Better Test Scores with TestGardener

J. O. Ramsay
Departments of Mathematics and Statistics and of Psychology
McGill University
J. Li
Ottawa Hospital Research Institute
M. Wiberg
Department of Statistics
USBE, Umeå University

March 2, 2020

This version is temporary because it is still evolving, and must not be distributed beyond those authorized to have it.

Contents

Chapter 1

Introduction

This book is about how to produce better test scores. Much better, in fact. So much better that, if it is you who are being tested, you should demand that these better scoring methods be applied to your data.

First, you'll want to know what it is that these new scores are improving. Then you'll want to look closely at exactly what we mean by better. You'll want to know how much better, of course; and then, if you're convinced that the improvement is worth the trouble to get it, you may want an explanation that reveals how better scoring works.

Perhaps, since test scores are numbers, you have already detected the possibly disturbing odour of mathematics. Maybe really deep mathematics. Don't worry, there are other outlets for the math, and in this book we will aim to avoid mathematical notation entirely. And we assume that you only bring secondary school mathematics to the task of reading about better scores. Well, except for the occasional footnote and some material in an appendix. But do be prepared for plenty of graphs, pictures, actual test questions and whatever else might be helpful.

The concept of *probability* will inevitably play a role in describing how tests work. But you will discover a new concept, called *surprisal*, that re-expresses the intuitions that we all carry about probability. And, happily, is rather easier than probability to understand and to manipulate.

In this introductory chapter, we will first define more carefully what we mean by test, and what sorts of tests we will expect to score better. It will also be worthwhile to reflect a little on why tests are so important, not only in advanced societies that organize the lives of people like yourself, but also in human social structures as far back as history will take us, and even into the fundamental aspects of the evolution of life.

But first, let's remind ourselves about how tests are usually scored.

1.1 Economic and Medical Perspectives on Testing

We know that testing can play a huge role in the life of an individual, from paving the way into a top university to consigning an unfortunate to a life of poverty and boredom. Tests help stroke victims to walk again and select fighter pilots. But what is their role in society as a whole?

The budgets of advanced economies are dominated by three big items: Education, health and defence. The United States is no exception; the 2019 budget allocates 1.1, 1.7 and 1.0 trillion dollars to each of these, respectively. Of these, education not only comes first, but plays a large hidden role in the other two. Training a doctor costs a fortune, and every soldier must be able to shoot straight, run fast and carry large weights long distances over dreadful terrain. Underlying all education is the process of monitoring progress and ascertaining when performance is satisfactory. For this we need tests, and we need them to do more than just pass or fail. They are also vital teaching tools, excellent motivators and highlight where more effort is required.

In both medicine and the military, much of the progress is achieved by many small improvements one at a time. A small increase in the effectiveness of a vaccine can avert an epidemic, and many a battle has been won by better training of recruits rather than by the genius of generals. A gain of say, 5%, can, when aggregated over thousands or millions of cases, represent a huge benefit to society at large. Sometimes society really hits a home run; the simple and cheap addition of seat belts in cars is estimated to have reduced fatal injuries by around 50%. If we can approach test scoring improvements of this magnitude, teaching and performance assessment will be dramatically safer.

Knowledge is very much like blood. Both are complex systems, and reducing either to one or a few score values hides a lot of detail, but still may be useful. In both cases the person being assessed has a right to be best assessment tools and procedures, given the consequences of wrong or misleading scores. And in both cases the persons being assessed have a right to privacy.

Test takers at all scoring levels are precious, and those at the lower levels need to be accurately assessed so as to avoid placing them in programs that they can't handle and to ensure that academic institution funds are used efficiently. We will see that the sum scores seriously over-estimate performance or aptitude level of low-competence test takers.

1.2 Meet the Sum Score

If the test is in *multiple choice format*, each question in the test is accompanied by, typically, four or five possible answers, only one of which is correct. In this case your score is simply the number of correct choices that you make, which can range from 0 to the number of questions. We call this the *sum score*, although it is also often called the *number right score*. A typical question might read:

What is a sum score?

1. *A collection of marks on a smooth surface.*

2. *The last piece of music in a symphony.*

3. *A mis-spelling of "some score".*

4. *The result of adding up a collection of numbers.*

Slightly more generally, if your response to each question is an open-ended fill-in-the-blank variety, then it is usually the case that a test scorer will assign one of a small number of pre-assigned scores, such as 0 to 4, to your answer, and the final test score is the sum of the question scores. Test designers call this a *constructed response* test. A typical question might be, "What is city is the capital of Sweden?"

A third common question type is one where you are given a set of responses, each with a pre-assigned score that you see, and you choose a score level that reflects your inner you. For example,"How much do you love me?" with five possible responses with scores -2, -1, 0, 1, 2. Again, the test score is the sum of the response scores.[1] The essential feature is that numbers are attached to responses, and these number are added.

We see that better scoring methods can be used for a much wider range of assessment processes than academic tests, but let's keep it simple and use the term *test* to refer to anything that can be sum-scored, *test taker* to refer anyone filling in one of these assessments and *test question* for any question in the test.

Almost all tests are sum-scored, including those tests designed by testing professionals to be administered to millions of students, such as the Scholastic Aptitude Test or SAT that is used in the United States for admitting students to college. And certainly classroom tests are sum scored, except for single-response essay style tests where a scorer is free to assign a score based solely on the basis of the scoring person's judgement.

It is not hard to see why we use sum scores. Adding is an easy calculation, although this is a minor virtue given the computing capacity of modern home computers. The summing process also produces the same answer on matter what the order is in which the questions are answered. Most importantly, sum scores are also easy to understand as measures of performance or status, and therefore readily accepted. Perhaps too readily. More sophisticated scoring procedures, such as those we advocate, may be be regarded with suspicion, or as tainted or taint-able, even if the score designer wears a white coat, works in a big lab, is a doctor of "score-ology,"[2] and has a large salary. Hence, a book of this nature may face serious challenges in bringing the reader on board for an alternative scoring method.

[1] Test designers have fallen into the bad practice of labelling things by the name of the person who they believe (often incorrectly) invented the idea, and consequently call these Likert Scale questions.

[2] Better known as psychometrics.

1.3 What's Not to Love About the Sum Score?

We start with a simple complaint: Sum scores ignore the variation in the performance of the questions themselves. They do not recognize that the answers to some questions tell us more about the test taker than other questions do. Questions can underperform in many ways.

- The wording of a question or its answer choices may be confusing.
- The question itself can be somewhat off-topic, so that getting it right may not relate to what the test is supposed to measure.
- Many of the wrong answers may be so obviously wrong that only one or possibly two are ever chosen, so that the question can be answered correctly by just guessing at the one or two that are remotely plausible.
- More than one of the answers nmight be right.
- None of the answers might be right.
- The answer scored as right may be wrong.
- The question is so easy that almost no one gets it wrong.
- the question is so difficult that virtually no one gets it right.

One really obvious variation among test takers is how smart they are. The information that an easy question yields about a high performance group of test takers will be minimal, since nearly all of them will get it right, and therefore the question will not indicate their relative standings among their elite peers. The same applies to a set of weak test takers answering difficult questions well beyond their grasp of the topic being tested. This means that test takers at the extremes of performance will typically have far fewer questions providing useful information than those with mid-level performance.

We in statistics call a special connection between test takers and test questions an *interaction* between them. Taking account of the interaction between test taker and question difficulty is fundamental to making best use of the test data. Sum scoring pays no attention whatever such interactions.

As an example of an interaction, consider sum scores pay no attention to the possibility that a question will be useful for certain test takers, but not others. Consider, for example, that a really basic question like, "Are newborn babies toilet trained" is relatively useless for a mother expecting her fourth child, or that a much harder question like, "How likely is it that my baby will born with jaundice?" will probably not be helpful for a Dad expecting his first. Or the question may involve material that

would be more difficult for one gender, for one or more ethnic or language groups, or some other subgroup of test takers.[3]

A test is a battle between two opposing communities, test takers and test designers, and they define define victory in opposing ways. Test takers aim to defeat all the questions and get a perfect score. But test designers construct questions in order to fail test takers, and are therefore totally defeated by the smartest examinees. Test questions are like wolves trailing a herd of caribou; some single out the newborn calves, the lame, the weak and the isolated, but there are also the alpha questions designed to slay the best. Using a scoring method that takes no account of question performance is like marching into battle without bothering to assess the enemy's defensive and offensive resources.[4]

1.4 What Makes a Test Score Better?

This is not such a simple question to answer, and in Chapter 9 we will return to the issue. At this point, however, we can reduce the question to two issues. The first is, "By how much will a test score vary for a specific test taker if a sequence of many different but somehow equivalent tests are taken?" The second is, "By how much will a test score vary if a specific test is administered to many different test takers who are somehow equivalent?" These are easy and perhaps obvious questions to ask, but the devil is in the detail of what we mean by "equivalent." And also, what precisely do we mean by the term "vary?" Later we'll teach you just enough statistics to understand what we mean by "vary."

But "equivalent" is impossible to realize in any real world situation. Instead, we will have to appeal to your intuition. We can also illustrate the idea by artificially generating data that closely resembles live test data on a computer using a mathematical model. We can then run such a program a few thousand times, and then produce graphs to display the resulting variation. We call this process *data simulation*, and this process can tell us a a great deal about the performance of a scoring method.

We will also worry about whether scores deviate *systematically* from what they should be. For example, the test that will be our prime example, the Swedish Scholastic Aptitude or SweSAT test, is a test of both quantitative and verbal skills given to over 50,000 Swedish secondary school students. Each subtest of the SweSAT has 80 questions. Only two students got perfect scores on the quantitative subtest. This somehow seems unreasonable. Surely each question in the test is not so insanely difficult that a bright high school student cannot be reasonably confident of getting its answer right! Something doesn't add up here. We will indeed discover that there is

[3]Psychometricians refer to this as *differential item functioning or DIF for short.*

[4]It's interesting that psychometricians use the Greek symbol θ to stand for performance level, which is also is the symbol for death or, in Greek, *thanatos*. Test takers aim to slay test questions and θ measures their performance.

something in the test that prevents the thousands of the most elite test takers from getting perfect scores.

We call this obviously systematic tendency to under-estimate the scores of the top students, or over-estimate the scores of weak students, *bias.* We will propose better scoring methods for reducing bias.

These two ideas of variation and bias are folded together into what we call the *accuracy* of a test score. The scoring methods that we promote will produce more accurate scores both in the sense of reducing their test-to-test and taker-to-taker variability, and their systematic bias.

1.5 How much Better Will These Test Scores Be?

An accurate score brings many benefits, some direct, and some indirect. Certainly if you are a potential near–miss in terms of college entry, you will want the most accurate score possible, so that some relatively small downward error will not deprive you of at least a chance to obtain a college education. But if you or your mother really has over-estimated your ability to thrive at college, you will also want to know since an unsuccessful year does not come cheaply. And, on the other hand, you could a potentially brilliant researcher, you will not want something random keeping you from being selected for a postgraduate degree program.

Tests also cost a lot of money and time to produce. Cost of the SweSAT is about two million dollars and two years of time. If it were possible to substantially reduce the size of the test and still get an acceptable accuracy that is, say, roughly what it would be if the test were sum scored, a test developer would want to know.

The SweSAT quantitative test that we referred to earlier is a carefully designed test, and representative of test design for large scale testing programs. Suppose that you are right in the middle of the cohort being tested, which means that 50% will score at or below your score. Then our computer simulation of multiple test reveals that the length of the test could be reduced from 80 down to about 49 questions, with the shorter better–scored test having about the same level of accuracy as the longer sum scored test. This means about 60% of the production cost and time. If, on the other hand, you are at the 25% level, with only quarter of the tested cohort below you, as test of only 23 better scored questions will do just as well. Finally, if you are in the elite subgroup of test takers where only 5% perform better than you, the test length can be reduced to about 35 questions.

Another way to look at the benefit is how long a sum-scored test would have to be in order to have the same accuracy as the 80-question better-scored test. For the 25%, 50% and 95% true performance levels, the sum scored test would have to contain about 260, 130 and 185 questions, respectively. Chapter 9 goes into these benefits more deeply, and displays benefit estimates for all four tests.

1.6 Meet the Weighted Sum Score

Our proposal for the improving the sum score is simple: Tweak it a bit, but keep the summing over questions because adding things produces the same answer no matter what the order the in which the things are processed. The tweaking consists in weighting questions according to how effective they are at a given performance level. What this means is that, instead adding 1's and 0's in the multiple choice case, we add numbers, for a specific test taker, that are:

- strongly positive if getting the chosen answer strongly suggests that the test taker is at a higher place on the score scale
- near zero if the choice, whether right or wrong, sheds little light on performance at that score level
- negative if the answer choice indicates that the test taker should be at a level.

A key aspect of our approach is that the weights or numbers that we add vary from test to test taker as well as from question to question.

We hope that you will read on to see how we devise this scoring method, why it works, and how much it would cost to use the method.

Well, actually, we can dispatch this cost issue immediately. We have developed a computer program, an"app" IT jargon, that is available for free and can score 100,000 exams in a couple of minutes using a laptop computer. Or you can score your exam choices on a smart phone considerably faster than you can read the result. You will meet the application *TestGardener* at the end of this book.

1.7 The Minimum Number of Test Takers

One might be tempted to use a sports metaphor. Like high jumping, for example. A trainer is going to ask a new trainee to jump over a bar, and will move the bar up until the jumper succeeds and fails roughly equally often. The training begins at that point.

But there is one important way in which this metaphor doesn't work. The high jump trainer knows exactly how difficult each bar setting is. But even the best test designers are not sure how easy or hard their questions are. It's only after some data have been collected from test takers that they can really be sure. Therefore, we have to use the data resulting from an administration of a test, not only test taker performance, but question performance as well. How we do this will take many of the coming chapters.

We therefore do need an adequate number of test takers for assessing question performance, and we consider that upwards of a few hundred or so will suffice. We hope to contribute something to the scoring of classroom and and upper-year college

testing in the future, but in this book we will confine our attention to medium- and large-scale testing situations.

1.8 A Story about Weights

This short story about a gym captures much of the detail in our better test scoring methodology.

A professor walks into a fitness center and asks about a weight training program. The trainer says, "The weight room is through that door. I'm quite busy at the moment, but go in there and shoulder-press as many of the dumbbell pairs as you can, and get back to me." The professor goes in, does what he's told, and comes back, reporting that he has successfully shoulder pressed 85% of the weights.

The trainer knows that he should do a better assessment, and also knows that there are far more light dumbbells than heavy ones in the room. He looks the academic over and sees a body that is in fairly good shape, but rather elderly and decidedly skinny.

He says, "Hey, prof, let's forget the 2 and 5-pound weights, and anything over 20 pounds would for you be dangerous. Here are pairs of dumbbells that are 8, 10, 12, 15 and 20 pounds. Give me five shoulder presses starting with the 8's and working up. Take as much time as you like between reps. I'll score your performance, starting it at zero. For each successful rep in this 8-pound sequence, I'll add 2 to your score. Then I'll move you on to the 10's. For each successful press, I'll add to the score the difference 2 between the weight that you've pressed and the previous weight. If you don't succeed on a rep, I will subtract that weight difference from your score. And so on through 12, 15 and 20 pound weights. When your score falls below zero, we'll know where to start the training program."

The elements in this story that we will use in our scoring process are these:

- The trainer knows that lifting 2 and 5 pound weights and failing to lift 20 pounders will tell him nothing useful.
- The weights themselves have numbers attached to them that can be added and subtracted.
- The trainer uses the difference between the current weight and the last weight as a score for each rep.
- A weight at which a shoulder press is unsuccessful is just as important as the weights that can be pressed.
- The point at which the professor's score hits zero defines his current performance level.

Similar games are used in a wide range of performance assessments. Consider an audition for entry to a music school. The assessor selects a range of pieces in ascending order of difficulty, and may even have attached a number to their performance level. The level at which the performance begins to break up indicates whether the candidate is ready for admission.

Or think of a racing cyclist training for the Tour do France. After a look at the weather conditions and the contours of his route, he sets up his initial choice of gears as he leaves his home to be somewhere a comfortable speed. With climbs and headwinds, he cycles down, and with descents and tail winds he gears up. His computer records the changes in gear ratio (the ratio of the number of teeth on the chain-wheel to the number in gear within the rear cluster). These changes are added up along the route. At a certain point, the up-change sums equal the down-change sums and the total is at or near zero. His speed at that point is his overall performance level.

Test questions are like weights, music performances and cycle rides, but what we lack for these questions is a system of change-like numbers that we can use to compose a score. That is where this book begins.

1.9 Where are We Going?

Chapter 2 specifies exactly what we mean by "test". This also includes most of the questionnaires that we encounter in a market place, health infrastructure and government agencies. We call such questionnaires *scales*. The chapter goes on to present in detail the three tests and a scale that we use in our exposition, as well presenting the data that define our illustrations. You will want to read this short and easy chapter.

You should also understand what defines the quality of a question, and since we use graphs to present information, you will benefit from getting to know those that we present in Chapter 2. We use the battlefield metaphor to discuss how we assess the performance of test questions. There we will see that a good graph can go a long way to inform us about which questions are the most important for assessing which test takers are at which level of performance. That is, we will see that no test question can be effective for all examinees, and that a quality test must contain questions that are effective all the levels of performance.

Chapter 4 also introduces the important concept of a *score index* as opposed to a *test score*. Basically the value of score index that belongs to a test taker points to the value of a better test score that the test taker is assigned. That is, the score index is like a system of mail boxes, each of which holds a test score. Please don't pass over Chapter 4!

Chapter 5 uses the graphs developed in Chapter 4 to provide an in-depth look at questions selected from all four tests in order to tell a few interesting stories, including who so few people get a perfect score on the SweSAT quantitative subtest.

Chapter 6 begins our treatment of the methodology that we used to make better scores. If you are inclined to take granted that all this technology works out as we claim, you could skip along to Chapter 9, which explains what one means by the "performance" of a score. The chapter goes on to use graphs to present the qualities of our better scores for each of our four sets of data. Chapter 10 may only appeal to the more technically minded who want a better understanding of how the process of better scoring actually works, and may be skipped as desired. Chapter 11, however, is designed to bring you to use our computer program or application *TestGardener*, which is accessible on our web site, to either run through some sample analyses, or even to analyze your own data should be have some. If you want better test scores, you'll want to enjoy this chapter, and test drive *TestGardener*.

If you do hang in for Chapters 6 to 8, you will learn some surprising and exciting new concepts. We will work very hard in these chapters to use only the mathematics that you learned in secondary school to explain these ideas, and we'll even promise to use only the math that we imagine you haven't forgotten. The new concepts include a transformation of the concept of probability into something equivalent, but is much more directly useful for devising best scores, and easier to understand than probability itself. Another object that tells a great story is a curve that we can view in a plot has a dramatic wobble in it that suggests that the acquisition of knowledge over a wide range of topics proceeds in two quite different phases. And in Chapter 7 the weight lifting story is retold as a method for computing a better test score.

this will have to be modified to accommodate Marie's new chapter, it's change in position.

Chapter 2

Tests and Scales: Essential Features

2.1 Introduction

In this chapter define and illustrate the anatomy of what we call a test. The test definition covers a rather larger range that you might suspect, of your idea of a test is a set of questions designed to find out how much you know about a specific topic. It extends to any situation in which someone presents you with a series of questions; and you either chose one of a limited set of answers, or construct your own answer and someone else places your answer in one of a small set of categories. The implicit assumption is that all of these questions relate to the same general concept, and therefore that it makes sense to reduce your choices down to a single number that reflects your status with respect to this concept.

We have to be careful with terminology, since many of the terms that we use can and are used for different ideas than we intend. So we define what a "score" is for both each answer for each question, and for the test as a whole.

Choices, whether by the test taker or by the person assigning an answer to a category, are inevitably associated with probabilities that these choices will be made. These choice probabilities will in turn depend on what the level of performance on the test a test taker has. Low-performing somewhat confused test takers will tend to have probabilities for choosing each answer for a given question rather evenly distributed. At the other end of the performance range, high-performing test takers will inevitably have high probabilities of choosing the right answer, and therefore much small probabilities of choosing wrong answers, or so the theory claims. Exceptions can occur, however.

