PART I

Introduction to Testlets

1

Introduction

To count is modern practice, the ancient method was to guess.
Samuel Johnson

The subject of this book is testing. The origin of testing lies within a magical epiphany. Almost four thousand years ago, an anonymous functionary was focused on a task set before him by the emperor of China. He needed a procedure to choose wisely from among all applicants for government positions. Perhaps, utilizing a model made famous by Newton three millennia later, he was lying under a tree when the fundamental principle of testing struck him:

> *A small sample of behavior, gathered under controlled circumstances, can be predictive of performance in a much broader set of uncontrolled circumstances.*

Testing is an ingenious solution to a practical problem. In an ideal world, before choosing among candidates for some position, we would like to know how they would perform in that position. Obviously, if we could place each of them in the position for a long enough period of time we would have direct evidence, but such a solution is usually not a practical possibility. But, even in a situation in which we had, say a year of trial, we would still not know for sure how they would perform after two years, or three, or more. In this instance, however, we would be pretty convinced of the validity of the task utilized to make the decision – it is, after all, the same job. But if the tasks on the job vary, the one-year trial period may not be long enough. In addition, one year may not be long enough to allow us to be sure – our judgments about the person's performance on each of the components of the position may not be as reliable as we would like. No matter what we do, we must make some sort of extrapolating inference.

The formal Chinese Civil Service tests used during the Chan Dynasty (1115 B.C.) were but a continuation of practices that were already a thousand years old. These were job sample tests assessing proficiency in archery,

arithmetic, horsemanship, music, writing, and various social skills. The candidate's total test score was a summary of his scores on these subtests, or what has come to be called testlets. The Chinese testing enterprise, as innovative and clever as it was, did not have a formal theory for how to combine the candidate's performance on these several testlets. The development of such a theory lay almost three thousand years in the future. This book describes a theory of testlets and the pathway that was traversed in its development.

We have found that an often-useful strategy in any problem of experimental design is to start our thoughts with an optimal design, without regard to practical constraints, and then shrink back to a solution within the ever-present practical limits. By proceeding in this manner, exactly what must be assumed to be true without evidence is made explicit. It doesn't take too much of a stretch of one's imagination to conceive of that ancient Chinese functionary, whose very bones are now long dust, reasoning in this manner. He must have wondered:

i. What were the key tasks that the selected persons would have to perform?
ii. How could those tasks be represented?
iii. How can each examinee's performance be rated?
iv. How accurate is this artificial set of tasks at predicting future performance? And, if he was as clever as we reckon he must have been,
v. How accurate would this artificial set of tasks be at predicting future performance if the task constructor made modest errors in each of these steps?

These ancient issues remain the key questions that face modern testing. The first step in preparing any test is deciding what information about the examinees the users of the test scores will want to know. If we are trying to choose among applicants to college, we might want to know something about the applicants' scholarly ability, their maturity, and their grit; in short, their ability to do the myriad of tasks it takes to succeed in a collegiate environment. We probably would not be much interested in their height, weight, or eye color.

The second question is how are the tasks that we assign related to the variables of interest (construct validity). Thus, the question of primary interest might be "how well can this person write descriptive prose?" After we agree to this, we then examine the various kinds of tasks that we could ask that would provide evidence about the person's writing ability. We could ask the person to write some descriptions; we might ask the person to answer some verbal analogy items; we might even ask the person to run 100 meters as fast as she can. Or we could construct a test that was a combination of some or all of these. Obviously, each of these tasks bears a differential connection to the original

question. In practice, the choice of representation depends on many variables, some scientific and some practical.

The third question can be restated into "how can this test be scored?" At this point, it is important to make a distinction between an item and a question. An item has a surface similarity to a question, but "item" is a term with a very specific technical meaning. The term "question" encompasses any sort of interrogative:

Where's the bathroom?
What time is it?
Do you come here often?
What's your sign?

These are all questions, but they should not be confused with items. For a question to be called an item, it needs an additional characteristic:

There must be a scoring rule

Tests can be scored in many ways. Traditionally, "points" were allocated according to some sort of rule. When it was possible, items were scored right or wrong and the overall score was the number or percentage correct. With items like essays that could not be divided neatly into binary categories, raters would follow some sort of scoring rubric to allocate points. In the latter half of the 20th century, formal statistical models were introduced to take some of the guesswork out of this procedure (see Thissen & Wainer, 2001, for a detailed description of test scoring).

