this fitted function – 331,943,913. The difference between the two is 494,632 people, or 0.1%, an error certainly comparable to the estimated error from the decennial census. Obtaining this estimate cost only about an hour of a statistician's time – totaling perhaps a couple of hundred dollars.

Fitting a connecting function, like the quadratic shown in Figure 1.2, can have many uses. By far, the most important purpose is to provide, at a glance, an accurate representation of more than a century's growth of a nation's population – Henry D. Hubbard (in the preface to Brinton, 1939) memorably characterized this use when he pointed out the following: "There is a magic in graphs. The profile of a curve reveals in a flash a whole situation – the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces."

Although the fitted curve provides only an approximation of the population between censuses, we can, by adopting very credible regularity assumptions, confidently use the curve to interpolate between censuses and obtain accurate estimates for any time in the 150-year range of the data.

A third use, and the one that we have illustrated here, is extrapolation 10 years beyond the 2010 census. Extrapolation, like interpolation, relies on

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

regularity assumptions, but those assumptions become more heroic the further the estimate is from the data. As we have seen, predicting the 2020 US population from prior census results ending in 2010 yielded an estimate that is likely accurate enough for most applications. Were we to use the same function to predict further into the future, we would be less sure, and our uncertainty would, naturally, expand with the size of the extrapolation. Of course, there are more data-rich methods that could improve the accuracy of such extrapolations by making their inevitable underlying assumptions more credible.[3]

And so, returning to the original question, is it worth $14.2 billion to just estimate this single number when it could have been determined as accurately in just a few minutes and be paid for out of petty cash? It doesn't take a congressional study group or the Office of Management and the Budget to tell us that the answer is no. If all the census gave us was that single number, it would be a colossal waste of census workers' time and taxpayers' money. However, the Constitution and acts of Congress require that the census enumerate the resident population and report on population sizes for very small geographical regions, including for 2020 about 11,000,000 census blocks and 73,000 census tracts. These population values for small areas are needed for apportionment of the House of Representatives and for apportionment of state legislatures and other governmental units. They, together with survey data from the American Community Survey, are also valuable for allocation of social services, among other uses. The census provides such small area estimates admirably well, but to be able to do so requires massive data collection and so incurs a huge expense. Yet the importance of providing accurate answers to many crucially important small area questions makes its impressive cost unavoidable.

There are the two key lessons we should take from this census example:

1. Obtaining an accurate estimate of the grand total is easy and cheap.
2. Obtaining accurate small area estimates is hard and expensive, and hence should not be attempted unless such small area estimates are important enough to justify the vast increases in the resources of time and treasure that are required.

[3] For example, the change in the population of the United States at any given time after a full census is the sum of four factors: (1) the number of births since the census, (2) the number of deaths since then, (3) the number of immigrants since then, and (4) the number of emigrants. When all of these are added together, we find that the net increase in the US population since 2010 is one person every 13 seconds, and so to get an accurate estimate of the total population at any moment, one merely needs to ascertain how much time has elapsed since the last estimate, in seconds; divide by 13 and add in that increment. Note that a single clerk with access to a pocket calculator in a minute or two could use this method to estimate the change from the previous census (to an accuracy of ±0.1%).

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

1.2 Back to Tests

Now let us return to tests. Instead of characterizing cost in terms of dollars (a worthwhile metric, for sure, but grist for another mill), let us instead use examinee time. Is it worth using an hour (or two or even more) of examinee time to estimate just a single number – a single score? Is the small marginal increase in accuracy obtained from a 60- or 90-item test over, say, a 30-item test worth doubling or tripling examinee time?

A glance at the gradual slope of the Spearman–Brown curve shown in Figure 1.1 as it nears its asymptote tells us that we aren't getting much of a return on our investment. And multiplying the extra hour spent by each examinee by the millions of examinees that often take such tests makes this conclusion stronger still. What would be the circumstances in which a test score with a reliability of 0.89 will not suffice, but one of 0.91 would? Off hand, it is hard to think of any.

But, returning to the lessons taught us by census, perhaps there are other uses for the information gathered by the test that require additional length – the equivalent of the small area estimates of census. In testing, such estimates are usually called subscores – small area estimates on aspects of the subject matter of the test. On a high school math test, these might be subscores on algebra, arithmetic, geometry, and trigonometry. For a licensing exam in veterinary medicine, there might be subscores on the pulmonary system, the skeletal system, the renal system, and so on. There is even the possibility of cross-classified subscores; perhaps one on dogs; another on cats; and others on cows, horses, and pigs. Such cross-classified subscores are akin to the census having estimates by ethnic group and also by geographic location.