This is the chapter in which we introduce our four tests (or three tests and a scale, actually). We primarily rely on graphs to show selected kinds of information about a test. For example, how does the number of test takers receiving a specific test score vary over the range of possible test scores? More specifically, at what test score value

do about 50% of the test takers score at or below? Or at what value is the percentage 5%, 25%, 75% and 95%? Knowing these amounts will give us a good idea of how difficult the test is, among other things. That the 50% success level is 36 questions and that only two out of over 55,000 test takers got perfect scores tells us that the SweSAT math subtest is one tough exam.

We speak of the performance of test takers, but test questions also have performances. Since their job is to fail test takers, the probability that the answer to a question will be wrong is a simple one-number measure of performance, and the highest probabilities of failure are obviously associate with the most difficult questions. But what we really want to know is the probability of failure is for a question among test takers with a specified performance level, and this is the main topic of the next chapter.

2.2 The Structure of Questions and Answers

Our first task is to be as clear as we possible can about what we mean by a "test" in the context assessing performance, or by a "scale" in assessing some status of a person other than what could be described as a performance. It is worth keeping a distinction between these two concepts, but in most respects they are identical and can be used interchangeably. And in particular, tests and scales share completely the structure of the data that they describe and the data analysis machinery that we are going to use to process the data. Let's keep it simple, tests assess performance, and scales assess any other characteristic of a person that we wouldn't call a performance. From a data perspective, they are the same thing. Consequently, we will permit ourselves to use the simple little word "test" most of the time, and reserve "scale" for non-performance assessment.

We want to lay out as carefully as we can what the structure of a test (or scale) is from the perspective of this book. A test is a set of tasks, and associated with each task is a set of events. Let's call the tasks *questions* and the events *answers*. The vast majority of answers are choices among a small number of possibilities. But we also include *scored questions* where the question is some sort of performance, such as writing a short essay, and the answers are the assigning by one or more raters or scorers, who assign the performance to one among a small number of categories.

Occasionally the answer is a measurement, such as the time taken to complete a one hundred metre dash, and we can use the term, *race* in this case. Measured answers require a special kind of processing, and they may also consist of only a single question. The analysis of races is beyond the scope of this book.

We especially have in mind questions and answers that can be presented to a relatively large number of test takers. Let say, for example, at least a few hundred, but also often to thousands of test takers. As a consequence, each question is usually expected to require at most a few minutes to complete, although scored performances can take considerably longer to complete.

Almost always there is some sort of order among the answers to a question. The simplest example is when one answer is considered to be correct or desirable, and the other wrong of undesirable, in which case there are only two states in the ordering. We shall see, however, that his does not mean that we should treat all of the wrong answers as being equivalent. Scales of the self-report variety have answers that explicitly ordered and that indicate the scale taker's assessment of her or his own status in some sense.

2.3 Scored Answers and Test Scores

The test design team usually aims to produce a single number from all the test taker's answer choices. Given the complexity of human systems, this is always a questionable practice, since it assumes that all the questions reflect the status of the same thing, roughly referred to as "ability", "performance" or "status." Test designers do this in order to propose a single number to some third party as a basis for making decisions concerning the test takers. Colleges and universities go even further by reducing four years of intense study and learning to a single grade point average. Often, too, the same number is used by many decision makers in a wide variety of decision environments. In many of these decision environments, such as corporate human resource departments, only a limited amount of time can be given the appraisal of each candidate, so that single number summaries are considered valuable as a basis for a decision. We wish that the process were more sensitive, but that's the way it too often is.

For better or worse, we are stuck, therefore, with reducing a test to a *test score.* The test score in turn is based on a score for each chosen or allocated answer to a question.

Almost always the test score is constructed by adding up the answer scores, and we call such a test score a *sum score.*

Answer scores are typically counting numbers or integers. For a test with the *multiple choice* format, the scores are almost always one for the correct answer and 0 for all the wrong answers, and adding these amounts to using as the test score the number of right answers, often called the *number right* score. Answer scores may also be as simple as $0, 1, 2, \ldots$ up to some best score value; or, for scales they answer scores may be signed numbers spread around zero, such as $-2, -1, 0, 1, 2$.

Answer scores like these are usually assigned by test designers to answers before the test is taken, and not changed after a test has been administered to a group of test takers. But exceptions do occur, and especially if it is discovered that test takers are not responding to a particular question as expected. We will shortly see examples of this.

It is also possible, too, to use the choice data that a test taker produces in more sophisticated ways, where test scores are only used to define a starting point for a

procedure that provides an alternative test score that has better properties. And this is the whole point of this book.

We will use the phrase *test score* frequently to refer to this process of adding up the scores of the chosen answers. But where we want to emphasize that test scores as we understand them are a result of test designers assigning answer scores and adding them, we will also say *designed score.*

2.4 Probability and Test Scores

In any choice situation, such is presented by the candy bar counter that we almost always pass in a food store or a pharmacy, we can imagine a probability for each choice. Probability is itself a kind of scoring system, where the scores are between zero and one, and add up to one across the available choices. Probability is a mathematical idea that most have learned to use in daily experience with things like weather prediction. Actually estimating a probability can require a lot of data, as the typical election survey illustrates. A conceptual key to estimating a probability is to collect together a group of choosers which can be regarded as essentially the same, but who for many complicated reasons do not all choose the same thing. Or probability can also be estimated by a single chooser making a choice many times, such as in the choice a beverage for an evening cocktail from a refrigerator.

If we can estimate the choice probabilities with some success, then we can replace the test designer's score for the chosen answer by the score associated with each answer multiplied by the probability that the answer will be chosen. The probabilities associated with all the possible answers to a specific question must sum to exactly one. We call the sum of the $probability \times weight$ values the *average question score.*[1]. We can express this in the following word equation:

$$Average\ question\ score = sum\ of\ (answer\ score \times probability\ of\ choice)$$

With these probability-based answer scores, we can go on to make a test score that is slso much better, in the sense that the variation in the score across a number of administrations of the test will be substantially smaller than that of the sum score. This is:

$$Average\ test\ score = sum\ of\ average\ question\ scores$$

We now look at some actual tests and scales where we have data on substantial numbers of test or scale takers.

[1]In mathematics and statistics the term is *expected score*, which is, unlike most mathematical terminology, also self-explanatory.

2.5 The Multiple Choice Test SweSAT

We will use two large-scale testing projects in Sweden to illustrate much of what we do. Both the Swedish Scholastic Assessment Test and the Swedish National Test in Mathematics are tests developed and administered by the Swedish national government to several tens of thousands of final year secondary school students each year.

We acknowledge here a precious gift from the Swedish people. As we indicated in the last chapter, the security side of testing is ever-present in all testing programs. And especially so in countries where grievances tend to be settled in courts of law at great costs to both parties. As a result, large collections of live testing data accompanied by the test questions themselves have been almost impossible to obtain by those of us involved in the testing corner of data science. Legal problems associated with question disclosure tend to be compounded by the corporate nature of the testing industry in countries like the United States. Much of the research on test development is funded and executed within testing competitive corporate entities that have little incentive to share among themselves or with outsiders.

But Sweden has taken a remarkably progressive view of testing science and the right of those being tested to access to the data they themselves produce. Each year the questions in these tests are disclosed after an appropriate lapse except for a few copywriting restrictions, and the data themselves are made available for research purposes after reasonable procedures for protecting individual test takers from identification.

These two tests are produced at Umeå University where one of the authors is a professor. We have had the gift of close collaboration with these testing agencies, including unrestricted permission to disclose and comment on any aspects of the tests that we see as problematical. Without this collaboration, this book would be unthinkable. With profound gratitude, we aim to pass this gift on to you, our readers, in the form of ideas and techniques for improving the analysis of testing data.

The Swedish Scholastic Assessment Test, abbreviated SweSAT, is typical of tests that are administered to a very large numbers of test takers. The SweSAT was designed to aid universities in selecting the best upper high school students to admit to their programs. What is being tested is their knowledge state at their level of education, but because the universities want the top performers, the questions are challenging.

The format for each question is multiple choice, where each question is accompanied by a set of answers, only one of which is correct. Typically the test taker achieves a score of 1 if the correct item is chosen, and 0 if any of the others are chosen. The test score is the number of correct answers, and we refer to such a score as a *sum score.*. In this case, the highest possible score is equal to the number of questions, and the lowest is 0. The test can be scored by a computer and therefore the scoring cost is negligible and the error rate is near zero provided that the the answer sheet is

properly filled in by the test taker. Our data come from two administrations, one at the end of 2013 and the other in 2014. The questions used in these administrations were different, and a total of 160 questions were used for assessment purposes for each administration.

The SweSAT has two sections assessing quantitative aptitudes and verbal aptitudes, respectively. We shall refer to these subtests as SweSAT-Q and SweSAT-V, respectively. Each section contains 80 questions. For each of these subtests, the questions are in turn organized into subsections. The quantitative subsections are:

- 12 questions on data sufficiency
- 24 about diagrams, tables and maps
- 24 questions involving mathematical problem solving and
- 20 questions requiring quantitative comparisons.

The verbal subsections are:

- 20 vocabulary questions
- 20 Swedish reading comprehension problems
- 20 English reading comprehension questions and
- 20 sentence completion questions.

The two sections are administered over five testing periods, each about an hour long. One of the testing sessions is given over to questions used for test equating or trial purposes.

The data that we are working with provide for each question and each test taker:

- the number of the answer that is chosen
- an indication that the question has not been completed or
- whether a response has been made that is regarded as illegible or otherwise impossible to interpret.

We analyzed these data in two ways: the first, that we call *binary*, involved using only whether the question was correctly answered, so that wrong answers, missing and uninterpretable responses are grouped together as simply not correct. The second type of analysis used what we call the full data, where we took account of which wrong answer was chosen, and also defined a special category for missing or illegible answers.

Sweden, like many European countries, has experienced a large influx of refugees, migrants and immigrants, so that reading and comprehension handicaps are known to be a serious issue. We have chosen to remove from the data all administrations

that took place in embassies or other locations not in Sweden, but this involved only a few hundred test takers. Most of our illustrations will be drawn from the remaining 53,768 students who were in the 2013 administration.

say a bit about the National Test Data, too

2.6 Plotting Test Taker Performance on the SweSAT

How well do those who take this these two subtests perform? We have at hand an obvious measure of performance, namely the counts of the number of correct answers, or what we call the sum score. We now present two plots that display how many test takers there are at each possible sum score value. These graphs will indicate that the SweSAT-Q was rather difficult and that the SweSAT-V was rather easier.

In Figure 2.1, we have plotted for each SweSAT subtest in a sequence of steps the number of students who obtained each of the possible scores. We see that the score on the SweSAT-Q in the right panel that occurred most often was 28, which is much less than the midway score of 40 questions correct. In fact, it also turned out that the score of 28 also separated the bottom 25% from the top 75% of the scores, and we have plotted the second vertical dashed line from the left at that point. The next dashed line to the right is at the score 36 that separates the scores into the bottom and top 50%, a score that is called the *median*, and which is still below the mid-score. The two right-most dashed lines at 45 and 60 correspond to the bottom 75% and bottom 95% of the scores, so that a test takers with 60 or more correct answers was in the elite 5%. Only two test takers were able to get all the questions correct.

The SweSAT-V subtest in the left panel was easier, with the most popular score being 40, and the scores being more widely spread over the score range. Moreover, the width of the score interval containing the bottom 5% was larger and that for the top 95% smaller than the respective intervals for the SweSAT-Q. We note that the score distribution indulges low performance test takers in the sense no one in the lower group scores anywhere near zero.

The SweSAT-Q test is for sure tough. But, as we move through the book, we will discover that the sum or number right score seriously underestimates how smart the top 5% of the test takers are. For example, our scoring method will assign the top score level to 76 students instead of 2, the number of test takers given a score of 80 on the SweSAT-Q.

Some questions have four answers and others have five. If a student chose an answer randomly, we would expect that on the average they would get a score somewhere between $80/4 = 20$ and $80/5 = 16$, respectively. In China students are trained to never leave a question unanswered, and if they have no idea, they are told to choose answer C. In the SweSAT-Q test, that strategy applied to all questions would give an average score of 24, which a gain of more than four points over what we would

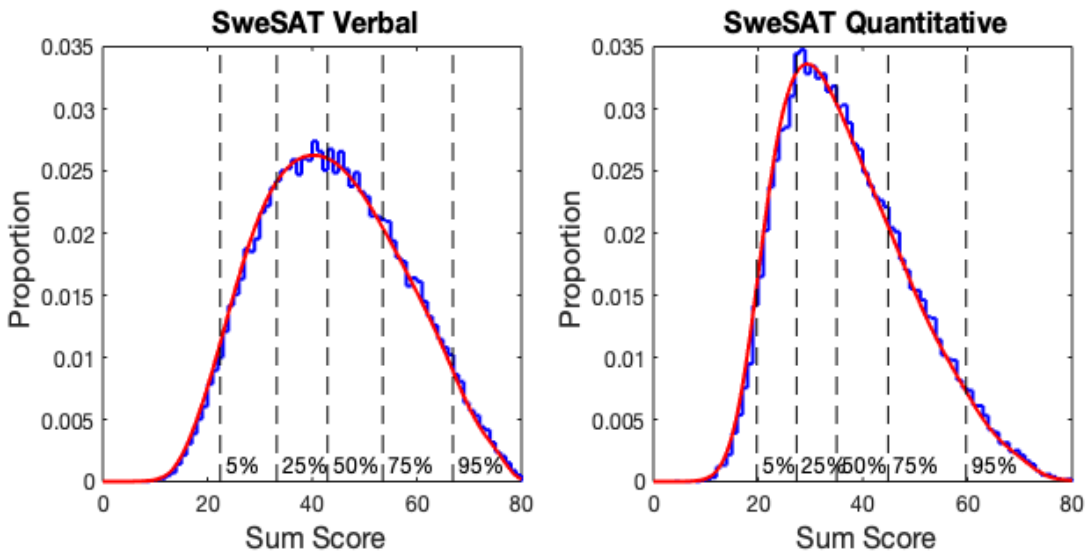


Figure 2.1: The left panel shows the distribution of possible scores for the verbal subtest of the SweSAT and the right penal shows the score distribution for the quantitative subtest. The stepped line in each panel shows the proportion of the test takers who obtain each of the 81 possible score values for the quantitative subtest of the SweSAT. The solid smooth red curve is a smooth curve representing these proportions.

expect to happen by chance.

What about the smooth red line? This is an important graphical device; replacing a rough line like the staircase bar levels by a smooth line helps us to see the overall trend. It also can be used as a kind of template, that we call a score density, and the area between the curve and the lower axis is exactly 1.[2] The quantitative curve density declines much more slowly on the right than on the left, conveying the idea that scores bunch up toward the left boundary. Not a happy day, we fear, for the majority of test takers, but just what the university admission offices were looking for.

2.7 Plotting Question Performance on the SweSAT

Questions, too, have an easy measure of performance. For a question, however, performance goes in the opposite direction, namely it is how many test takers fail to answer a question correctly. We will find a high-jumping or a weight lifting analogy useful, and so here our plots will resemble either a set set of high jump bar placements or a set of weight sizes. Figure 2.2 plots how hard a question is as the height of a horizontal bar. The top bars indicate that there are a couple of questions that fewer than 15% of the test takers can handle, and the bottom bar corresponds to a question that roughly 70% can get right.

When we turn to the verbal subtest of the SweSAT, we may be surprised to see that this easier test nevertheless has a hardest question that is even more difficult than its SweSAT-Q counterpart, and an easiest question that easier than its SweSAT-Q counterpart. This may be due to the fact that it's possible to devise wrong answers for verbal questions that are more effective at seducing test takers away from the right answer at both extremes of performance.

While neither the sum score or the better scores that we propose are affected in anyway by the order in which the test taker works on the questions, it is conceptually helpful to imagine what effect going through the questions in order of difficulty. That is, these figures invite us to think of questions as a kind of ladder where the easy questions are like the steps or rungs at the bottom of the ladder and the hardest questions are top rungs. If we view a particular test taker's success on the questions in this order, an alternative measure of performance is the difficulty of a subset of questions where there is a success/failure proportion of around 0.5, corresponding to a 50/50 odds ratio. This resembles the weight-lifting task where the person begins with weights that are expected to be hoisted without too much effort and proceeds

[2]The red curves are included in the plots in order to give us a simple visual summary of how the scores are distributed. These are called *probability density curves* in the statistical literature. The area between the horizontal axis and each curve is one, corresponding to the fact that probability one refers to all of the possible data.

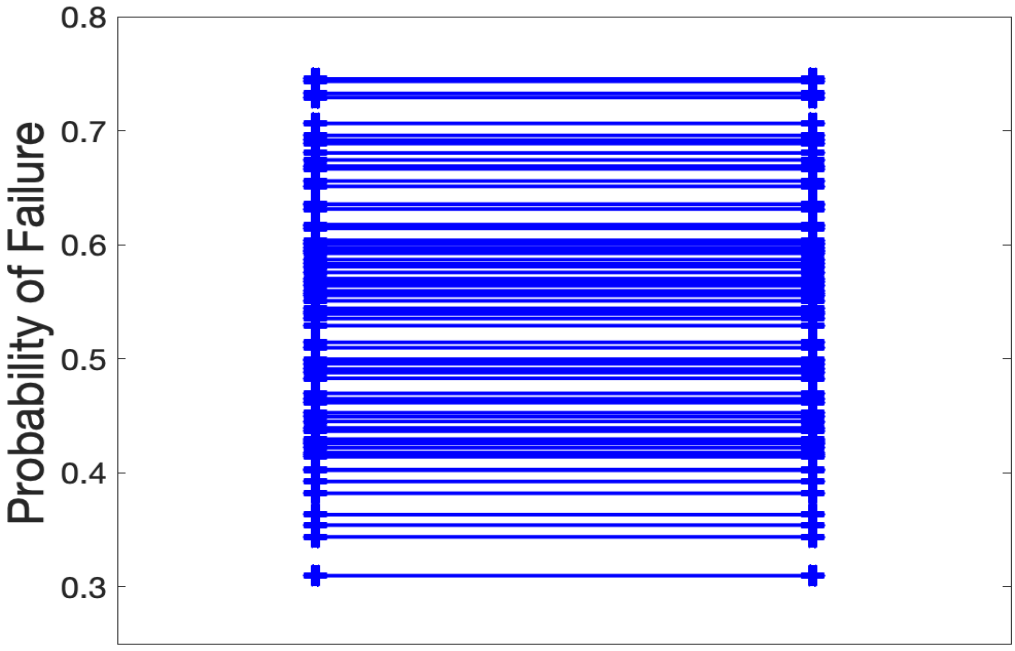


Figure 2.2: The height of each of these bars indicates the probability that a test taker will not answer a SweSAT-Q question correctly.

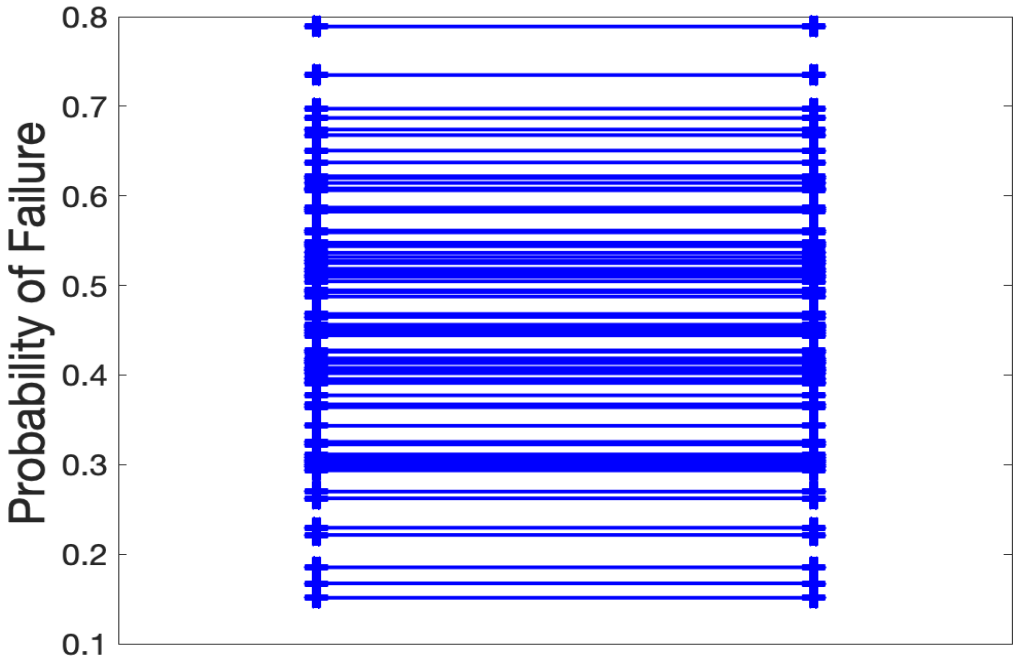


Figure 2.3: The height of each of these bars indicates the probability that a test taker will not answer a SweSAT-V question correctly.

through the weights until weights begin to be too heavy to press.