The fourth question, accuracy of prediction, is much more complex than might first appear. It is the fundamental issue of test validity and is composed of many pieces. Taking the simplistic view of validity as just a question of predictive accuracy assumes that one can determine a single criterion. This is rarely, if ever, the case. Moreover, even when it is the case, there is rarely a single way to characterize that criterion. Few would deny that medical schools should admit those applicants who are likely to become the best physicians. But how do you measure that criterion?

In addition, most tests have multiple purposes, and the validity of the test's score varies with the purpose. Entrance tests for college are often aggregated and then used as measures of the quality of precollegiate education. While it is reasonable to believe that the same test that can compare individuals' college readiness also contains information about their high school education, a special purpose test for the latter (probably focusing more on specific subject matter – as with the Scholastic Assessment Test (SAT) II) seems like a better option. But preparing separate specific tests for each of many closely allied purposes is

often impractical, and hence we usually approximate the test we want with the test we have.

And, finally, the fifth question, the question of robustness, which is the driving issue behind the developments described in this book: How well is any particular test score going to work if the various assumptions that must be made in its construction are not quite correct? Obviously, this is a quantitative question. If the inferences being made are inexact, the testing instrument can afford small infidelities. Millennia past, in Han Dynasty China, the selection process that utilized their tests was so extreme that anyone who survived the screening had to be very able indeed. There were thousands of candidates for every opening, and so even though there were many worthy candidates who were turned down, the ones who were chosen were surely exceptional. The screening test did not have to be great to find extraordinary candidates.

This same situation exists today. The screening for National Merit Scholars begins with about 1.5 million candidates and eventually winnows down to 1,500 winners. The tests that accomplish this do not have to be very valid – just a little – in order for the winners to be special. Of course there are many, many worthy candidates who are not chosen, but no one could argue that any of the winners were not worthy as well. Herman Wold, a member of the committee that chooses Nobel Prize winners in economics, once explained that choosing winners was easy, because in the tails of the distribution data points are sparse and there are big spaces between them. When that is the situation, the choice algorithm, and the evidence that it uses, does not have to be precise.

Testing has prospered over the nearly four millennia of its existence because it offered a distinct improvement over the method that had preceded it.

But as time has passed and test usage increased, the demands that we have made on test scores have increased, as has the "fineness" of the distinctions that we wish to make. As this has happened, tests and how they are scored have improved. These improvements have occurred for three principal reasons:

1. The demands made on the test scores have become more strenuous.
2. We have gathered more and more evidence about how well various alternatives perform.
3. Our eyes have become accustomed to the dim light in those often dark and Byzantine alleys where tests and psychometrics live.

Let us examine each of these in more detail:

Tests have three principal purposes: (i) measurement, (ii) contest, and (iii) prod. For the first 3,950 years of their existence, tests were principally contests. Those with high enough scores won (got the job, received the scholarship, were

admitted) and everyone else lost (didn't get the job, didn't get the scholarship, were rejected). For some of that time, tests also served as prods: "Why are you studying math tonight?" "I have a test tomorrow." In the modern world, with the enormous prevalence of tests, "prod" is still often the primary purpose. A test as a measurement tool is the most modern usage. It was only in the past century that a statistical theory of test scores was developed to provide a rigorous methodology of scoring that allows us to think of test scores as measurement in a way that was more than rudimentary in character (Gulliksen, 1950/1987 Lord & Novick, 1968; Thissen & Wainer, 2001). These formal developments occurred in response to users of tests who wanted to know more than merely who won. Test scores became one measure of the efficacy of the educational system. They were asked to express the effectiveness of children's learning progress. And, by comparing scores of different subgroups within society, they were used to express the fairness of that society.

As the usage of tests for these many purposes increased, evidence accumulated that supported some previously held notions, contradicted some others, and revealed some things that were never suspected (e.g., the precise character of the heritability of some mental traits (Bock & Kolakowski, 1973)). The multiple-choice item was developed for ease of scoring, but as evidence accumulated, it was discovered that it could provide valid evidence of skills and knowledge well beyond what its inventors could originally have even dreamed. One example of this (there are many others) is revealed in the study of writing ability (Wainer, Lukele, & Thissen, 1994). Few would think that multiple-choice items were a sensible way to measure writing ability, yet in fact when three 30-minute tests were given, in which two were essays and one was a mixture of verbal analogies and other typical multiple-choice verbal items, it was found that examinees' scores on the multiple-choice test correlated more highly with either of the two essay test scores than the essay test scores did with one another! What this means is that if you want to know how someone will do on some future essay exam, you can predict that future essay score better from the multiple-choice exam than you could from an essay exam of the same length.