Thus, the production of meaningful subscores needed for important purposes would be a justification for tests that contain more items than would be required merely for an accurate enough estimate of total score. What is a meaningful subscore? *It is one that is reliable enough for its prospective use and one that contains information that is not adequately focused (or is overly diluted) in the total test score.*

There are at least three prospective uses of such subscores:

(i) To provide institutions with information used for major decisions such as admission, licensure, and immigration qualification
(ii) To aid examinees in assessing their strengths and weaknesses, often with an eye toward remediating the latter
(iii) To aid individuals and institutions (e.g., teachers and schools) in assessing the effectiveness of their instruction, again with an eye toward remediating weaknesses

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

In the first case, subscores need to be highly reliable, given the life-changing decisions dependent upon them. The demands on reliability *increase* when subscores must exceed a fixed value and when multiple subscores are involved.

In the second case, helping examinees, the subscores need to be reliable enough so that attempts to remediate weaknesses do not become just the futile pursuit of noise. And, obviously, the subscores must contain information that is focused on performance on the specific topic of interest and is not diluted over the broad range of topics contained in the total test score. We might call these two characteristics of a worthwhile subscore *reliability* and *specificity*. But for a subscore to have a specific focus apart from the total score, its information must be somewhat orthogonal to the total score; hence we shall designate this characteristic of a useful subscore *orthogonality*. Shortly, we will provide an introductory discussion of each of these two important characteristics and then tell the full story in Chapter 3. But first, let us drop back in time and see where the concept and use of subscores came from.

1.3 A Brief Account of the Long History of Tests and Subtests, Scores and Subscores

The use of mental tests appears to be almost as ancient as Western civilization. The Hebrew Bible (in Judges 12:4–6) provides an early reference to testing in Western culture.[4] It describes a short verbal test that the Gileadites used to uncover the fleeing Ephraimites hiding in their midst. The test was one item long. Candidates had to pronounce the word שיבולת (transliterated as *shibboleth*). Ephraimites apparently pronounced the initial *sh* as *s*. The consequences of this test were quite severe (the banks of the Jordan were strewn with the bodies of the 42,000 who failed). Obviously, any test that consists of but a single item can have no subscores. But there were much earlier tests that were longer and had subscores.

In his 1970 *History*, Philip DuBois reported that tests had been around for millennia, and whenever they consisted of more than a single item, the appeal of computing subscores has been irresistible.

There was some rudimentary proficiency testing that took place in China around 2200 BCE, which predated the biblical testing program by almost a thousand years! The emperor of China is said to have examined his officials

[4] Judges 12:6. "Then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. Then they took him and slew him at the passages of Jordan: and there fell at that time of the Ephraimites forty and two thousand."

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

every third year. This set a precedent for periodic exams in China that was to persist for a very long time. In 1115 BCE, at the beginning of the Chan dynasty, formal testing procedures were instituted for candidates for office. Job sample subtests were used, with proficiency required in (1) archery, (2) arithmetic, (3) horsemanship, (4) music, (5) writing, and (6) skill in the rites and ceremonies of public and social life. The Chinese discovered the fundamental truth that underlies the validity of testing – that a relatively small sample of an individual's performance, measured under carefully controlled conditions, can yield an accurate picture of that individual's ability to perform under much broader conditions for a longer period of time. The procedures developed by the Chinese are reasonably similar to the canons of good testing practice used today. For example, they required objectivity – candidates' names were concealed to ensure anonymity; they sometimes went so far as to have the answers redrafted by another individual to hide the handwriting. Tests were often read by two independent examiners, with a third brought in to adjudicate differences. Test conditions were as uniform as could be managed – proctors watched over the exams given in special examination halls that were large permanent structures consisting of hundreds of small cells. The testing process was so rigorous that sometimes candidates died during the course of the exams.

This testing program was augmented and modified through the years and has been praised by many Western scholars. Voltaire and Quesnay advocated its use in France, where it was adopted in 1791, only to be (temporarily) abolished by Napoleon. It was cited by British reformers as their model for the system set up in 1833 to select trainees for the Indian civil service – the precursor to the British civil service. The success of the British system influenced Senator Charles Sumner of Massachusetts and Representative Thomas Jenckes of Rhode Island in the development of the examination system they introduced into Congress in 1868. This eventually led to George Hunt Pendleton proposing the eponymously entitled US Civil Service Act in January 1883.

The US military has arguably one of the most widely used and consequential testing programs in the United States, in terms of both number of examinees and length of time it has been in use. It also has been carefully thought through and researched over the better part of a century. Few testing programs can match the careful seriousness of its construction and use. We feel this makes it a worthy and informative illustration of the use of tests and subscores in support of evidence-based personnel decision-making. In the following narrative, we first trace the history of military testing in the United States, then move to a discussion of decisions based on total test

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

scores and the use of subscores, and then conclude with an evaluation of the success of this approach as a guide to others who would like to use both test scores and subscores to generate evidence supporting claims about individuals and groups.

1.4 The Origins of Mental Testing in the US Military

During World War I, Robert M. Yerkes, president of the American Psychological Association, took the lead in involving psychologists in the war effort. One major contribution was the implementation of a program for the psychological examination of recruits. Yerkes formed a committee for this purpose that met in May 1917 at the Vineland Training School in New Jersey. His committee debated the relative merits of very brief individual tests versus longer group tests. For reasons of objectivity, uniformity, and reliability, they decided to develop a group test of intelligence.