This is rather like an extended play-off series among many ice hockey teams designed to identify a small number of top performing teams that deserve to play against each other for the grand prize. It also hints at the possibility that, for a top notch math student, the easiest questions are a waste of time in terms of assessing their true ability. Or, for a struggling student near the bottom, the harder questions are not telling us anything useful. We shall pursue this idea in Chapter **??** where we devise a more accurate performance measure.

The question difficulty ladder would also work like the automatic transmission in a motor vehicle. The transmission upshifts rapidly through the lowest gears due to the low resistances to acceleration presented by the initial gear levels. It ceases to shift upwards when the resistance to acceleration remains at a certain level, and it shifts downward when, for example, a steep hill raises the resistance beyond what the usual gear is designed to handle. Just as a test taker deals easily with the initial questions, but ultimately slows down as the question difficulty begins to be more that he or she can handle.

2.8 The Constructed Response National Mathematics Test

The Swedish National Test in Mathematics is an example of the other major type of test of performance in classrooms. In this test each question requires that the test taker do some work and then write in an answer sheet the answer. Heres the first question in the test:

Write down the expression that is missing in the brackets in order for the equivalence to be true in the expression $(\quad).(x-5) = x^2 - 25$.

Open-ended questions like this usually take more time to answer than the multiple choice format, if only because the answer must be written down rather than merely checked off. The student receives a score of 1 if the answer is correct and 0 otherwise.

Many of the questions can have more than two scores, and in one question possible scores range from 0 to 4. Such questions tend to require more work or ingenuity, and test takers can be give partial credit for displaying some of the required steps along with their answer.

This test is also developed at Umeå University, and is designed to be administered and scored by teachers as an aid to assigning a final letter grade at the end in the final year of secondary school. Full instructions on what would be required for a given score are supplied to teachers. A version of the test is produced each semester, the time required to produce the test is two years, and the cost of producing each version is about $1,600,000 US.

The version of the National Math Test for which we have data has 25 questions,

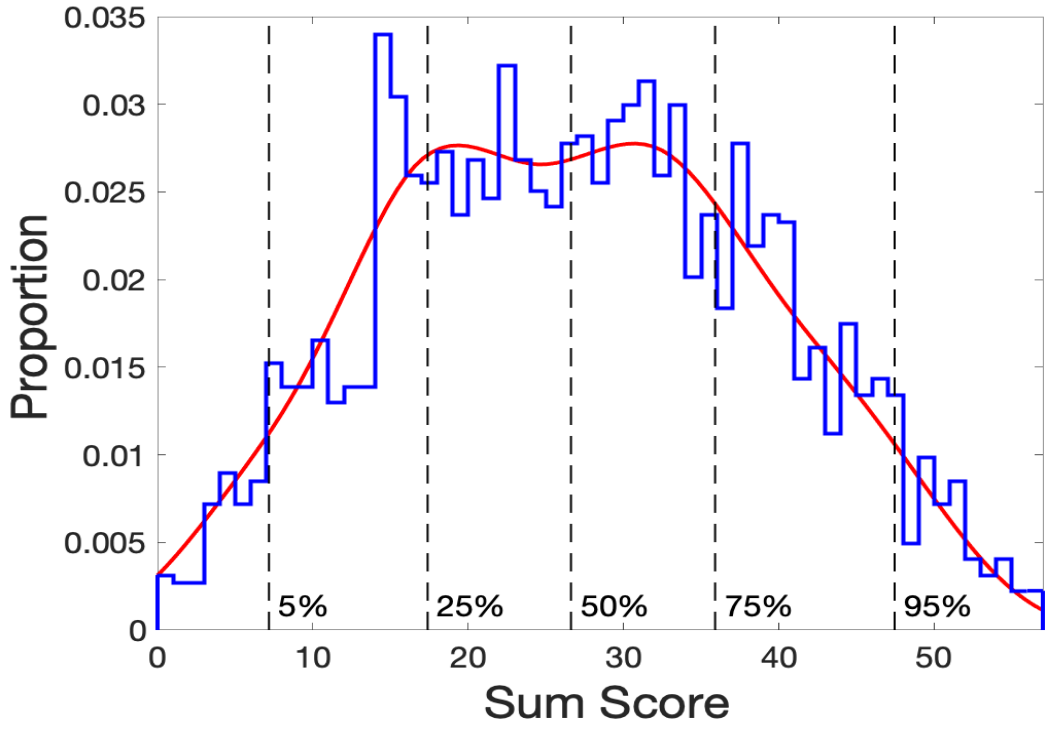


Figure 2.4: The stepped line shows the proportion of the test takers who obtain each of the 57 possible score values for the National Math Test. The solid smooth curve is a smooth fit to to these proportions.

some of which have more than one part. The total number of parts is 32, and the highest possible score is 57. We have data from 2,235 test takers.

In Figure 2.4 the proportions are far more variable because the number of scores at each score level is only about 1/20th of what we had for the two SweSAT subtests. The distribution stylized by the smooth curve is much more symmetric and is centred on the median score which is quite close to the mid-score value. We can tell that this is an easier test by the fact that more students are able to get scores of either 0 or 57 and the corresponding 5% and 95% lines are much closer to the boundaries.

Juan is confused by the last sentence

2.9 Which Test Question Format is Better?

This matter has been hotly debated ever since the first multiple choice tests began to appear at the beginning of the 20th century, with the supporters of the multiple choice format usually being on the defensive. Getting into this issue would be a major distraction for this book, but the literature on the topic makes interesting reading.

Teachers of mathematics have been especially critical of multiple choice questions, and have argued that constructing responses is an important part of the teaching process, as well as of assessing performance. The National Math Test is set up to be scored by teachers, so that the cost and time required for the scoring process is negligible. The scoring of constructed response questions inevitably involves subjective decisions, and this favours the automatically scoreable multiple choice format.

explain negligible

What does concern us, however, is whether adding up the scores is the most efficient way to measure performance. We dont think so, and we will show that for either format taking into account question performance in the scoring process can bring a remarkable improvement in score accuracy.

2.10 The Symptom Distress Scale

The term "scale" is used widely, probably because it is an easy way to say "questionnaire". The essential difference between a test and a scale is that for scales we allow the test taker to tell us about themselves, rather than forcing them to make a choice that reveals rather than directly telling us. We could easily turn a test into a scale by replacing the answer choices for questions like SweSAT-Q 55 by something like *(1) I haven't a clue, (2) I'm not sure but I'm leaning, (3) greater than 2 can't be right but it is almost so, (4) insufficient information sounds good, and (5) None of these damned answers are right!* Questions like these are called *self report* questions, and are appropriate where there is every reason to tell the truth and no incentive to lie.

The Symptom Distress Scale is widely used in nursing practice and research to assess the degree of distress felt by patients. The scale requires the rating of the intensity of the 13 types of distress using five categories. The categories are given numerical weights from 0 to 4 corresponding to the intensity or frequency of the distress. The 13 types of distress are as follows:

1. Inability to sleep
2. Fatigue
3. Bowel-related symptoms
4. Breathing-related symptoms
5. Coughing
6. Inability to concentrate
7. Intensity of nausea distress
8. Frequency of nausea distress
9. Intensity of pain
10. Frequency of pain
11. General outlook on life
12. Loss of appetite
13. Deterioration of appearance

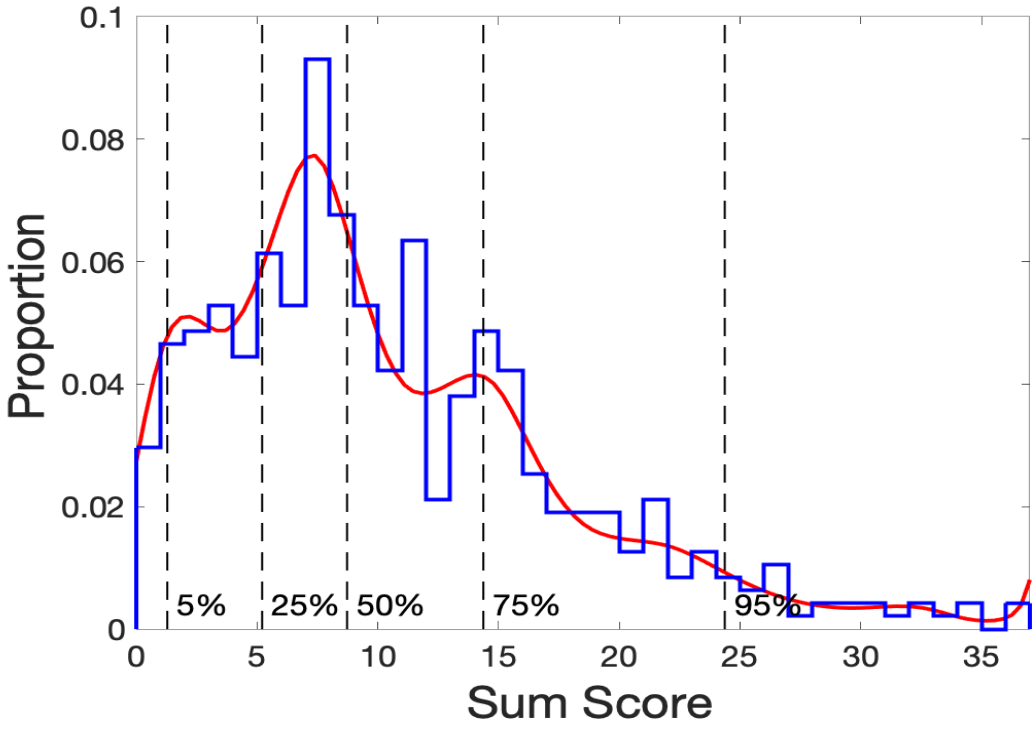


Figure 2.5: The stepped line shows the proportion of the patients who received each of the 37 possible score values for the Symptom Distress Scale. The solid smooth curve is a smooth fit to to these proportions.

Figure 2.5 reveals that the distress scores are mostly in the mild to medium range of distress. Almost three-quarters of the patients produced scores equivalent to ratings of one or zero on all types of distress. This is of course fortunate for the nursing staff, who can consequently focus almost all of their attention of those few experiencing severe distress.

Chapter 3

How Tests are Constructed and Analyzed

3.1 Introduction

When we construct a medium- or large-scale test there are many aspects which we have to take into account. A central part is the construction of the test and its questions. To develop questions is a process which will be discussed in this chapter. It involves different decisions about what knowledge to measure as well as how we should measure this knowledge. The how, refers to which questions and which question formats are suitable to use to examine what knowledge a test taker has. In this chapter, we will talk about the question development process and how pretesting of question can be used to improve the quality of the test questions. We will also discuss the design cycle for a test, the comparison of test scores and which scores the test takers are given.

3.2 Question development

The construction of test and questions for medium- and large-scale testing are typically done by persons who has this as part-time job or as a full-time job. Those who develop questions for the Swedish National Test in Mathematics are either teachers who send in suggestions of questions or former mathematics teacher who work full time with constructing the test. A reason for using former and present mathematics teachers is that these persons are well aware of the curriculum which the tests intend to measure. These people also have knowledge and experience of what students considers to be easy and difficult mathematics in the classroom. In Sweden, there is no particular education for designing knowledge tests thus those who work with designing these tests get internal education about the process, how to construct questions and analyze the questions and the tests.

For constructing 12 Swedish National Tests in Mathematics for different grades and mathematical courses there are about seven persons who work full time to construct questions and about 2.5 persons who work full time with administrative, result gathering and graphical display at a university department. For the SweSAT, there are also several former teacher who work with constructing test questions, for example former Swedish, Mathematics or English teachers. But SweSAT also have used part-time persons who sends in suggestions of questions. One of the authors of this book used to send in about 40 quantitative questions a year during a 10 year period. Of course only a smaller part of these suggested questions were used in the final version of the SweSAT but by having a large number of suggested questions to choose from, those who develops the test can aim to put together a balanced test with respect to content, difficulty and question format. At present, there are about 11 persons who work full time at a university department to construct two versions of the SweSAT per year. In addition there are a number of paid external reviewers of the questions and external persons who sends in suggestions of questions.

The question formats used in the test are chosen by those who are in charge of the test they have decided the format after studying similar purpose tests, involving experts in the area of interest (e.g. mathematics) and possibly trying different question format on the intended audience. Once the question format is decided, it is usually the same over a large number of administrations, which is the case for the Swedish National Test in Mathematics and the SweSAT. A reason that the question format is the same over different administrations is to facilitate comparability between different test administrations. When the different test question formats have been decided, guidelines are developed so all who create test questions follow the same rules. This is true in both tests that we are using in this book. An important aspect in these guidelines concern that test questions are gender and ethnic neutral. The question format gives the structure of the question and its specification. For example question 63 is one of 12 questions in the quantitative subtest with the same question format and this question reads as follows.

A bicycle dealer have five single colored bicycles for sale. There are both male and female bicycles. The colors on the bikes are black, blue, red or green, and two of the bicycles have the same color. Which color has these two bicycles?

(1) The male bicycles are in three colors.
(2) One of the female bicycles is red but the other have another color. There are no black or green female bicycle.

Sufficient information for the solution is obtained
A. in (1) but not in (2).
B. in (2) but not in (1).
C. in (1) together with (2).

D. in (1) and (2) respectively
E. not through any of the two claims

Question 63, is a typical example of a question which has a specified question format. The response alternatives are always the same with this question format, so the person who creates a new question only need to think of the context, what to ask the question about and how much and which information they should give in the claims (1) and 2). Note, the person developing the questions selects which response alternative will be correct depending on how much information is provided. Those in charge of the SweSAT however, usually strive to have a certain distribution of the correct response alternatives to make sure that not all correct alternatives are for example response alternative "C".

As for the verbal subtest in SweSAT, one part contains 20 questions of Swedish words. In these questions, the question is a single Swedish word and the five response alternatives consist of one correct synonym to the word of interest and four incorrect alternatives. Incorrect alternatives are referred to as distractors as they try to distract the test takers from the correct response alternative. When constructing these word questions the challenge is to come up with good distractors. A good distractor should be attractive to lower performing students but still appear plausible in the context. This means that those persons constructing these questions usually read many texts in the subject area of the word, for example texts from newspapers or books, to find words which are in the same context but have a different meaning than the correct answer. Those constructing the possible answers usually also try to make the correct answer move around among the possible response alternatives. For a five response alternative question, i.e. possible answers are A-E, one strategy taught at preparation courses for large-scale tests is to always choose response "C" if you are unsure which is the correct response alternative. In SweSAT, you should always guess if you do not know the correct answer as there are no deduction of scores for incorrect answer.

When question suggestions have arrived, a test developer (i.e. one of those former teacher who now work full time with designing the test) make a first check to ensure that the proposed test questions are suitable for the test of interest. Next, the question is reviewed by a review panel consisting of experts in the subject area, language experts, experts of the test in question and some of the review panelists should be expert of the population of interest. For the Swedish National Test in Mathematics the subject expert is typically a mathematics teacher who review the content to make sure that it fits the curriculum the test intend to measure and make sure that the question has a solution. The different experts review the test questions from the following aspects:

1. Appropriate,
2. Accuracy,
3. Language and grammar,

4. Question construction problem,
5. Offensiveness or appearances of bias,
6. Readability.

The first aspect "Appropriate" refers to the fact that the test question should meet the intention of the test, the content should match the intention of what knowledge the test is trying to measure and have the correct question format specified by those in charge of the test. For example, a mathematical question should mirror the knowledge area for the test (e.g. multiplication) and not any other mathematical knowledge area (e.g. division). If we take again question 63 in the SweSAT, here one makes sure that the submitted question has the two claims and built in this given item format.

The second aspect "Accuracy" means that the questions should be correct, and the question should not have any obvious mistakes. This include to make sure that the question is possible to solve. It should also not contain contradictory statements. The third aspect refers to the check to make sure that the test questions should be free from grammar mistakes and misspells and the language should be appropriate for the group the test is given to. It is important that the question does not favor someone from a specific culture or a specific group by using a jargong that is now known for those who are taken the test.

The fourth aspect "Question construction problem" refers to both mathematical notation used in the question but also that the question has only one correct answer and not several correct answers if it is a question with one correct answer and several distractors, that is a multiple-choice question. The fifth aspect refers to the fact that it is important that no group of people feel offended by the question. An offensive question could be that one preserving gender roles or discredit some group due to their origin. For example, assigning women to do the laundry and assign immigrants to low paid work in a question. The context the question is set in could also be of importance. For example if a mathematical question is about different sizes of screws - it is possible that the context favor more boys than girls. This aspect also means that it is important to examine that the question works similar for different groups, for example for boys and girls. The final aspect "Readability" refers to the important task of making sure that the question is easy to interpret and reads well and does not have an overcomplicated language for the group the test is given to.

Note, if the question has a fixed number of response alternatives as in the SweSAT, the review panel also check the questions different response alternatives for all the six aspects as well. If the questions require constructed responses as some of the question in the Swedish National Test in Mathematics, the review panel should try the possible solution and make sure that the question only has one solution. The questions which are not satisfactory according to these six aspects are either thrown away or rewritten. If the questions are rewritten, then the review panel needs to go through all the six aspects again. When the review board is satisfied with a number

of possible questions, those questions are sent for pretesting. Pretesting means that a number of potential test takers is given the question and the test developers evaluates how these persons score on the question. Details about how to pretest questions are described in the next section.

After the pretesting, it is common practice that the review panel review the test questions again from the aspects but keeping in mind how the test takers performed on the question. This is standard procedure when developing the SweSAT. Note, due to economic reasons, it is possible that in other tests that the review panel is sometimes used only before the pretesting of questions or only after the pretesting of questions.

3.3 Pretesting questions

A large part of designing a test is to write suitable questions in terms of the previous described aspects. In order to be sure that the questions are useful for its intended purpose they are often pretested before they are used in an ordinary test. Pretesting of questions can be done at a separate occasion, like for the Swedish National Test in Mathematics or the pretesting of questions can be embedded into the ordinary test as is the case for the SweSAT. The Swedish National Test usually tests their questions in a number of school classes before it is used in a regular test. A problem with this approach is that the test takers may not be as motivated to perform their best on the questions as they are aware it is just a pretest and thus the quality of the pretesting may be lower than if one pretest at the same occasion as the regular test. If possible, it is thus better to embed the questions one would like to pretest within a regular test. This approach however makes the regular test longer so the test taker may experience fatigue of testing. One challenge is to hide the pretest questions so that no test taker can guess that they are pretest questions and just skip them.

If the pretest questions are included in the regular test, the questions can be in either a separate subtest in the test form or embedded among the regular test questions. In the SweSAT the pretest items are given in a specific subtest which contain 40 presting questions. This subtest is either a verbal or a quantitative subtest which mirror these regular subtests in the SweSAT. The test takers thus get two regular subtest each having 40 verbal questions, two regular subtests each having 40 quantitative questions, and one extra subtest with pretesting questions which is either quantitative or verbal. The test takers does not know which subtests are the regular subtests and which subtest is the pretesting subtest. Note, the test takers test scores on the SweSAT is only calculated from the four regular subtests, and not the subtest with pretesting questions.

The pretest questions are then analyzed with respect to different statistical properties. This analyze include to examine how difficult the question is in the group it was tested in by examining how many test takers got the question correct. It also include to examine how how much it discriminates between low and high performing

test takers and if the question appears to be easy or difficult to guess the correct answer to. When we have sketched the questions response functions in the previous examples, those questions with steep curves are more discriminating and those with a flat response curve are less discriminating. It is important to know how well the question discriminates as many tests have as its purpose to distinguish those who has a certain level of knowledge from those who lack certain levels of knowledge. If the whole test score scale is supposed to be used for latter purpose it is important that there are questions covering the whole score scale. If the main purpose instead is to identify top performers it is important to include more difficult questions. It is also important to screen the questions performance in comparison to which group a test taker belongs to, for example with respect to gender, ethnicity or language. A good mathematics question should be answered similarly by test takers with same knowledge level regardless of their gender, ethnicity and language.[1]

If multiple choice questions are pretested, it is important to check which response alternatives the test takers are preferring. This means to not only examine which test takers managed to choose the correct response alternative but also to make sure that the distractors work well so that low achieving test takers are choosing those instead of the correct response alternative. If it is a constructed response question, then we are interest to examine if the instruction is clear so that only one correct constructed response can be given.