The reason for this remarkable result illuminates the value of multiple-choice exams. Any test score contains some error. This error has several components. One component is merely random fluctuation; if the examinee took the same test on another day he might not respond in exactly the same way. A second component is associated with the context of the exam; if one is asked to write an essay on the American Civil War, it might not be as good as one on World War II, because the examinee might just know more about the latter topic. But if the test is supposed to measure writing ability, and not subject matter

knowledge, this variation is non–construct-related variance, and hence in this context is thought of as error. A third component of error is scoring variance; essays are typically scored by human judges, and even the best-trained judges do not always agree; nor would they provide the same score if they were to rescore the exam (Bradlow & Wainer, 1998). Hence, an essay's score might vary depending on who did the scoring. A fourth component is construct bias; the test might not be measuring the intended thing. So, for example, suppose we are interested in measuring a teenager's height, but all we have is a bathroom scale. We could use the child's weight to predict her height (taller children are generally heavier), but the prediction would probably contain a gender bias because teenage girls typically weigh less than boys of the same height. Hence boys would probably be underpredicted, and girls overpredicted.

A multiple-choice test of writing may contain a bias that is not in an essay test, because it is measuring something different from writing, but the size of two of the other three components of error variation is almost surely smaller. There is no judge-to-judge scoring variation on a multiple-choice exam. And, because a 30-minute multiple-choice exam can contain 30 or so items, a broad range of topics can be spanned, making it unlikely that large idiosyncratic topic effects would exist. Whether there is a difference in day-to-day variation in student performance on any specific essay versus a multiple-choice test is a topic for experiment, but is not likely to be substantial.

So now the answer is in sight. The multiple-choice writing test predicts future essay scores better than does an essay score of the same length because the construct bias is small relative to the other sources of error. Of course, the other errors on essay tests can be reduced by increasing the number of essays that are written; by making each one longer; and by having many judges score each one. But in most circumstances, these options make the test too cumbersome and would make the test too expensive for the examinee if such a policy were implemented. And, even if so, a longer multiple-choice exam might still be better.

Being able to measure the wrong thing accurately as opposed to the right thing badly has been the saving grace of modern testing so long as the wrong thing is not too far wrong. It has also allowed psychometricians to work apparent magic and connect tests on different topics so that they can be placed on comparable scales. This is possible because of the generally high correlation among all categories of ability. Students who are good at math are generally good at science; those who have well-developed verbal skills are also good at history, sociology, and geography. A large number of 20th-century tests have relied heavily on the high covariation among different tests to allow scores on very different tests to have approximately the same meaning. For example, the

achievement tests given to prospective college students (SAT II) are in many subjects (e.g., math, chemistry, physics, history, French, Spanish, German, and Hebrew). The scores on each of these tests are scaled from a low of 200 to a high of 800. Users treat the scores as if a 600 on Hebrew means the same as a 600 on French or physics. But how could it be? What does it mean to say that you are as good in French as I am in physics?

One possible way to allow such a use would be to scale each test separately, so that they each had the same mean and standard deviation. But this would be sensible only if we believed that the distribution of proficiency of the students who chose each examination were approximately the same. This is not credible; the average person taking the French exam would starve to death on the Boulevard Raspail, whereas the average person taking the Hebrew exam could be dropped into the middle of Tel Aviv with no ill effects. Thus, if this hypothesis of equal proficiencies were not true, then all we could say is that you are as able relative to the population that took "your" test as I am to the population that took mine. This limits inferences from test scores to only among individuals who took the same form of the test, which is not powerful enough for current demands.

Instead what is done is to lean heavily on the covariation of ability across subjects by (for language tests) using the verbal portion of SAT I as an equating link between the French and Hebrew exams. In a parallel fashion, the mathematical portion of SAT I is used as the common link between the various science exams. The same idea is useful in assessing the prospective value of an advanced placement course for an individual student on the basis of her performance on other tests (Wainer & Lichten, 2000).

This high level of covariation between individual performance on tests, of often substantially different subject matter, led early researchers to suggest that performance was governed by just a single, underlying general ability that came to be called "g." Almost a century of research has shown that although this conjecture was incorrect, it was only barely so (Ree & Earles, 1991, 1992, 1993). The unified structure of proficiency has helped to give tests of widely different characters enough predictive validity to be of practical use.