The criteria they adopted (described in detail on page 62 of Philip DuBois' 1970 book on the history of testing) for the development of the new group test were:

1) Adaptability for group use
2) Correlation with measures of intelligence known to be valid
3) Measurement of a wide range of ability
4) Objectivity of scoring, preferably by stencils
5) Rapidity of scoring
6) Possibility of many alternate forms so as to discourage coaching
7) Unfavorableness to malingering
8) Unfavorableness to cheating
9) Independence of school training
10) Minimum of writing in making responses
11) Material intrinsically interesting
12) Economy of time

In just 7 working days, they constructed 10 subtests with enough items for 10 different forms. They then prepared one form for printing and experimental administration. The pilot testing was done on fewer than 500 subjects. These subjects were broadly sampled and came from such diverse sources as a school for those with intellectual disabilities, a psychopathic hospital, a reformatory, some aviation recruits, some men in an officers' training camp, 60 high school students, and 114 marines at a navy yard. They also administered either the Stanford–Binet intelligence test or an abbreviated form of it. They

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

found that the scores of their test correlated 0.9 with those of the Stanford–Binet and 0.8 with the abbreviated Binet.

The items and instructions were then edited, time limits were revised, and scoring formulas were developed to maximize the correlation of the total score with the Binet. Items within each subtest were ordered by difficulty, and four alternate forms were prepared for mass administration.

By August, statistical workers under E. L. Thorndike's direction had analyzed the results of the revised test after it had been administered to 3,129 soldiers and 372 inmates of institutions for mental defectives. The results prompted Thorndike to call this the "best group test ever devised." It yielded good distributions of scores, and it correlated about 0.7 with schooling and 0.5 with ratings by superior officers. This test was dubbed *Examination a*.

In December of the same year, *Examination* a was revised once again. It became the famous *Army Alpha.* This version had only eight subtests; two of the original ones were dropped because of low correlation with other measures and because they were of inappropriate difficulty. The resulting test (whose components are shown below) bears a remarkable similarity to the structure of the modern Armed Services Vocational Aptitude Battery (ASVAB).

Test	Number of Items
1. Oral Direction	12
2. Arithmetical Reasoning	20
3. Practical Judgement	16
4. Synonym-Antonym	40
5. Disarranged Sentences	24
6. Number Series Completion	20
7. Analogies	40
8. Information	40

This testing program, which remained under Yerkes' supervision, tested almost 2 million men. Two thirds of them received the *Army Alpha*, and the remainder were tested with an alternative form, *Army Beta*, *a* nonverbal form devised for illiterate and non-English-speaking recruits. Together they represented the first large-scale use of intelligence testing.

The success of the *Army Alpha* led to the development of a variety of special tests. In 1919, Henry Link discovered that a card-sorting test aided in the successful selection of shell inspectors and that a tapping test was valid for gaugers. He pointed out that a job analysis coupled with an experimental administration of tests thought to require the same abilities as the job and a

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-41368-8 — Subscores
Shelby Haberman, Sandip Sinharay, Richard A. Feinberg, Howard Wainer
Excerpt
More Information

validity study that correlated test performance with later job success yielded instruments that could distinguish between job applicants who were good risks and those who were not. L. L. Thurstone developed a "rhythm test" that accurately predicted future telegraphers' speed.

Testing programs within the military became much more extensive during World War II. In 1939, the Personnel Testing Service was established in the Office of the Adjutant General of the Army. This gave rise to the *Army General Classification Test* (AGCT) that was an updated version of the *Army Alpha.* The chairman of the committee that oversaw the development of the AGCT was Walter V. Bingham, who served on the 1917 committee that developed Alpha. This test eventually developed into a four-part exam consisting of tests of (1) reading and vocabulary, (2) arithmetic computation, (3) arithmetic reasoning, and (4) spatial relations. Supplemental tests for mechanical and clerical aptitude, code learning ability, and oral trade were also developed. By the end of the war, more than 9 million men had taken the AGCT in one form or another. The navy and the army air forces participated in the same program but with some different tests that they required for their own purposes.

In 1950, the *Armed Forces Classification Test* was instituted to be used as a screening instrument for all services. It was designed to ensure appropriate allocation of talent to all branches. This was the precursor of the *Armed Forces Qualification Test* (AFQT) that led in turn to the *Armed Services Vocational Aptitude Battery* (ASVAB).

1.4.1 The ASVAB and Scores Derived from It

The ASVAB consists of nine subtests, each of which is scored separately. Each of those scores range from 1 to 100 and is scaled so that the mean is 50. The nine subtests are:

1. Arithmetic Reasoning
2. Mathematics Knowledge
3. Word Knowledge
4. Paragraph Comprehension
5. General Science
6. Electronics Information
7. Auto & Shop Information
8. Mechanical Comprehension
9. Assembling Objects

The scores on the first four of these subtests (Arithmetic Reasoning, Mathematics Knowledge, Word Knowledge, and Paragraph Comprehension) are combined

© in this web service Cambridge University Press & Assessment www.cambridge.org