3.3.1 Reasons to pretest questions

There are several reasons to pretest questions before they are used in a regular test. The number one reason to pretest questions is to get information on how the question will work among test takers before the questions are used in a regular test. Thus, pretesting helps to assure quality in a test and to build test which are more similar to each other. By presting the questions we can put together a test for which behaviour we in advance know before it is administered to the test takers. It is also helpful to know the questions statistical properties in advance, in order to make tests which are similar over different test administrations.

There are several national and international associations who are working with guidelines and instructions for questions in standardized tests and they all recommend pretesting questions before they are used [2]. Prested questions which perform well in terms of their content and their statistical properties can then later be used in a regular test form. Prested questions which does not perform well is either thrown away or changed and then reviewed and pretested again.

[1] Most test agency use both classical test theory and item response theory indices to examine the questions in this step.

[2] In Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014) it is strongly recommended to pretest items. Most organizations involved in high-stakes testing use these standards to guide them when developing test questions.

3.4 Design cycle for a test

To design a test demands a number of phases which can vary slightly depending on the test. In this section we will discuss the following phases;

1. Decide the purpose of the test.
2. Prepare test specifications.
3. Construct questions.
4. Review test questions.
5. Pretest questions.
6. Review test questions.
7. Design validity and reliability studies for the final test form.
8. Develop guidelines for scoring, administration and interpretation of test scores.
9. Compare the test to previous test forms.

To decide the purpose of the test, means to decide if it should be used to compare test takers with each other[3] as in admissions tests or to compare the test takers to some well-defined criteria[4], for example a specific mathematics grade. If the test should be used for selecting high performing test takers, like in an admission tests it is more important to have questions for higher qualified test takers than for lower qualified test takers. If the test instead should be used for grading, we need test questions which cover the whole ability range in order to be able to set grades over the whole grading scale.

To prepare test specifications means that we must make a plan of what to test and how much different parts of the material of interest should be tested. As a help to structure the test questions one can use different taxonomies [5] In Table 3.1 we show one way how to structure different types of knowledge. The idea is to decide what level of knowledge the test should be about and then ask questions which belong to each of the dots in the table. This means that you in the first dot will find a question which is about factual knowledge - or simply facts and something that the test takers only need to remember, for example that 2 multiplied by 2 is 4. The questions categorized in the top left are usually the easiest questions, and the hardest questions are found in the lower right corner.

[3] norm-referenced test

[4] criterion-referenced test

[5] A well-known taxonomy which are often used when developing questions and which we discuss here is Bloom's revised taxonomy (Krathwohl, 2001).

Table 3.1: Table of how one can classify different aspects of knowledge.

Cognitive Dimensions	Different knowledge			
	Factual	Conceptual	Procedural	Metacognitive
1. Remember	.	.	.	.
2. Understand	.	.	.	.
3. Apply	.	.	.	.
4. Analyze	.	.	.	.
5. Evaluate	.	.	.	.
6. Create	.	.	.	.

Conceptual knowledge means that the test taker needs to understand concepts, principles, theories, models and classification within the subject. In mathematics, this means understanding concepts and recognizing their applications in various situations. For example how to use a square root. Procedural knowledge means that the test taker needs to solve a problem through the mathematical skills with for example help of a computer, calculator or a pencil and paper. The last type of knowledge - metacognitive knowledge - refers to how much the test taker is aware of his or her own knowledge. Typically questions here are self evaluation questions like, can you rate your knowledge in algebra. This result is then compared with the students actual knowledge of algebra. The student is said to have a high metacognitive knowledge if the estimate of the knowledge and the actual knowledge coincide. Note, metacognitive knowledge is not always tested - for example it is not part of neither the SweSAT or the Swedish National Test in Mathematics. It is however not uncommon to ask these kind of questions when students are learning mathematics in the schools within the textbooks.

The rows represent the thinking process or dimension the test taker need, which we refer to as the "cognitive" dimension. The words are self descriptive, where it is easiest to learn to just remember, like for example learning the table of multiplication, and the hardest is "create" where you create new mathematics. The idea is to be aware what parts the test covers and where one may need more questions.

These kind of tables are important if we want to make a similar test for another administration and want to make sure that the test are built in a similar way. Thus, the idea is to decide which proportion of test questions should be in each dot of the table and that these decided proportions should be stable over different test versions in order to assure comparable test versions. The test questions are then prepared, reviewed and pretested as described previously.

When we have found good test questions they are put together and given to the test taker as a regular test. After the test taker has completed the test, the results of the test is examined. Besides examining each test question and how the test taker

performed on the test one are usually also conducting studies of validity, i.e. studies to make sure that the test really measured what it was intended to measure. This is standard procedure in the SweSAT, especially when larger changes have been made. As SweSAT is a college admissions test, several researcher have studied how students perform at the university programs they were admitted to as a function of how well they did on SweSAT. Some of these researchers have also compared the performance with the test takers grades from high school as you can be admitted to university either on your SweSAT results or on your high school grades. The overall conclusion from these studies is that SweSAT is a valid test for university admissions as it appears that those who perform well on the SweSAT also perform well at the university.

Finally, when constructing a test it is important to develop specific guidelines for how the test should be scored and to decide what kind of test score or test scores should be reported back to the the test takers. As for the SweSAT, the verbal test and quantitative tests each have 80 multiple choice questions. These questions are scored 0 for incorrect and 1 for correct answer. These scores are then summed together to form a sum score between 0 and 80 for each of the test forms. However, this sum scores are not the scores the test takers use for applying to university as we will see in the next section.

Another important decision is if the test should be used for grading, as the Swedish National Test in Mathematics it is important to decide the rules and levels for the different grading levels. Those who creates this test decides the grade levels in discussion with the National Agency of Education which is the authority which is responsible for what is taught in schools in Sweden.

It is also important how the administration of the test should be done including if some questions should not be disclosed. In the Swedish National Test in Mathematics the questions are not disclosed. A reason is that former test versions are sometimes used if a test version get misused or questions are leaked on internet intentionally by test takers. As these tests are also used for grown-ups who go back to school to get a high school diploma the tests are not only given on one single day but are rather given continuously over the year. Thus several old test versions can be used if one is afraid that the current test version has been leaked to the public. SweSAT on the other hand, is administered once every half year where all regular test questions are revealed the same day that the test is given. A reason for this is that transparency is important in Sweden and it should not be a secret how an admissions test to university looks like. This means that for those test takers who take the SweSAT for the first time, the old SweSATs can be downloaded from the internet including the answers to facilitate the training for the test. Only pretesting questions and questions solely used for comparison of test scores are not disclosed to the public.

3.5 Comparing test scores

In order for test scores to be useful in medium and large testing programs, we need to have tools to compare the test scores between different test versions and different test administrations. This is particulary important with a test like SweSAT where the test result is valid for five years and thus it is possible that different test takers applying for university have taken up to 10 different test versions. It is thus essential that test scores from different test versions are made comparable.

Although we try hard to develop similar tests in terms of difficulty it is not always the case that we managed to do so. As it should not matter for the test taker which test version he or she receives, we may need to make the test scores comparable by using some statistics tools. The idea is to make different test forms comparable by placing them on the same scale. To our help we can use a number of different methods developed for the purpose of placing test versions on the same scale [6].

In order to put different test versions on the same score scale we need to use some shared features between the test versions. The shared features could either be to use the same or equivalent test takers, or to give a smaller set of questions, i.e. an anchor test, to some of the test takers who take the different test versions. Note, there are also different methods for putting the test versions on the same score scale before the test is administered or after it has been administered.

The oldest methods to compare scores from different test versions include methods based on adjusting for different test score means, or adjusting for both different test score means and different variations of test scores. We can also claim that test scores that the same percentil of test takers have got on two test versions are comparable, in which case we are using *equipercentile* comparison of test scores.

Table 3.2 shows the distribution of test takers over bins of test scores of the quantitative test. Also shown are the cumulative percentage of test takers having at least a certain sum score. For example, 45 percent of the test takers have a score of 33 or less on this test version. If we would have a similar table over another test version it may have shown that 45 percent of another group of test takers have a score of 32 or less on that test version. If this is the case, we can say that the test scores 33 and 32 are comparable by using equipercentile comparison of test scores [7].

To make scores comparable between different test versions, SweSAT gives a smaller group of test takers (about 2,000 test takers) a verbal anchor test and another equally small group of test takers a quantitative anchor test. Each of these anchor tests consist of 40 questions. The test takers scores on the anchor tests are then used in order to make the scores comparable. However we cannot ask a test taker to walk around and

[6] This statistical procedure is called test score equating and the interested reader is referred to González & Wiberg (2017) where different methods are described and how to perform them are given in details.

[7] This can be formerly described for the two test versions X and Y with distribution function F and G. Then the equating transformation can be written as $\varphi(x) = G_Y^{-1}(F_X(x))$.

Table 3.2: Sum scores, scaled scores, number of test takers (N), percentage of test takers within each raw score interval and cumulative percentage of test takers.

Sum score	Scaled score	N	Percent N	Cumulative percent
0-16	0.0	792	1.5	1.5
17-18	0.1	921	1.7	3.2
19-20	0.2	1667	3.1	6.2
21-22	0.3	2379	4.4	10.6
23-24	0.4	2923	5.4	16.0
25-26	0.5	3229	5.9	21.9
27-28	0.6	3750	6.9	28.8
29-30	0.7	3592	6.6	35.4
31-33	0.8	5264	9.7	45.1
34-36	0.9	5013	9.2	54.3
37-39	1.0	4500	8.3	62.6
40-43	1.1	5123	9.4	72.0
44-46	1.2	3413	6.3	78.3
47-49	1.3	2753	5.1	83.4
50-52	1.4	2363	4.3	87.7
53-55	1.5	1893	3.5	91.2
56-59	1.6	1925	3.5	94.8
60-62	1.7	1069	2.0	96.7
63-65	1.8	730	1.3	98.1
66-68	1.9	472	0.9	98.9
69-80	2.0	579	1.1	100.0

claim that the score of 32 which he or she got on one test version is just the same as 33 on another test version as that would make it complicated with many different test versions. Instead of using the raw scores the solution is to use scaled scores.

3.6 Scaled scores

Large-scale testing typically wants to give test takers their results within the same score range over different test administrations. In order for this to happen, it is common to give the test takers scaled scores instead of raw scores. This means that the raw scores which have been made comparable by some kind of comparison method are transferred into scaled scores. These scaled scores are the actual scores given to the test takers and which are typically used later for college applications. For example, in the past each subtest on the US SAT aimed for a mean scaled score of 500.

In Statistics, we use four different scale levels. We use a nominal scale if we are

only interested in categorizing an attribute, as for example if you are asked to fill out your gender on a survey. We use a ordinal scale if the response alternatives have a rank order, as for example if we are asking how much pain you feel on a scale from 1 to 5. We know that a pain of 3 is higher than a 2, but it is not necessary that the distance between a 2 and a 3 is the same as the difference between 4 and 5. Next, we use an interval scale if the distance is the same between the different scale steps, as in temperature. However, we cannot say that 10 degrees is twice the temperature of 5 degrees. Finally we have the ratio scale which has an absolute zero. A raw score on a test is usually considered to be on a ratio scale. If Sara gets a sum score of 10 on a test, and Lisa gets a score of 20, Lisa is said to have twice the test score of Sara. Note however, when scores are transferred to scaled score they might also be transferred to another scale level. Thus a raw score on a ratio scale might be transferred to a scaled score on an interval scale or an ordinal scale.

For the SweSAT, each of the verbal and the quantitative raw sum score scale of 0 to 80 is transferred into the scaled score from 0.0 to 2.0 with increment 0.10. The scale transformation varies over administrations with the aim of having the same proportion of test takers at each scaled score step. For an example of how this transformation looks like for SweSAT, refer to Table 3.2, where the scaled scores are given in the second column for the test version we used in this book. Note, this table will looks slightly different for different test administrations as the idea is to keep similar amount of test takers within each scaled score.

When a test taker have gotten its raw scores from the two subtests transferred into two scaled scores, the scaled scores are added together and averaged. Each SweSAT test taker thus get one scaled score which contains its verbal and quantitative score. Thus, the final scaled score scale range from 0 to 2.0 with increment 0.05. This scaled score is the actual score which the test taker use when applying for a university program.

Chapter 4

Graphing Question Quality

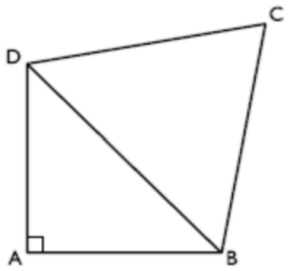
4.1 Introduction

Scientists who work with numbers discovered long ago that a plot of data could get a message across a lot faster than a table containing the data themselves. In fact, it quickly emerged in the 18th and 19th centuries that data plotting could become an art form, and now there is a large literature on the topic. In this section we design our main graphical tool, which we call the *question profile.* We will use this plot and variations of it throughout the book in order to understand aspects of how a test question performs the task of locating a test taker on the score scale.

We choose as our main example in this chapter a somewhat challenging Euclidean geometry question in the SweSAT quantitative subtest. It is displayed in Figure 4.1. A correct choice requires a knowing of the meaning of some technical words: "equilateral," "circumference" and "quadrilateral", familiarity of the Theorem of Pythagorus, the concept of a square root, and the ability put these facts together to solve an actual problem.

We begin with the straight-forward plot in Figure 4.2. For each of the 81 possible sum score values, the proportion of test takers choosing the correct answer is plotted on the vertical axis, and the value of the sum score on the horizontal axis. This is an example of what is called at *bar graph.* We see that none of the over 55,000 test takers even has a sum score less than 11. It appears that all of the test takers with sum scores greater than 72 answered the question correctly. The proportions for test takers at 20 or below flop around quite a bit, but after that the proportions increase steadily, as we would expect given the large number of test takers. At the central sum score value, 40, the proportion appears to be about 1/2 or 0.5. But is the central score value actually interesting? f we refer back to Figure 2.1, we note that the central sum score itself, below which half of the test takers are found, is only 35. At that level only 40% choose the correct answer. We congratulate all those getting it right with certainty, and especially after a return to Figure 2.1 reveals that these folks were the top 0.1% of the class.

AB=AD=1 cm. The triangle BCD is equilateral. **What is the circumference of the quadrilateral ABCD?**



A. 5 cm
B. 6 cm
C. $(2+2\sqrt{2})$ cm
D. $(2+3\sqrt{2})$ cm

Figure 4.1: Question 46 in SweSAT-Q, the quantitative subtest of the SweSAT. The correct answer is C.

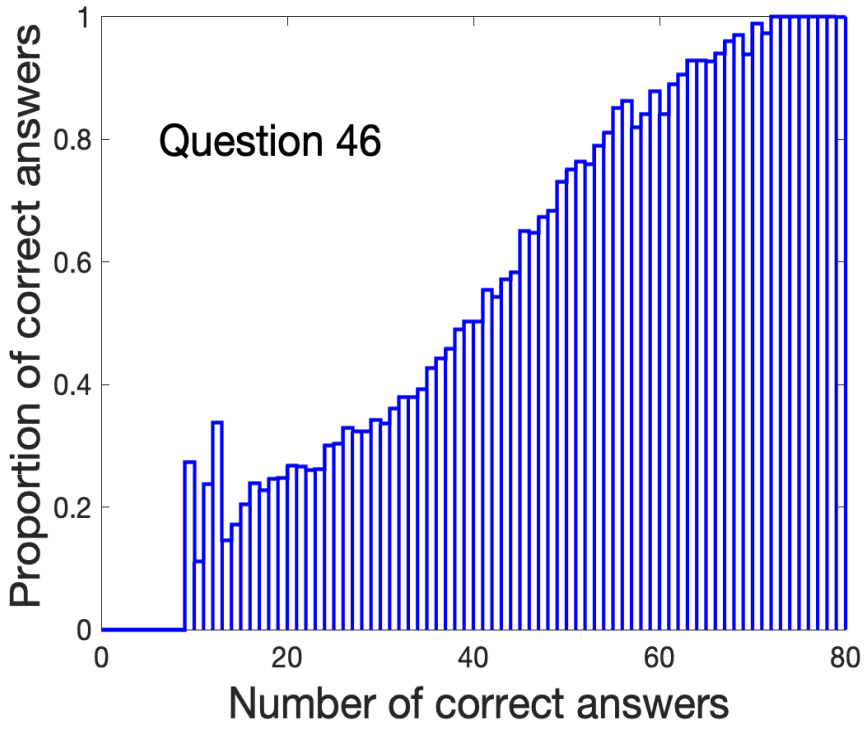


Figure 4.2: The proportions of test takers at each sum score value who chose the correct answer for question 46 on the quantitative subtest of the SweSAT-Q.

The plot is of course effective, but we are bothered slightly by the fact that we had to refer to another plot in order to understand this one. The score *values* that we use here for the horizontal scale (called the *abscissa*) don't see to be quite what we require. There are a large number of test takers getting sum scores in the centre of the score range, but much fewer in the extremely low and extremely high regions, so that a lot of data is squashed together in the center and a relative few observations are spread out at the extremes. These features have made it difficult to understand how difficult the test is. There is also redundant information. Of course the top scorers get a question right; this is what being a top scorer means. Somehow we aren't really learning anything for scores higher than 72.

We introduce at this point the simple strategy of allocated the data to *bins* such that each bin contains approximately the same number of test takers. Then we compute the *proportion* of test takers getting the question right, and we add these proportions to the plot, with result shown in Figure 4.3. We still use the sum score values for the horizontal direction, but now we show position along this scale in terms of bin boundaries. Now we can see immediately that the first and last bin cover a lot of score values, 18 in the lower bin plus 14 in the upper bin equals 32 or 40% of the horizontal space allocated to about 3.5% of the data. This does not seem like an efficient use of space!

But data-binning does reveal interesting things that were hidden in the first plot. Most noticeably, we now discover that the proportions increase in a nice smooth way that just begs to be reduced to a curved line. Replacing jumpy data points by smooth lines is something that statistical graphics specialists have perfected, and Figure 4.4 leaves the point in the plot but adds a smooth line, and not only for the correct answer (the blue curve) but also for the incorrect answers. We love this curvy line because the points cluster tightly around it, and to such an extent that in future plots could drop the points and just look at the line.

Moreover, adding the wrong answer information enriches the plot considerably. We see most of the test takers choose more or less equally amount the three wrong answers, and are therefore unable to distinguish among them. Indeed, the bottom 25% are choosing equally among all the answers, and therefore most likely just guessing. We do note that, among the top 5%, only one wrong answer, number 4, is chosen. Perhaps it's because they know that $\sqrt{2}$ must have something to do the answer because they can use the Theorem of Pythagorus. Perhaps we should call those in the top interval "Pythagoreans."

We dropped the boundaries of the bin because the points pretty much tell us what they did. We also added something that eliminates going back to Figure 2.1, namely some dashed vertical lines showing us what sum score values correspond to five meaningful percentages of test takers at or below a marker line. Now the difficulty of the question pops out at us right away since we can immediately see that the central test taker only has a 40% chance to getting this one, only the top 5% get it right with real assurance. Both Figure 2.1 and this figure remind us that a lot of the SweSAT

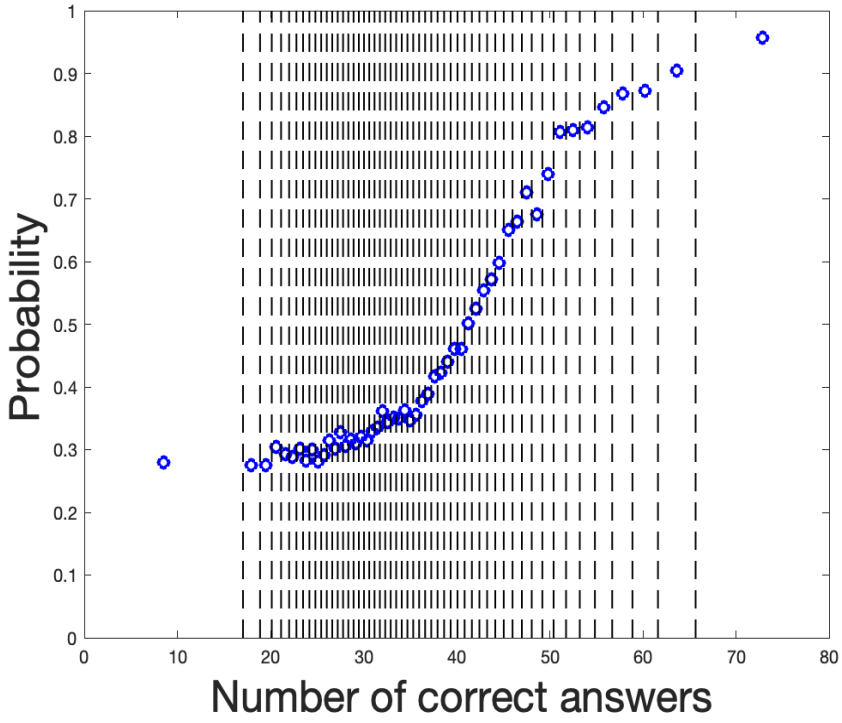


Figure 4.3: The circles are the proportions of test takers in each of 55 bins for question 46 of the SweSAT-Q. The circles are located at the mid-points of the bins, and each bin contains about 1000 test takers.