After the value of mass-administered standardized testing was confirmed with its initial military uses in the first quarter of the 20th century, its use ballooned and now invades all aspects of modern life, from school admissions, to educational diagnosis, to licensure. It has even played an integral role in a U.S. Supreme Court decision of whether or not a convicted murderer was smart enough to be executed (*Atkins v Virginia,* 2002). As testing has become more integral to education and licensing, ways to increase its efficiency have been sought. As we discussed earlier, the movement in the first half of the 20th

century toward a multiple-choice format has been one successful improvement. The movement toward computerized test administration at around the beginning of the 21st century has been another.

Computerizing test administration has been done for several purposes. Two reasons that loom especially large are (i) for testing constructs difficult or impossible to test otherwise (e.g., testing proficiency in the use of CAD-CAM equipment) and (ii) for providing a test customized for each examinee. It is the second purpose that is relevant to our discussion at this time. Providing customized exams is the purpose of what has become known as computerized adaptive testing, or CAT (Wainer et al., 2000). In CAT, the test is broken up into smaller units, traditionally items, and testing begins with an item of middling difficulty being presented to an examinee. If the examinee gets the item correct, a more difficult one is presented; if the examinee gets it wrong, an easier one. In this way, the examinee is faced with fewer items that are of little value in estimating his or her ability and the estimation of ability proceeds more quickly to an accurate estimate. The increased efficiency of such a procedure is of obvious value for usages like instructional diagnosis, when an instructor would like to know, with reasonable accuracy, what the student doesn't know. And the instructor would like to know this with as short a test as possible. To build a test with enough items for accurate diagnosis in all areas would make it unwieldy in the extreme. But by making it adaptive, a student could demonstrate mastery quickly and obviate all diagnostic questions meant to ferret out exactly the area of misunderstanding. Adaptive tests can also provide greater precision of estimation within a fixed amount of testing time. Current experience suggests that CAT can yield the same precision as a fixed-length test that has about twice the number of items.

For these spectacular outcomes to occur, CAT requires an enormous bank of items for the testing algorithm to choose from as it builds the test. It also needs a measurement model (a theoretical structure) that can connect the various individual tests in a way that allows us to make comparisons among examinees who may have taken individualized test forms with few, or even no, items in common. The formal test scoring models that were originally developed are grouped under the general label "true score theory" and focused on the entire test as the unit of observation. This approach will be discussed in Chapter 2. True score theory carried testing a long way but was insufficient for situations like CAT in which test forms are constructed on the fly. In this situation, the "test" was too large a unit and instead the individual item was considered the building block of the theory, which was thus named "item response theory," or IRT. We will elaborate on IRT in Chapter 3.

IRT is a family of models that try to describe, in a stochastic way, what happens when an item meets an examinee. To do this, it makes some assumptions about the character of the response process. These assumptions range from models that have very restrictive assumptions to others for which those assumptions are relaxed considerably. But until recently, all IRT models had conditional local independence as one of their primary requirements. What this means is that an examinee's performance on any item depends only on the examinee's ability and that item's characteristics, and that knowledge of the examinee's performance on another item (or items) does not add any further information. Extensive evidence has shown that this assumption was usually accurate enough for the sort of tests being built and the inferences being made. But it did force some limitations on test construction that were unfortunate. For example, local independence did not hold for typical reading comprehension items that consisted of a reading passage and a set of four to six associated items. The U.S. military in the CAT version of their entrance test (CAT ASVAB – Armed Services Vocational Aptitude Battery) solved the problem by using only a single item with each passage. This was far from an optimal solution since it was inefficient to use up a considerable amount of examinee time in reading the passage, and then extract only one bit of information from it. Reducing the length of each passage considerably ameliorated this, but it was then discovered that the construct being measured with this new formulation was more heavily tilted toward vocabulary and away from the comprehension of descriptive prose. The Educational Testing Service on many of its tests employed an alternative strategy. It limited the number of items per passage to four to six and assumed conditional independence, hoping that it would be accurate enough; and for many of the tests it was.

But as evidence accumulated and tests were being asked to enable more delicate inferences, it became clear that the assumption of local independence was often untenable. Test developers were finding that with increasing frequency it was important to build a test out of units that were larger than a single item. Units that might be a statistical graphic or table followed by a number of questions; a passage in a foreign language presented orally and accompanied by a number of questions; a musical piece played and then several questions asked. The constructs that were being measured required such aggregations of items, which by their very nature, were not conditionally independent. It soon became clear that a measurement model was required in which the fungible unit of test construction was smaller than the whole test yet larger than a single item. Wainer and Kiely (1987) argued for such a model, proposing the *testlet* as the unit of test construction. The testlet was meant as an extremely flexible unit.