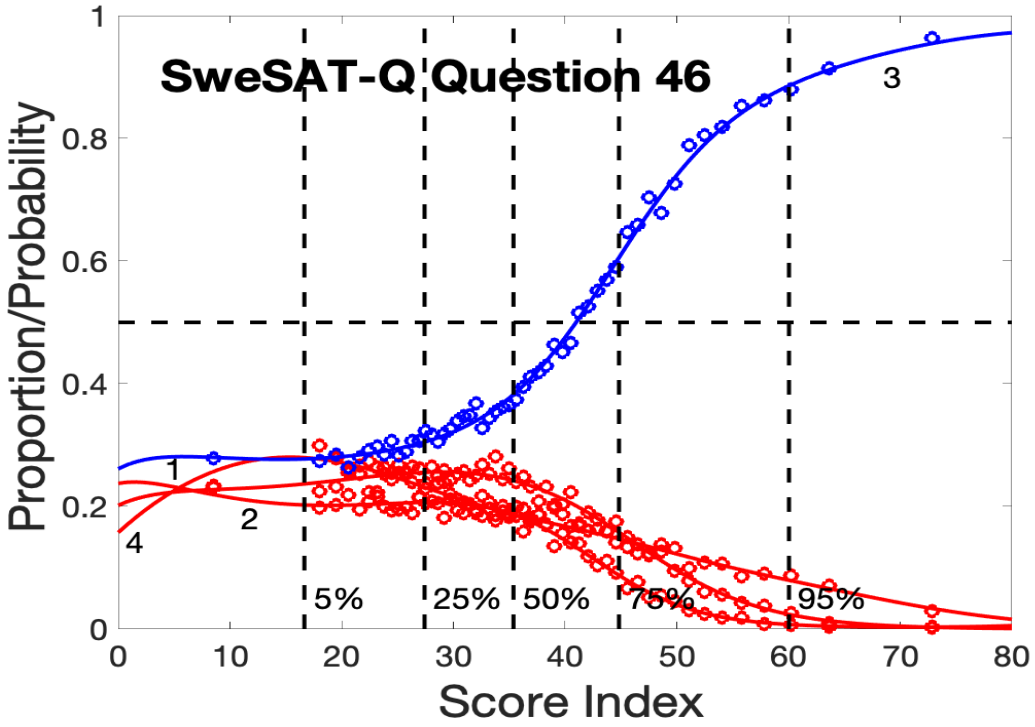


Figure 4.4: The circles are the proportions of test takers who chose the correct answer for question 46 in the SweSAT-Q. The circles are located at the mid-points of 55 bins, each containing about 1000 test takers. The smooth curve expresses how the proportions tend to change over score values. The proportions and probabilities for the correct answer are in blue, and for the wrong answers in red.

test takers are applying to university programs where a facility in math is perhaps not that important. You, dear reader, may be one of them. We'll try to keep this in mind.

We leave this section wondering why we use score values as the horizontal axis. This, we now see, assigns too much visual importance to the bottom and top 5%, and might be hiding some interesting information by squashing the overwhelming majority of test takers into the central 60% of the graph.

But we did change the sum score, that increases in a staircase fashion, to something that increases smoothly by fitting a curved line to the data. That is, we changed *sum score proportions* to *sum score probabilities.*. This has a profound implications that we will use to advantage. The probability of choosing an answer according to the smooth curve now also increases smoothly, and captures that belief that somewhere and sometime, for any two test takers, no matter how close their probabilities of choosing answer are, there is a test taker that has a probability value between theirs. In other words, as the probability of choice changes, there are no gaps.[1]

4.2 Introducing the Score Index

Notice that we changed the label for the horizontal axis in Figure 4.4 to "Score Index." We did this because we won't always use the number of correct answers to indicate test performance. For example, if we were comparing two tests with different numbers of questions, we might find the percent scale, running from 0 to 100, rather more useful. The percent scale would also be better if one test were easier than the another, or measured something entirely different, as do the SweSAT-Q and SweSAT-V subtests.

In short, we have invented a new thing when we went from the observed sum score as the horizontal variable to a smoothly increasing line. We are going to call this new thing a *score index.* We use the term *index* because for any index value, we can compute a test score by using the probabilities that the index value selects. More on this in Chapters 6 to 8. The term "index" also rightly suggests that there are other possible indexing systems that will do the same job. Finally, the score index can be used to impose a unique rank ordering of a set of test takers. There are many thousands with one of the score values, such as 30, in Figure 4.3, but we shall replace sum score values by smoothly changing score index values such that the chances of ties in score index values are remote.

We have, however, retained one feature of sum scores. There is always the potential for pileups of scores at lowest and the largest index values; whether a lowest index value of 0 or anything else, or a largest possible index. It is not unusual to see multiple zero scores for, say, constructed answer tests, corresponding to test takers whose level

[1] Way back in Chapter 1 we mentioned using the Greek letter θ to stand for a level of performance. We now insist that θ shall stand for the value of a score index. Moreover, we refer to a proportion or a probability associated with a score index value as $P(\theta)$.

of knowledge is below anything required to define an answer. Nor is it unusual to see multiple largest score values for a relatively easy test, or when a very large number of test takers is involved. We don't see this happening for the SweSAT quantitative subtest, but it can be a strong feature in other situations. Test takers with extreme score are, in a sense, outside of the range of performance that the test measures. Within either the lowest or the highest perfect score group, there is simply no further information available to discriminate among those who have these extreme scores.

4.3 Another Score Index: Percent Rank

Competitively minded test takers are often more interested in how many test takers had scores that were either better or worse than their score. In sports, the concept of *rank* is used to sort athletes in this way, with rank one being the holy grail that all athletes dream about. In this book we want to use large numbers for good things and small numbers for not so good outcomes. So, let's define *score rank* in our context as

$$rank(score) = number\ of\ test\ takers\ with\ scores\ at\ or\ below\ score$$

This definition of rank is like the ratings of movies or hotels, rather than the more traditional use of rank 1 to indicate the best or biggest. This keeps the best on the right of our plots and the worst on the left. If we use N to stand for the total number of test takers, score ranks in our sense will be spread fairly evenly between 1 and N.[2]

Two different tests will often have quite different numbers of test takers, so in order to more easily compare question performances between different tests, we can instead use the *percent rank* instead, defined as:

$$percent\ rank(score) = 100 \times rank\ score/N.$$

The advantage of percent rank as a horizontal variable is that these values tend to spread themselves evenly over the performance continuum, which now runs from 0 to 100. This contrasts with the uneven distribution of scores in Figures 4.2 to 4.4. Since the distribution of test scores will inevitably vary from one test to another, we use percent rank in order to hide this variation so as to make more visible features in our plots of question performance that we need to focus on.

Figure 4.5 displays question 46 performance using the percent rank score index in order to avoid a potential source of confusion. Now only $2 \times 5\% = 10\%$ of the test takers occupy this space, so that for most purposes we only want to give a quick glance to these end intervals.

[2]If we apply this definition to sum scores, score ranks will increase in a stepwise discrete way, and also very large numbers of test takers in the center will have the same score ranks. But we in the data analysis community have a little trick up our sleeves to deal with that. We add a tiny random number varying just a bit around 0 to each score before computing score ranks. This is called *jittering* in the data graphics community. Sure, this isn't fair to those unluckily receiving a negative jolt to their sum score, but at the level of a graph, we wouldn't notice the difference. Graphs are about seeing the big picture and hiding the fine details.

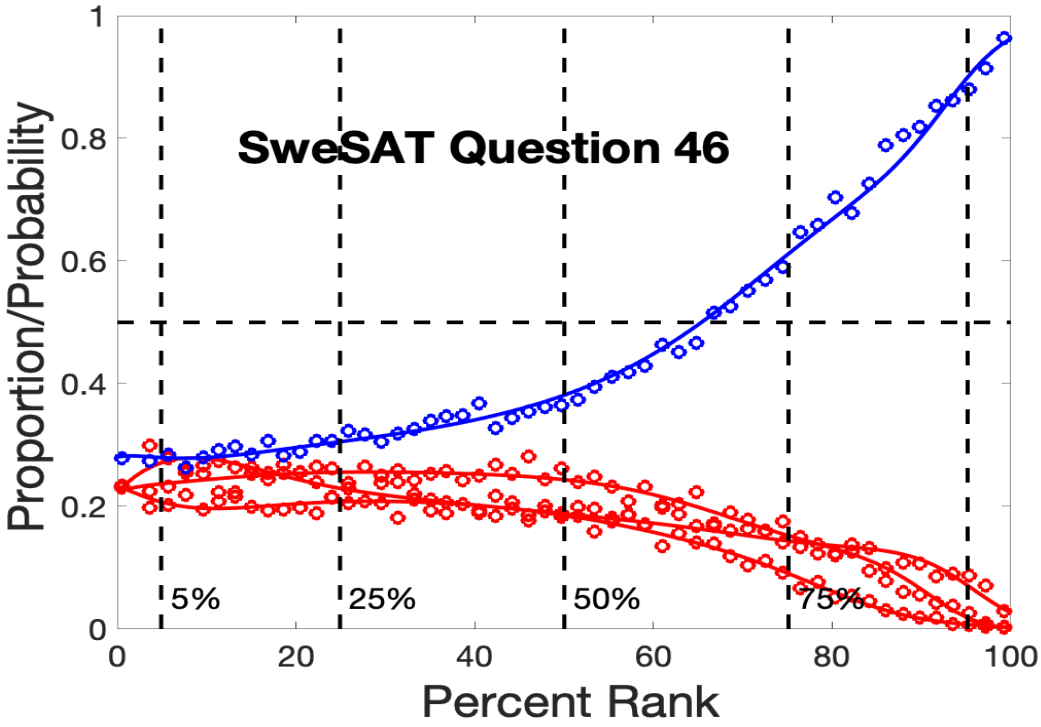


Figure 4.5: The data in Figure 4.4 displayed over the percent rank score index. The circles are the proportions of test takers who chose the correct answer for question 46 in the SweSAT-Q. The circles are located at the mid-points of 55 bins, each containing about 1000 test takers. The smooth curve is designed to represent how the proportions change over percent rank values.

The five marker percent dashed lines are now where we expect them to be, and we are not so distracted by what the question performance curve is doing over the bottom and top 5% of the test takers. When we compare this figure to the previous version, we see that the curve shape has changed. We can now see rather more detail in the curve over the central 50% of the test takers, which now exactly occupies the central 50% of the horizontal axis. In particular, it is evident that the question performance curve is sharply increasing only over the top 50% of the test takers. In fact, we now see, question 46 is actually a rather challenging question, and those getting it right at with a high probability are in fact well within the top 50% of the test takers. This is exactly what we want to see.

There is also an advantage brought by the score index percent rank that is not obvious in the plot. The 55 bins that we used to define the points that in turn defined the smooth curves are now all of equal width, as well as having roughly equal numbers of test takers. This helps the process of defining the smooth like to be more stable, and especially near the two boundaries.

What we have lost in Figure 4.5, namely information about the distribution of the sum score score index, is unimportant in the context of evaluating the performance of a question, and hiding this aspect produces a simpler and more informative plot. Good graphical displays should convey only one message at a time, and the message should leap off the page for us. We will continue to use percent rank as the score index when viewing more question performances in this and the next chapter.

4.4 What the Score Index Does

The score index is not a test score. Instead, it is a device for archiving or storing test scores. It resembles a file folder system, but one that is infinitely long. Each score index value points to a single test score value. We saw in Section 2.4 that, when we use probabilities to define average test scores, test scores can take on any number within a restricted range. This implies that adjacent score index folders point to test scores that are arbitrarily close together. Consequently, score index values are themselves numbers with values in a restricted range.

The job that the score index does is to ensure that test scores evolve smoothly as a test taker moves along in time and experience. Among an arbitrarily large group of test takers, the sorted test scores increase in the same way that decimal numbers do. By "evolve smoothly" we mean something that behaves like a line or a string: (1) positions on the line are ordered from smallest to largest, and (2) the score index is *continuous* in the sense that there are no gaps or breaks anywhere along its length. In fact, we want will see later that these curves are even smoother than merely being continuous. At no point does their rate of increase change abruptly. This will be important because, as we will see later, the rate at which a curve increases what defines a better test score.

What makes a score index different from a test score? First, the test designer

plays no role in its definition. Test designers can change the weights or scores that they assign to answers either before or after a test is administered, and the result is a different test score, but not a different score index. That is, the test designer can change that test score to which a score index points, but not the index itself.

Second, there is an infinite number of choices of score indexing systems. We have already used three: (1) the discrete integer-valued sum scores, (2) the continuous version of the sum score, and (3) the percent rank. In fact, any smooth transformation of a score index that preserves its order is also a score index. By "smooth" we mean that the transformation does not introduce any gaps or pile up scores at a single point.

In order to understand how flexible a score index is, let's propose an example. Suppose we run the score index from 0 to 1. Now, suppose it occurs to us that knowledge is more like an area than it is like a straight line. That is, we not only learn more and more things, but what we know spreads out over a wider and wider range of subjects, somewhat like the Mississippi river delta. Why not, then, square the numbers in the 0-1 interval? If we do, our progress, thought of as area, will start and finish at the same places, but pick up the pace of learning continually as we become familiar with technical terms and the main concepts. Here we have executed two transformations: We have changed the upper boundary of a sum score by dividing its value by the number of questions, and then we have squared the resulting numbers that range from 0 to 1.

Figure 4.6 shows how the two index systems would report progress on question 46 of the SweSAT-Q. Note that what the blue linear score index curve passes through probability 0.5 at score index 0.5, while the the red area score index curve does the same at score index $0.25 = 0.5^2$. The same is true across the whole index range; whatever level the blue curve is at for a given score index, the red curve takes the same value at the square of the blue score index. We see that the square index captures the idea that area increases more rapidly than length. We also see that the curve changes shape, but that the values within the curve do not.

This flexibility in the choice of the score index makes it sound rather arbitrary, but in fact this flexibility opens up an impressive list of opportunities, and we shall see some of these later in the book, just as we have seen that the percent rank score index simplifies a question profile in an important way.

Here are a couple of images that we like. A score index is like a bank vault in the sense that it protects the information in the test from test designers. The score index is also like a library. Library shelves hold books, and the books are shelved in the same order everywhere in a library system, so that the book seeker knows where a topic is. But book shelves themselves can be in quite different arrangements. In one library, they may be tall in order to fit the books into a cramped space, while in another larger space they may be spread out over a larger floor area. What doesn't change is the order of the books within the shelves. The score index is like a library shelf system; the continuous sum score values that we used first were stacked up in the

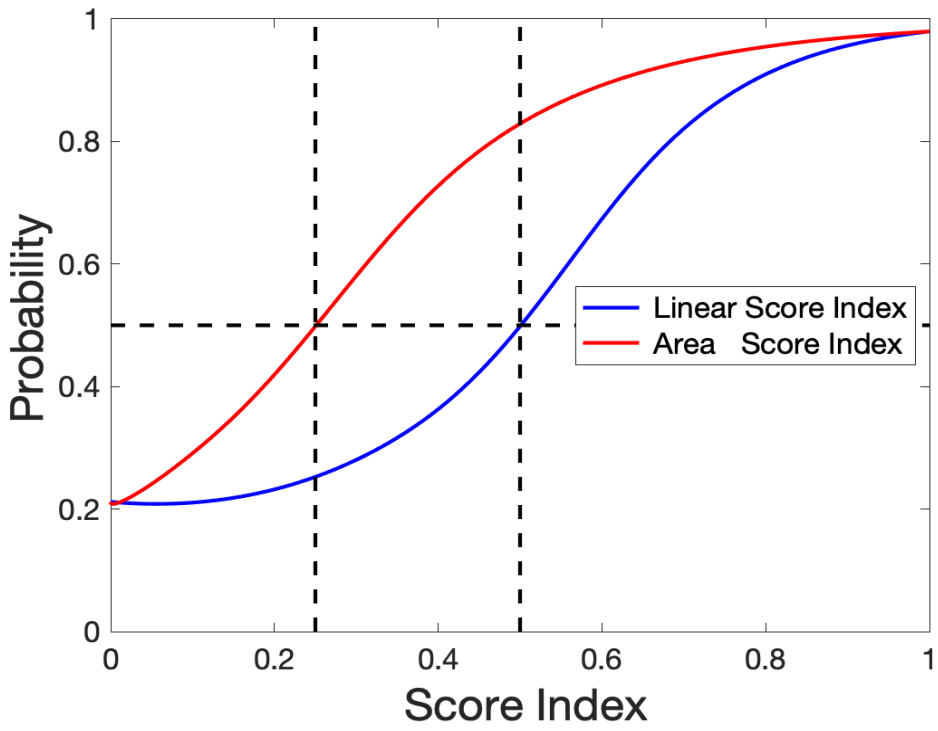


Figure 4.6: The blue curve displays the evolution of the probability of getting question 46 right in the SweSAT-Q shown in Figure 3 in Chapter 4, but using as score index the test score divided by 80. The red curve shows the progress if we treat knowledge area and square the score index.

middle like tall shelves; but the percent ranks that we used next, and that we prefer, were spread out evenly over their range. However, like the books, the probability values that the score index points to are in exactly the same order.

4.5 Some Varieties of Question Profile Shapes

Figure 4.7 displays aspects of question performance in a schematic fashion by using a smooth curve. The figure shows nine question profiles: three easy (top panel), three mid-range(center) and three hard (bottom). Within each panel you will see three types of curves assigned colours according to how rapidly they pass from probabilities near zero to probabilities near one. The steepness of a probability curve will play a major role in how we compute more accurate test scores.

Notice that the solid red curves in the plot come close to telling us which part of the performance scale the test taker inhabits. If the person passes the easy item but fails the two other questions, we can be reasonably confident that their percentage will be somewhere between 25 and 45. If all three are passed, we can strongly suspect that the persons performance is above 75%. Collectively these three red templates divide the scale into four regions, corresponding to the question scores (0,0,0), (1,0,0), (1,1,0) and (1,1,1), where 0 means fail and 1 means pass. A test composed of questions this powerful would not need to be very long before we had all the precision that we needed in a test score.

There are, however, $2^3 = 8$ possible triples of right/wrong question scores, and the four shown above will confuse things because they will violate the principle that scores get better as test taker gets smarter. The dotted green curves are apt to produce many of these ambivalent outcomes because, as they rise lazily from 0 to 1, it is quite possible that an easier item will be failed and a harder one passed. A test made up of these types of profiles will have to be substantially longer before we can safely assign a performance level to a test taker. Longer tests are expensive, and somebody is going to have to pay more for such an inefficient test. The dashed blue curves are more less what we see in practice.

At this point we have three criteria for an effective question:

1. *Probability of a correct answer increases with performance or ability.*

2. *We can place the mid-point of the curve (probability 0.5) where we like so as to control the mix of easy, moderate and hard questions.*

3. *The rise in probability should be steep at the mid-point so as to provide relatively unambiguous evidence of performance level.*

How do the SweSAT questions fare relative to the properties that see in Figure 4.7? Figure 4.8 shows all 80 question profiles for the SweSAT-Q. This is a lot to look at, to be sure, and we will switch to examining individual question performance

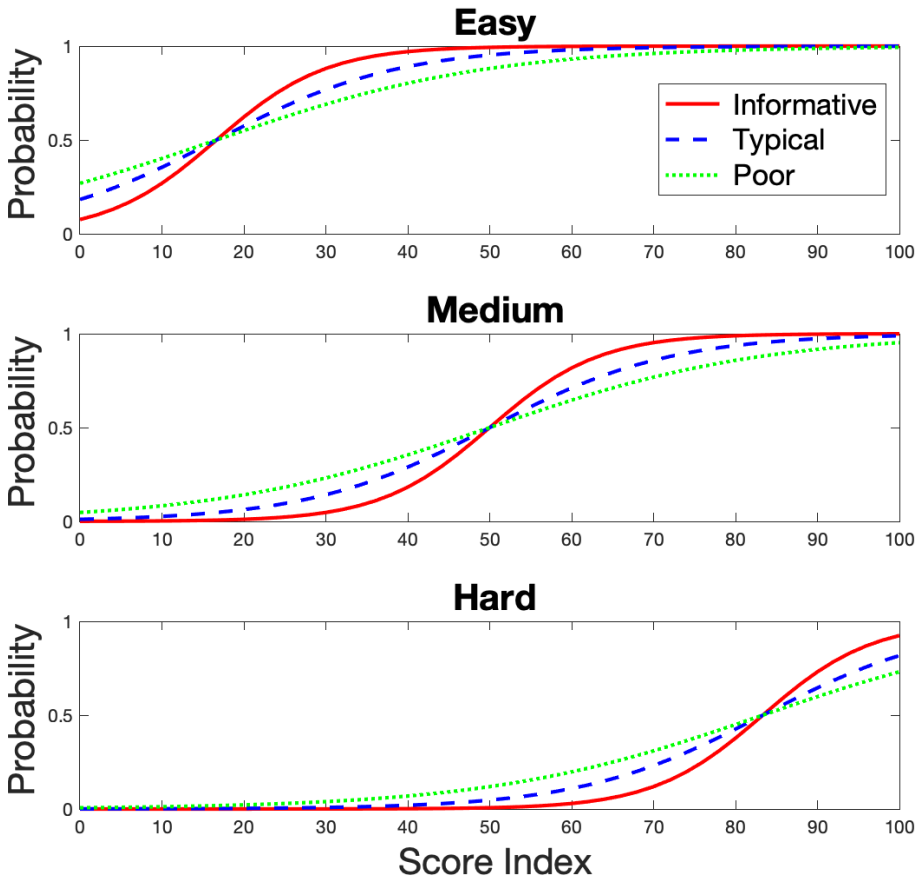


Figure 4.7: Schematic probability correct curves displaying how proportion correct varies over the total test score. The top panel shows three curves for an easy question, the middle panel shows three medium difficulty curves, and the bottom panel is for hard questions. Within each panel the curves vary in how much information they supply about test taker performance: red for highly informative, blue for moderately informative and green for relatively uninformative.

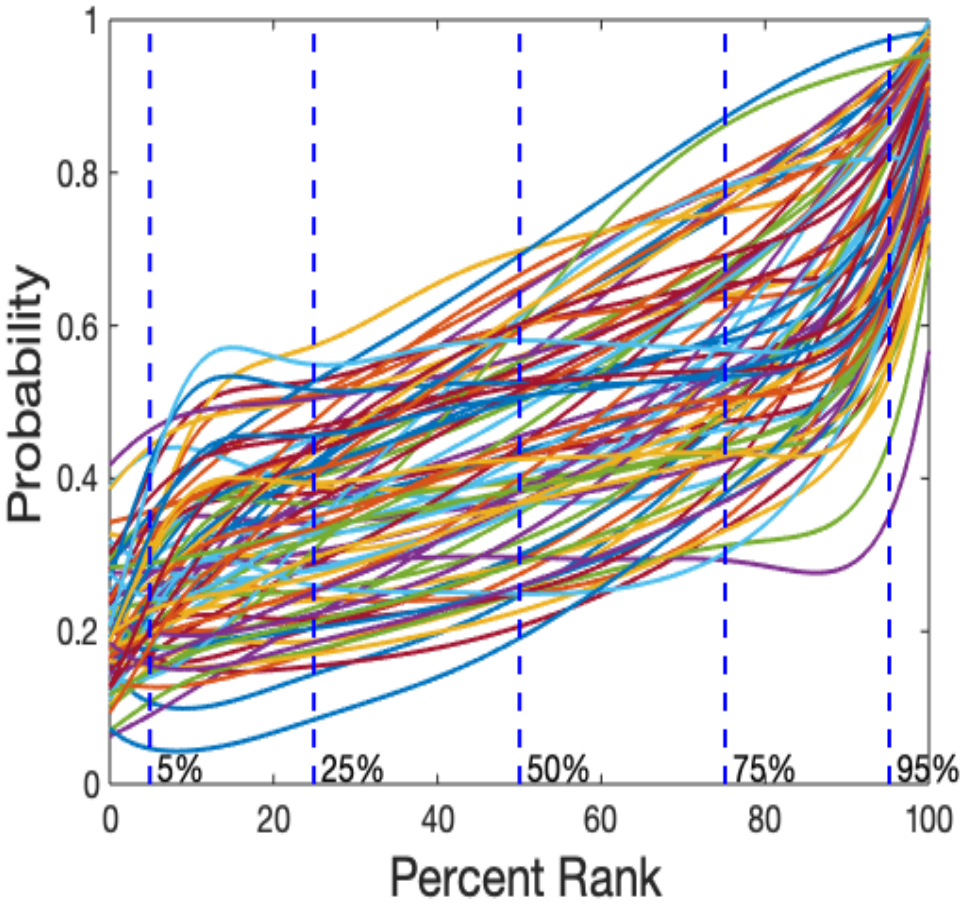


Figure 4.8: Proportion correct curves displaying how proportion correct varies over the total test score for all 80 questions on the SweSAT-Q.

shortly. But here we see at least that the majority of the probability profiles are of at least medium quality.

When we look at a lot of curves, our eyes tend to focus on the weird curves, and here perhaps especially the curves that never get close to probability one, like the purple curve that just barely beats the 50/50 probability 0.5, and is practically flat for 75% of the test takers. For example, at the 100% end of performance, we see a purple curve that only reaches probability 0.55 on the right axis. This is question 39. We will discuss why this question appears to be so hard for even the top SweSAT-Q test takers.

4.6 SweSAT-Q Question 55

Here's a question that we find especially interesting, and that illustrates some of the investigations that we can undertake once we have a good question profile setup.

Question 55 is in the quantitative comparisons section for the second half of the quantitative subtest, and is the following:
Let x be a positive real number. Let

$$f(x) = \frac{1}{x} + x.$$

The answers are:
1. $f(x) > 2$
2. $f(x) < 2$
3. $f(x) = 2$
4. *The information given is insufficient for answering the question.*

The fourth answer was scored as correct by the test designers. Figure 4.9 is the plot that we will use to assess this question.

Answer 4, scored as correct, is excellent in terms of its steepness. It is a trifle easier than would be ideally appropriate for a test taker at the 50% point in the score distribution. But test takers choosing this answer are providing strong evidence that they belong in the top half. The wrong answers 2 and 3 do a fine job of drawing the weakest test takers away from both the right answer, and as well the wrong answer 1.

But we have a problem! The probability of choosing answer 1, that $1/x + x$ is greater than 2, is around 0.2 all the way across the performance scale. Its easy to see that, if $x = 1$, then $1/x + x = 2$, so that even the arithmetically challenged ought to be able to see that there is at least one value of x for which 2 and the function have the same value. We would surely think that the brightest bin folks would have dismissed that answer out of hand, but that is not what happened. About 15% of them opted for answer number 1. Because of this, the probability template has the rather serious defect of preventing many hundreds of the brightest test takers from getting a perfect score. What went wrong?

We think that none of the answers are right. In fact, the question provides enough information to tell us that $f(x)$ is positive everywhere that $x > 0$, and also that at $x = 1$ we have $f(x) = 2$, which is its unique minimum value.[3] The test taker

[3] A simple calculation, available to most bright math students at this level, goes as follows:

1. subtract 2 from the right side of the equation to get $1/x + x - 2$, which is the difference between the expression I and 2,
2. multiply this quantity by x, which you can do because it is positive, so that we have $1+x^2-2x$,
3. notice that this is $(x-1)^2$,

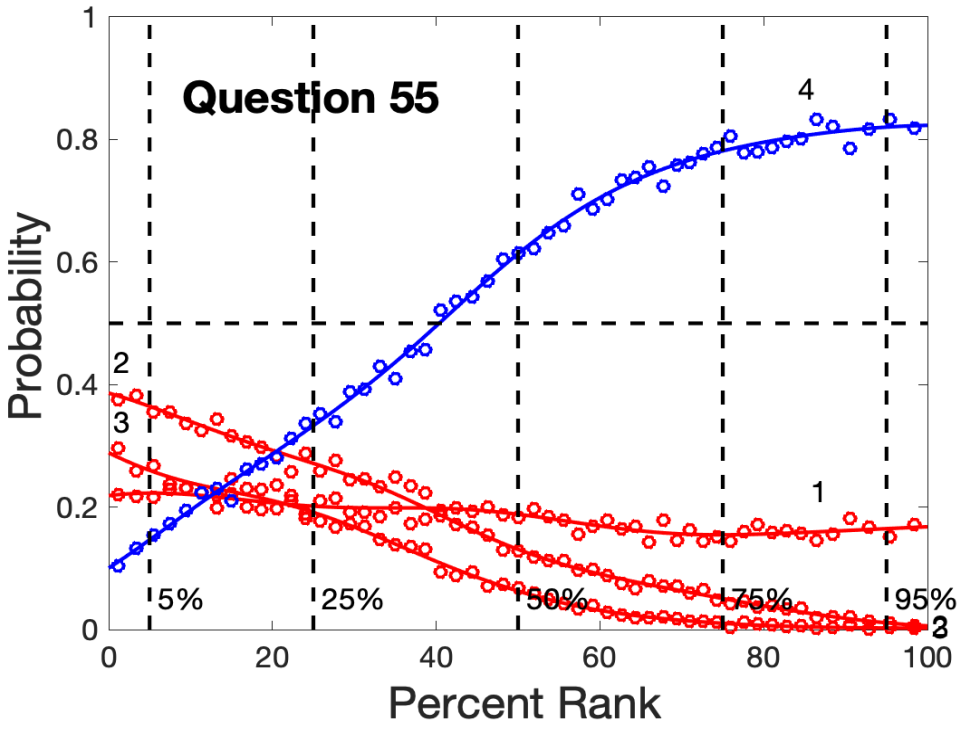


Figure 4.9: Proportion of choice curves displaying how proportion of choice varies over the total test score for all five answers for question 55 of the SweSAT-Q. The circles are the proportions observed with each of 55 bin locations and the solid curves are fitted to these data. The correct answer curve is in blue and the wrong answers are in red.

could be excused for thinking that cryptic phrase, "Information is insufficient," can't be right. Faced with this dilemma, the choice "greater" seems the least wrong, and especially when it is expressed in natural language rather than in the precise notation of mathematics.

We now add a fourth criterion for a good question, which question 55 fails to satisfy:

1. *Probability of a correct answer increases with performance or ability.*
2. *We can place the mid-point of the curve (probability 0.5) where we like so as to control the mix of easy, moderate and hard questions.*
3. *The rise in probability should be steep at the mid-point so as to provide relatively unambiguous evidence of performance level.*
4. *If a right answer curve does not reach one on the right, then it should at least be heading in that direction.*

4. divide by x so that the final expression $(x-1)^2/x$ now see is positive everywhere except at $x = 1$, where it is zero.

Chapter 5

Exploring Question Profiles

5.1 Introduction

In this chapter we broaden our appreciation for the information in question profile curves by having a look at questions from a wider variety of test and scale data. The first section continues to display question profiles taken from the two subtests of the SweSAT. Two effective questions from the SweSAT-Q subtest are followed by a seriously pathological question from the SweSAT-V. Section 5.3 offers a look at a test using test taker–constructed responses rather than test designers' multiple choice format. The final section leaves the academic testing environment and considers what is often called a *scale* or a *rating scale*. Here the scale responders rate the intensity of a set of experiences each of the levels among which the choice is made is associated with a score, which in this case ranges from 0 to 4.

We remind ourselves that the correct answer curve for academic tests are in blue, wrong answers are in red and missing or illegible choices are in green. But for the scale data, where there is no right answer, we simply plot each response in a different colour. We will drop the proportions from the plots in order to minimize distracting clutter, and only supply the smooth curves which fit the data.

Our comments on the question profiles are contained in the captions of the plots.

5.2 SweSAT Questions

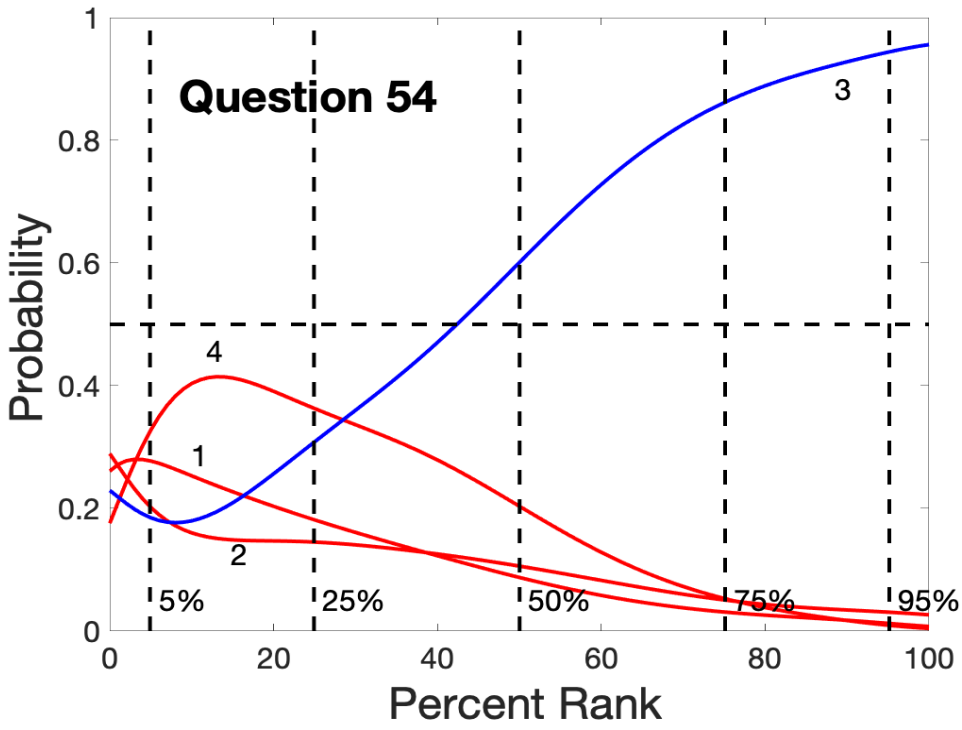


Figure 5.1: Question 54 in the SweSAT-Q displays a horizontal line and a line intersecting it from above at angles labelled $5x$ on the left and $4x$ on the right. *Answers are: (1) x greater than 20, (2) 20 greater than x, (3) x equals 20 and (4) Information is insufficient.* This is an effective easy question. It follows all the rules: the increase in probability is fast over the lower 50% of the test takers, and close enough to one to be satisfactory for the top 25%. The probability of the lowest 25% choosing the right answer is well below the chance or guessing level of 0.25 because wrong answers 1 and 4 draw them away. In short, someone failing this question is quite likely to be in the bottom 50% of the test takers.

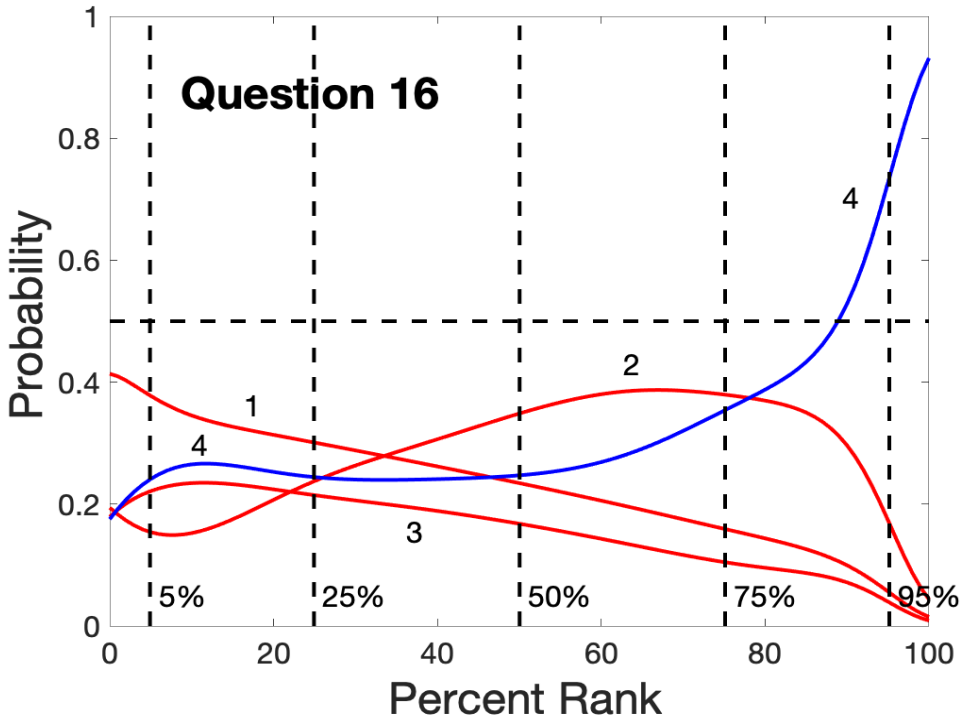


Figure 5.2: Question 16 in SweSAT-Q. *Two real numbers are related by x = -y. Answers are: (1) x greater than y, (2) y greater than x, (3) x equals y and (4) Information is insufficient.* This hard question behaves as a good question should. Wrong answers 1 and 3 draw the weakest test takers away from the right answer. Those with median scores find answer 2 more seductive. We wonder if there is a strong tendency for a test taker to avoid choosing "Insufficient information" unless no other answer seems plausible. In any case, getting this question right is good evidence that the test taker is in the top 25%.

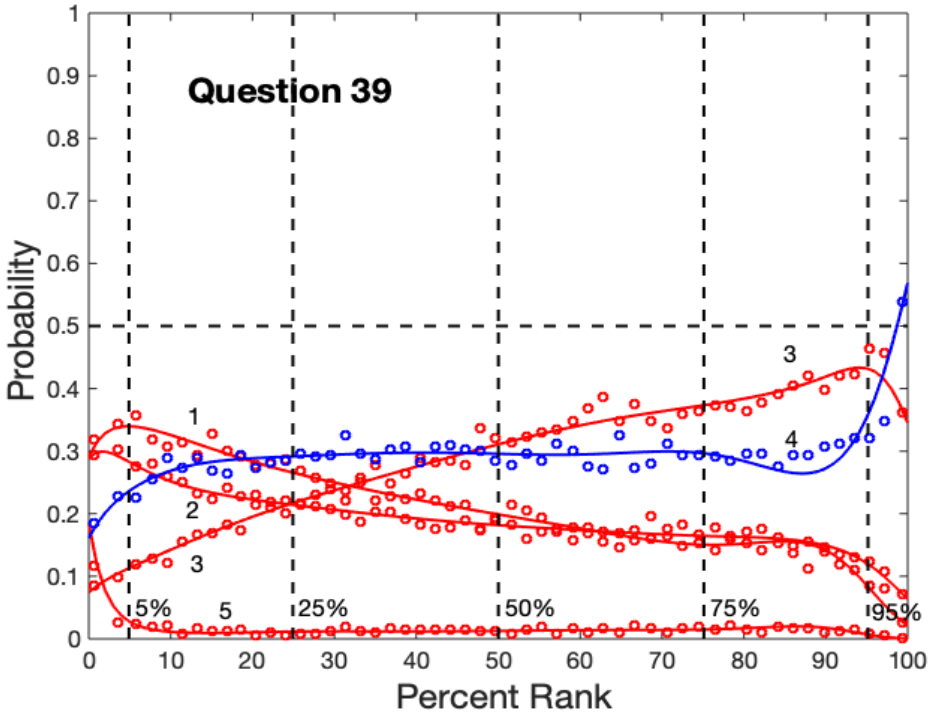


Figure 5.3: The probability of getting this question right for our best test takers would be only 0.55. Why? Unfortunately we cannot show the question because of copy-write issues, but we can say that the question involved extracting a quantitative answer from a complicated table. Answer 4 was scored as correct, but the answer 3 curve behaves more like a right answer curve. Can it be that answer 3 is also correct? In any case, none of these curves achieve the sharp increase in probability that we are looking for in a highly informative question.

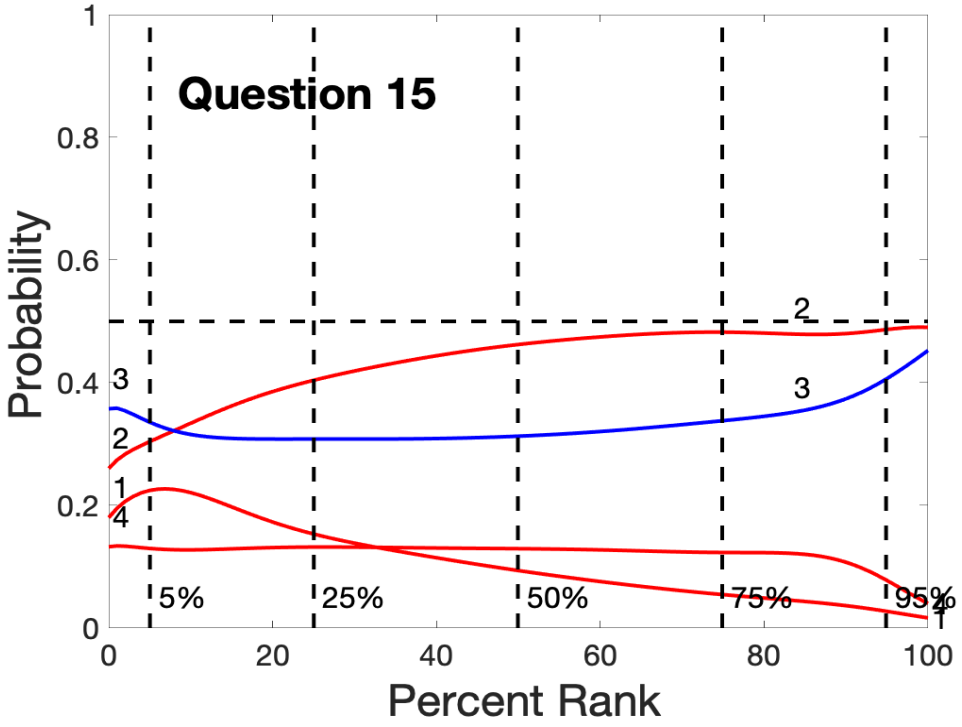


Figure 5.4: Question 15 of the SweSAT-V is a reading comprehension question in Swedish about DNA that we are not bothering to display. Here we see that no answer has a steeply increasing slope, as we also saw for SweSAT-Q question 39 in Figure 5.3. Again see that there are two answers that compete with each other for test takers over all performance levels. The correct answer 3 finally slightly dominates answer 2 among the top 5%, but fails to get anywhere near probability one. Could it be that verbal question 15 is like the SweSAT-Q questions 39 and 55 in having a second plausible right answer?

5.3 The National Math Test

We indicated in Chapter 3 that the National Math test differed from the SweSAT in requiring test takers to provide rather than recognize an answer, and it also in many items allowed for a range of values for the answers. This test is an example of what we call a *constructed response* test rather than a multiple choice test because the test taker, not the test designer, provides the answer. The test designer does provide a series of numerical values rather than just either 0 or 1 for many of the items, in order to give partial credit for wrong answers that nevertheless demonstrate some understanding. This is a teacher-graded test.

In this figure we have taken a different approach to displaying how precisely these curves are defined by the data, which involve 2235 test takers and 32 questions. The vertical lines superimposed on the curves at the bin centres define intervals that would contain the true rather than estimated value of the curve 95% of the time. These are called *confidence* intervals, and they nicely indicate what the range of reasonable plausible curve values would be given this amount of data. The precision is not nearly as good as for the SweSAT curves because only there are only a 1/20th of the number of test takers.

What seems striking about the constructed response curves of many of the National Test questions such as 15 and 30 is how rapidly the probabilities change over certain regions. This is because there is little opportunity for a constructed response question to get the right answer by guessing. Notice, too, that the right answer curves do get rather close to probability of one on the right or high performance side of the plot. Constructed response questions take longer to answer than multiple choice questions and can be expensive to securely score for large numbers of test takers, but they are more clear-cut or informative than multiple choice questions, and this can imply that the test itself may be more informative than a multiple choice test of the same length.

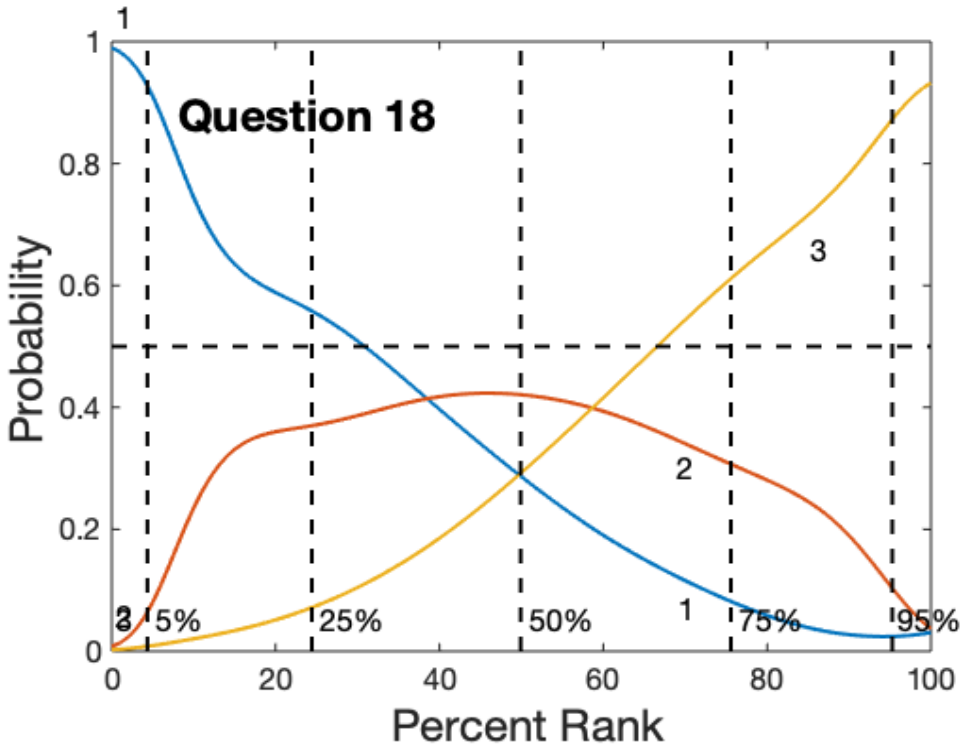


Figure 5.5: Proportion of choice curves displaying how proportion of choice varies over the total test score for all three answers for question 13 of the National Math test. *The question is: Simplify the expression* $(a^2-2b)/4$ *as far as possible if* $a = 2x+1$ *and* $b = 2x - 1.5$. The vertical lines around the curves indicate the precision with which the curve is estimated. This is a question of medium difficulty since it only requires the expressions for a and b and simplification. Curve number 1 is for an answer that gets a score of 0, and we see that for the top 25% of the students hardly anyone gets this score. A score of 1 is given primarily to the central 90% of the students, and the top score of 2 is associated with the top 75%.

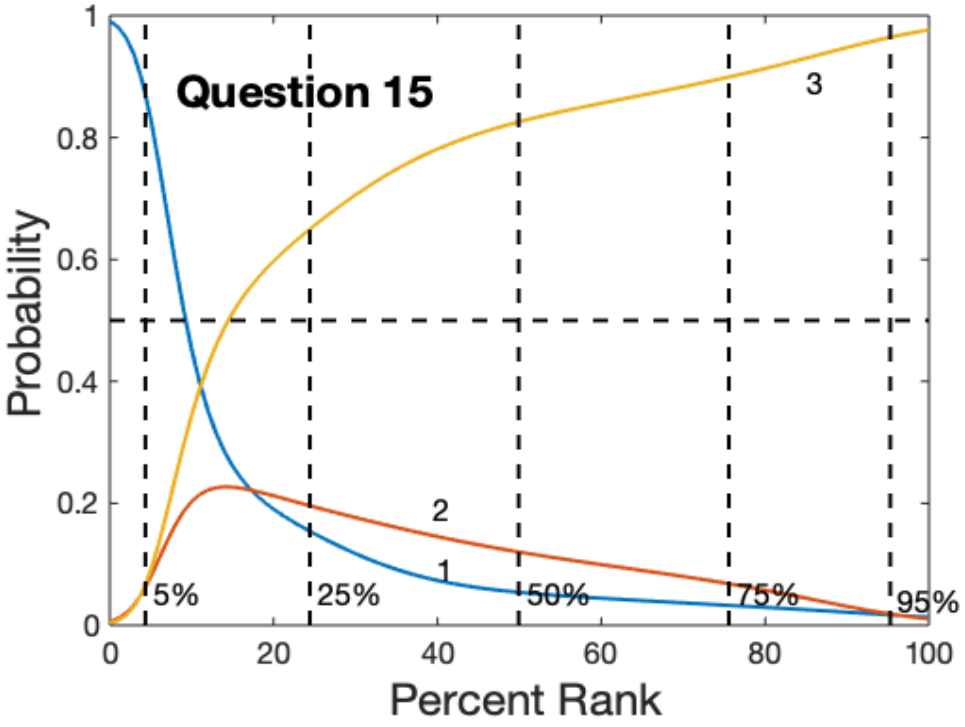


Figure 5.6: Proportion of choice curves displaying how proportion of choice varies over the total test score for all three answers for question 11 of the National Math test. *The question is: Solve the simultaneous equations* $y - 2x = 5$ *and* $2y - x = 4$. The vertical lines around the curves indicate the precision with which the curve is estimated. Here we see a fast drop in the 0 score curve, so that the top 75% of the test takers receive scores of 1 or 2.

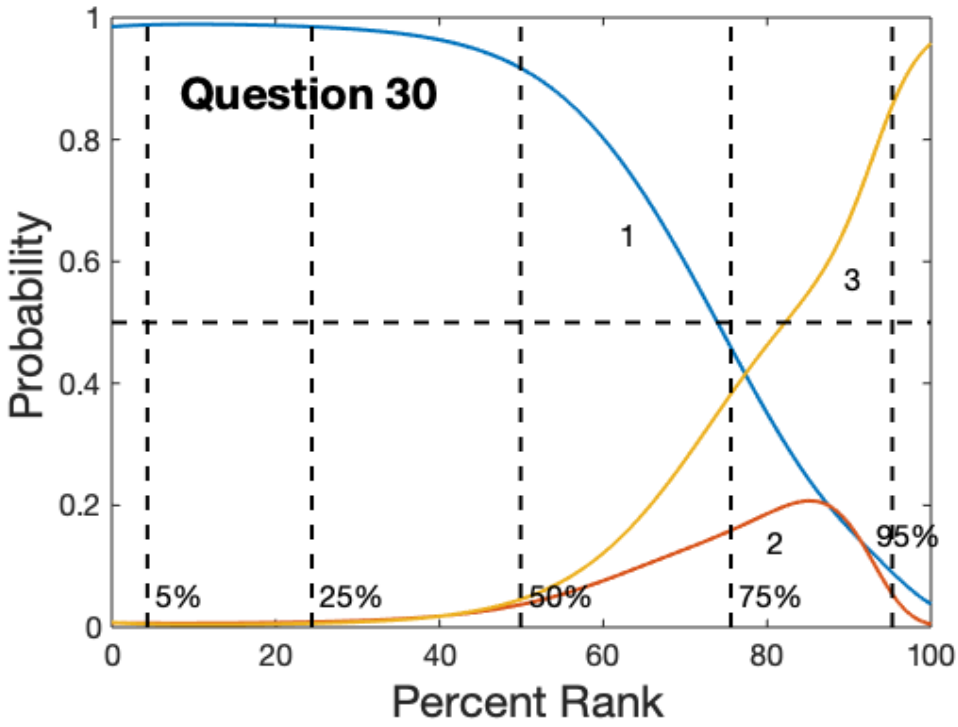


Figure 5.7: Proportion of choice curves displaying how proportion of choice varies over the total test score for all three answers for question 23 of the National Math test. The question is rather difficult because it requires solving two linear equations, but in a somewhat confusing format. The question is: *It holds for a function f that, where $f(x) = kx + m$, we have the two relations (1) $f(x + 2) - f(x) = 3$ and (2) $f(4) = 2m$. Find the function f.* The vertical lines around the curves indicate the precision with which the curve is estimated. Here we see a indication that a nonzero score implies that the test taker is in the top 50% and that a perfect score tends to indicate someone who is in the top 25%. The fact that all three curves are flat below the median test score indicates that this question will be useless at positioning people in that range.

5.4 The Symptom Distress Scale

This short 13-question self-report scale is typical of tens of thousands of questionnaires developed each year among the social scientists to permit scale takers to reveal how they feel they are with respect to some aspect of themselves that could usefully be expressed as a single scale score. We saw in Chapter 2 that the questions all pertain to some experience that could be a result of having some form of cancer and that could be treated by a professional care giver. There were 473 patients in the survey that provided the data. While this not as many as we had for the Swedish tests, it is sufficient to define aspects of the shapes of the curves.

Because the answers in scales are usually presented in increasing order of intensity, satisfaction, or whatever is being assessed, the first curve tends to be high on the left boundary and slope down to zero. In this questionnaire, the order is the intensity of symptoms, and those who choose the first category are reporting lowest level of intensity. The higher the overall distress level is, the less the probability is that the first answer will be chosen. At the other extreme the last answer tend to be chosen only by those with the most intense level of distress, and as a consequence the curve is the flip of the first curve, moving from a low level to a high level. For answers between these extremes, the curves tend to peak at intermediate points and then return to zero as the overall distress increases. What we want to see is how rapidly the choices of the answers move from the least distress to the most.

With self-report scales, there is inevitable wide variation over those completing the scale as to how likely they are to report a specific level of the property. The toleration of pain, for example, depends on many factors including how much pain a person considers normal and relatively tolerable. This subjectivity in the definition definition of answers tends to imply that no curve will change sharply either up or down. Consequently, we can expect that the random variability in scale scores will tend to be high relative to that of a comparable performance curve.

In Chapter 2.5 we saw that the only the top 20% or so of the respondents to the Symptom Distress Scale reported scores beyond the center of the scale range, which was from 0 to 37. The second or “mild” category was the most frequently chosen. We can be thankful that the intensity of distress from this awful disease is so low for so many.

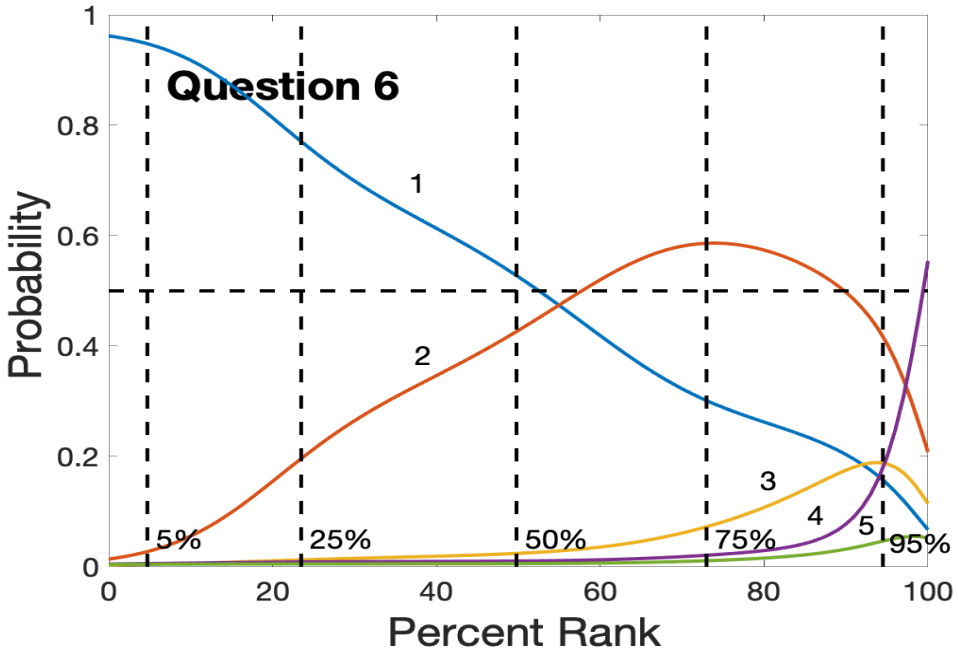


Figure 5.8: *Inability to concentrate.* We all have trouble concentrating from time to time, and it is not surprising that the choices of the first category declines slowly and tends to be chosen by all but the top 25% of patients. Likewise, given that having any form of cancer is likely to be distracting, the second mild lack of concentration is chosen over a wide range of distress levels. Only the top 5% report the two fourth level, and hardly anyone reports the highest level.

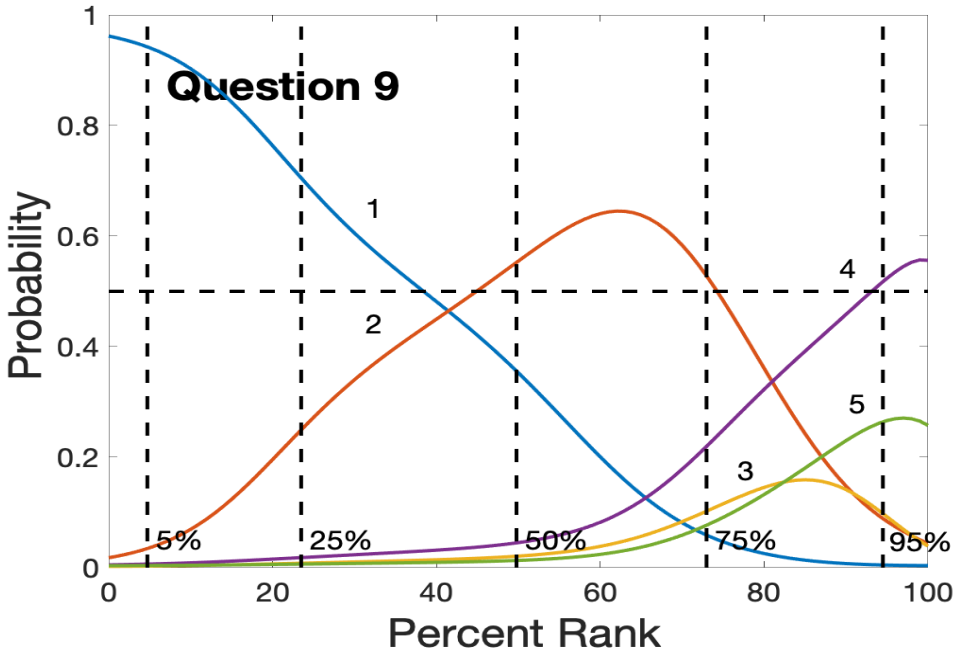


Figure 5.9: *Intensity of pain.* Pain intensity ranges from none to moderate for most respondents, and the two highest categories are used by only the top 25%. Interestingly, the middle category is used by hardly anyone, suggest that pain tends to be in essentially three states: mild, present but tolerable, or intense.

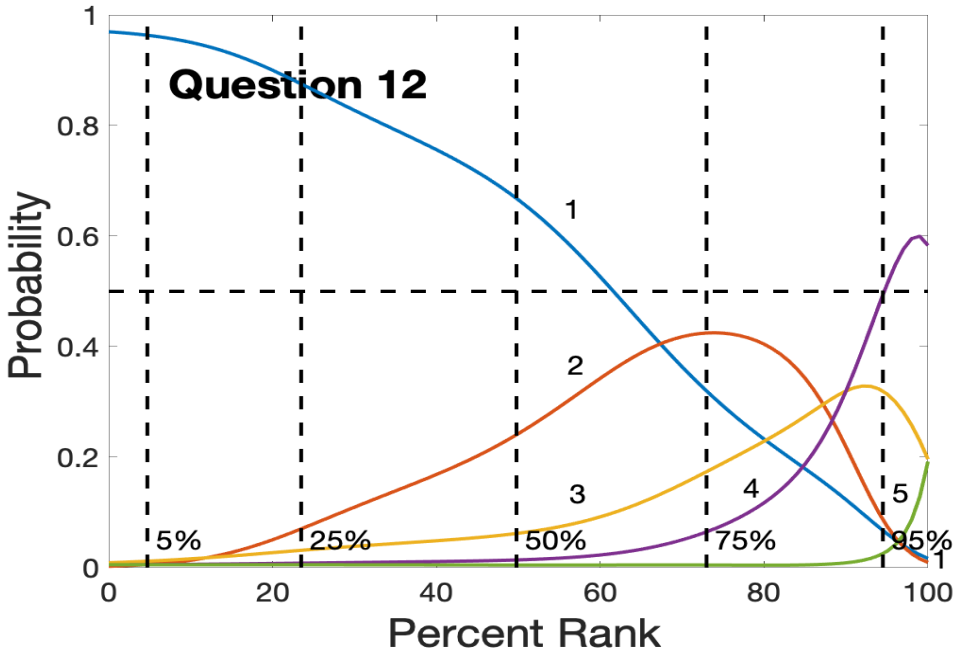


Figure 5.10: *Loss of appetite.* The bottom 50% seem to experience little loss of appetite, but among the top 50% all categories are chosen with noticeable frequency. It would appear that appetite is an experience to which they are finely sensitive. Could the quality of hospital food have something to do with this?

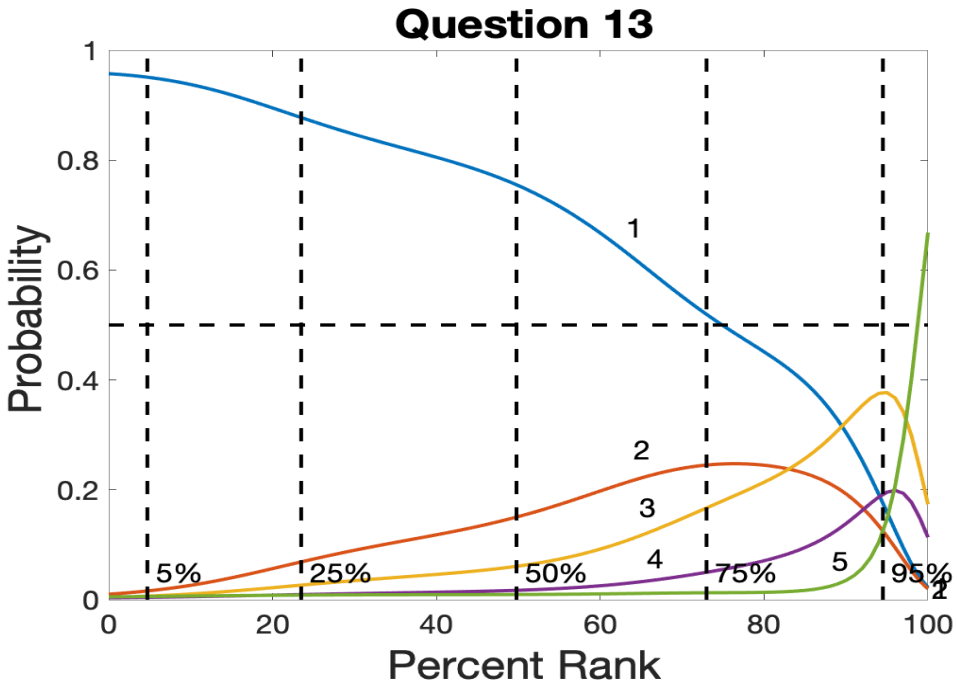


Figure 5.11: *Deterioration of appearance.* This is extreme for only the top 5%, and most are not affected at all or only mildly so.

5.5 How and When are Test Data Informative?

Now that we have taken a look at four sets of tests, three of which are tests of performance involving right and wrong answers, and one of which is a scale involving answers indicating levels of experience, we can reflect on the broad question of when test data are useful and when they are not. We use the term "informative" here because tests and scales seek to position each test taker on a line. The lines that we have considered reflect knowledge of mathematics for the SweSAT and the National Mathematics tests, and level of distress of a patient under nursing care for the SDS scale. A test is informative for a given test taker if the data define a position on the line relatively accurately. That is, we feel relatively secure about how well a test taker has been positioned relative to his fairly close neighbours' positions on the line.

We have, rather vaguely at this point, associated "informative" with questions whose answer curves or profiles are rapidly changing level. In the next chapters we will seek to pin down the notion of change, but at this point a fuzzy idea of change suffices. By contrast, a specific answer curve is uninformative over a designated range of positions if it does not change level much, so that its curve is more or less flat over that range. A question is uninformative over the range in question if all of its curves are uninformative. Review question 39 of the SweSAT-Q for question 15 of the SweSAT-V if you need reminding about what an uninformative question looks

like.

The more questions there are in the test that are informative for a given test taker's position, more accurately that position will be defined by the test. This has some surprising and also important implications. The closer the test taker gets to either lower or upper boundary on the score index, the fewer the number of informative questions there are apt to be. If they are estimated to be on either boundary, the test is essentially incomplete because there are no questions beyond the boundary to indicate where the test taker should be positioned. Therefore, test takers having either zero or perfect test scores are "beyond the test," and the test itself only tells us whether they are below or above the test.

A test is often used to assess a special group of test takers. For example, we can imagine that the nurses who look at data provided by the Symptom Distress Scale are more interested in patients in great distress who can benefit from treatments that provide relief. If there are relatively few patients in this category, the scale will have rather limited value and provide only vague indications. On the other hand, the SweSAT subtests are designed to highlight test takers with the aptitude to benefit from expensive and time-consuming higher education. The value of the SweSAT depends on how many takers of this test actually have that level of aptitude.

Chapter 6

From Probability to Surprisal

6.1 Introduction

In our preceding chapters we have concentrated on graphical tools that can show us how well a test or scale question or one of its answer is doing its job. We concluded that "doing its job" has something to do with whether the probabilities curve that we inspected were rising or falling, since a region over which the probability level was unchanging was telling us little or nothing about about where a test taker should be placed on whatever scale index we used to describe performance. For example, we saw in our plots question 46 from the SweSAT math subtest that probability of choosing the right answer was only increasing for the top 50% of the test takers. On the other hand, for the somewhat infamous question 55, over this higher performance region the "right"answer probability was not only not changing, but was also not getting anywhere near probability one for even the top 5% of the test takers.

This chapter is the first of three where we develop the concepts that allow us to put these intuitions to work to produce better test scores. We will continue to draw only on math skills that you acquired in secondary school, and we will also continue to present the necessary ideas in graphical form. If you want to skip these three chapters at this point, we understand. But if you want to understand why our better scores are so much better, you will need to come back to them.

We will now consider whether the concept of probability was a historical accident, and we will find that a simple transformation of it, that we call *surprisal*, is a measure of size, which we call a *magnitude*, and therefore is a quantity like most others that we use freely in everyday life. Specifically, surprisal is a measure of *information*, and it precisely the information in answer choices that we need to use effectively. But if you are the kind of person that reads footnotes and knows about logarithms, you will also see the math that underlies our scoring system.

6.2 Probability Curve Slope

Probability informs us about how often something will happen. We often estimate probabilities by counting the number of times the event has happened in the past, and then divide by the total number of times we have observed whether it has happened. The resulting ratio is called a *proportion*, and is an *estimate* of the probability value that we seek. This estimate becomes a probability when the counts become really large. We use probability to measure how often an answer to a question will be chosen by test takers at a particular performance level.

In this chapter we introduce an alternate way of expressing our intuitions about whether this (test taker/test question) event will be a success. We call this the *surprisal* of the event. We do this because (a) we conjecture that surprisal will make probability easier to think about, and (b) because the mathematics is made easier to understand. And, most importantly, surprisal will be the counterpart of the weight of the dumbbell in the weight-lifting story back in Chapter 1.

We noted earlier that an effective question requires that the probability of choosing an answer increases steeply somewhere in the range of performance levels. We use the term *slope* for the speed with which a moving on a curve is moving up at a particular point. A rapid increase or a high slope at a specific performance level will effectively separate test takers below that level from those above it. This steepness principle applies just as much to wrong answers as it does to right answers. If the wrong answer probability descends sharply at a point, and someone at that performance level chooses that wrong answer, we will have strong evidence that test taker's performance is below that point.

Assessing steepness by visual inspection of the probability curve is made difficult, however, when a probability curve approaches either zero or one, where it has little room to exhibit a rapid change. We need to transform probability in a way that still tells how often an answer will be chosen, but that removes upper limit of one that characterizes probability. This is exactly what converting probability to surprisal achieves.

6.3 Why is Probability so Difficult to Understand?

If either you have taken an introductory statistics course or you have taught one, you will know that probability is a nemisis for many students.[1] Most of the quantities that we manipulate in our day-to-day experiences share a few basic characteristics. We refer to quantities with these properties as *magnitudes*. Counts of discrete things

[1]You are less likely to know that a Nobel Prize was awarded to Daniel Kahneman in 2002 for his work with Amos Tversky on why nearly everybody has trouble with probability, and especially the probability of rare events. Their theoretical work, called *prospect theory*, showed how we are maladapted to manipulating probabilities, and explained why so often we make irrational decisions concerning rare events.

also share these properties, except that they are not continuous because they increase in a staircase fashion. We first look at the characteristics of magnitudes, and then discuss how probability values depart from magnitude properties.

6.3.1 The Magnitudes of Everyday Life

We order our lives using a limited number of quantities. These quantities are referred to by scientists as magnitudes, and they share these properties:

- The physical properties that we experience from moment to moment, such as distance, speed, mass, energy, heat, electric charge, pressure, force and, most importantly, money, all have a state that we call "zero." From a psychological perspective, zero and minuscule quantities are usually regarded as without interest or value. That is, as ignorable.

- Otherwise these quantities are positive.

- They have a nonzero magnitude which is assigned the value one and called its *unit*. The choice of the unit magnitude value is arbitrary and therefore selected for convenience.

- Physical quantities, unlike counts, are usually continuous, meaning that we can imagine infinitely small changes in them.

- Most magnitudes have no upper limit, or, like velocity, have upper limits so far away from our experience that we ignore them.

- When multiple magnitudes with the same unit and counts come from independent sources, they can be added and subtracted as we please without any change in their status as quantities. In fact, addition and subtraction are the *only* way that we can combine them.

- We do divide one difference by another, as we shall soon see, but then the result has no unit and is no longer a magnitude.

The conditions that a property be either zero or positive can be relaxed. Take time for example. Time can be a magnitude provided that we have a specific starting time in view, like midnight, when the starting gun went off, or 0 years Anno Domine (AD). We call this type of time *elapsed* time. But often we just want to think of time as stretching back forever and forward (hopefully) also forever. This type of time we call *duration*.

Temperature is another example. Heat has a zero, called *absolute zero* or *zero degrees Kelvin,*[2] and is therefore a magnitude. But it is awkward to keep referring to something that is so far away from anything we can experience. For that reason, we pick something that we experience, and assign zero arbitrarily to it. In the Celsius or centigrade scale, the freezing point of water plays this role, and absolute zero is -273 degrees C. As for Fahrenheit, who know why zero is where it is? Best to ask Siri.

But otherwise, time and temperature satisfy the *ruler criterion*:

A fixed increment means the same thing wherever the increment is applied.

As a carpenter would say, "If the board is 1/4 inch too long, it's just plain too long, no matter how short or long the board is."

6.3.2 Probability is not a Magnitude

A proportion is a ratio of two counts, the denominator of which is the largest possible count. Probabilities are what proportions become as the counts involved increase without limit. The ratio of days without rain this month to the number of days in the month is a proportion, and as such ranges between zero and one.

Here are the basic properties of probability:

- Probability has a zero, but it is usually regarded as unattainable and, if the probability is close to zero, either as a catastrophe or as an event of otherwise great interest.
- Probability has an upper limit of 1, and the event associated with this value is often regarded as of little interest. This contrasts with magnitudes where the bigger they are, the more excited we become.
- If probabilities arrive from independent sources, we multiply them to get the probability that the events occur simultaneously. Unfortunately our brains do not perform multiplication well, and are even worse at division.
- Probabilities are always ratios, which is why they do not have a unit of measurement.
- We do add probabilities, but only when they are segments of the same totality that has probability one. In this case, though, these segments are not independent of each other since nothing can be in two segments at the same time. A day with rain can't be a day without rain.

[2]Scientists are a somewhat vain bunch, and like to use each other's names for things. Here the Scot, Lord Kelvin, bless his heart, gets a boost. Mathematicians, though, are the worst at using proper names for often obvious things, and seem somehow to have complex foreign-sounding names, too.

Table 6.1: The relationship between the numbers of consecutive heads in a coin toss and their probabilities.

Number of Heads	Probability
0	1
1	1/2
2	1/4
3	1/8
4	1/16
5	1/32

6.4 Transforming Probability into Surprisal

The modern use of the term "probable" appeared during the evolution of the mathematics of gambling in the eighteenth century. Games of chance by their nature involve long sequences of repetitions of bets and other events related to money. Or, we might now say, are tests of a particular sort. The data that probability theory was designed to explain were counts, and often counts of rare events. These rare events, such as a big win at Monaco, were eagerly awaited; and, assuming that the games were fairly played, certainly surprising.

Consider, for example, coin tossing where the coin has not been tampered with and the coin tosser does not know the orientation of the coin before the toss. Let's identify probability one with the certainty that the coin will be tossed, but has not been yet.

Then we know that the probability of getting one head is 1/2, and that of two consecutive heads is 1/4. That is, the two-head event involves multiplying the probability of a single head times itself.

Now let's call the number of heads *surprisal*,[3] so that the surprisal of one head is one and of two heads is two, and etc. Surprisal zero is identified with the fact that the coin has not yet been tossed. As we imagine more and more consecutive heads, we get a sequence of probabilities and surprisals that looks like those in Table 6.4.

The number of heads is of course a count, and therefore counts can be added. We notice that the probability of a trial involving 2 heads plus that involving 3 heads, or 5 heads, has the same probability as that of 5 heads thrown consecutively in a single trial. We aren't particularly impressed by two heads, but five heads does capture our attention. In fact, the famous level of probability 0.05 used by scientists to declare a result "significant" lies somewhere between 4 and 5 heads in a row.

Now let's turn surprisal into a continuum. Let's call a suprisal the probability of throwing m heads in a row some percentage of the time and $m+1$ heads in a row the

[3]The term was first introduced in the context of the physics of heat by Tribus (1961) and is now widely used in the physical sciences.

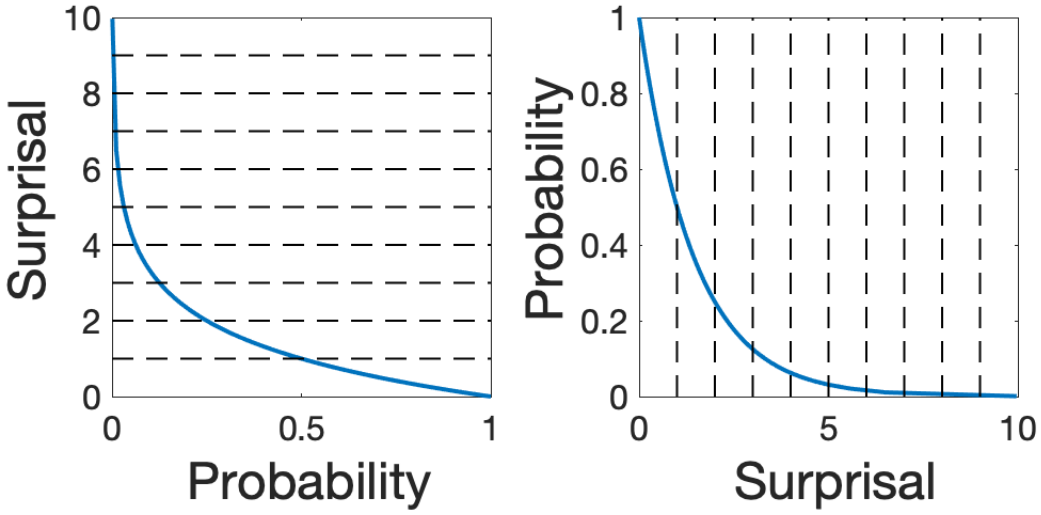


Figure 6.1: The left panel show how surprisal depends on probability, and right the inverse relation of how probability depends on surprisal.

rest of the time. With a little experimentation, we will see that the surprisal of 0.05 is about 4.32; and the surprisal of 0.01 is about 6.64, this probability being considered the "slam–dunk" for proving that the outcome of an experiment didn't just happen by chance. If we turn the experimentation over to a computer, in no time we will have the relationships between probability and surprisal that we see in the two panels of Figure 6.1.[4]

Here is a little word equation to express what surprisal is as a function of probability:

$$surprisal(probability) = an\ average\ number\ of\ consecutive\ heads$$

so that, for example,

$$surprisal(0.05) = 4.32.$$

The pair of parentheses on the left side of this equation indicate that surprisal is

[4]Surprisal is a part of the mathematical theory of information that is widely used in many fields of science, including the study of the transmission of signals across networks. In that larger context surprisal is referred to as *self-information*. Like proportions or probabilities, surprisal values are associated with a score index values, for which we use the symbol θ. The letter S has so many different uses that we have decided to use W instead for a surprisal value, and $W(\theta)$ for a surprisal value associated with a score index value of θ. Think of W as standing in for "Wow!" and the number of bits that it represents. The surprisal transformation $W(P)$ of probability P is $W(P) = -\log_2(P)$ where 2 is called the *base* of the logarithm. The *inverse* transformation from surprisal measured in bits back to probability is $P(W) = 2^{-W} = (1/2)^W$.

a *function* of probability, so that for every probability value there is a corresponding surprisal value, and the precise value that is being changed into probability is contained within the parentheses.

Defined using coin tosses, surprisal is a magnitude with the value being the *bit*, the basic element in computer storage. We could just as easily used any other easily replicated event as the unit magnitude of surprisal, such as the single toss of a six upon throws of dice. In that case the unit would correspond to a base probability of 1/6. The insurance industry often uses rare events as a base for surprisal because they tend to trigger insurance claims. The Ottawa River in Canada is supposed to reach a certain flood level on average only once in a hundred years. But somebody forgot to factor in global warming, and in fact that level has been surpassed twice in the last three years, with a destructive tornado thrown in for free.

During the writing of this book, a Boeing 737 Max 8 jet crashed killing all passengers. There was not much response beyond a routine investigation. Two months later it happened again. Immediately all 737 Max 8 aircraft were grounded. When the event is the crash of a new airliner, the basic unit is one crash in, say five years, which the air travel industry might accept as an acceptable frequency. One in that period would be surprising, but two in a couple of months was beyond shocking.

If the theory of uncertainty had unfolded somewhere else than in the gambling parlours of France, it would be unsurprising that surprisal would have emerged as the fundamental concept underlying statistics. If that had happened, we would be calculating uncertainties by adding and subtracting and for sure the field would be have been a lot kinder to the introductory statistics student. Surprisal and probability are two names for the same thing, and we think that we would often be better off with surprisal.

But probability does have an important role; we use it to *predict* events that are not rare. For example, most of us are comfortable with a weather forecast that uses percents to tell us how likely it is to rain or snow. And we would also be comfortable with a prediction that a bright student has a 90% chance of answering a question correctly and a not-so-bright student has a 60% of chance choosing a wrong answer. Let's put it this way: If you can easily see a probability in a graph, chances are you will be happy with it. But, of course, we could easily get used to saying that the bright getting the question has a surprisal of only about 0.15, and that the surprisal of the other student getting it wrong is about 3/4 of a bit. Both surprisals are lower than the 1 bit associated with 50/50 odds.

6.5 Comparing Sum Score Surprisal Distributions

Our first application of surprisal is to comparing the distributions of sum scores for the SweSAT-V and SweSAT-Q. We can now compare the two panels of Figure 2.1 using a single bar chart showing the surprisals side by side. In order to not overload our eyes, we replace the 81 possible sum score values by 20 bins, each containing

four consecutive sum score values.[5] Then we compute the proportion of sum scores occupying each bin by dividing the number of sum scores in a bin by N, the total number of test takers. Finally, we convert these proportions to surprisal values. In Figure 6.2 these surprisals tell us for each bin how surprising it is that a test taker's sum score would be in that bin. We notice right way that we are not nearly as surprised to see sum scores in the middle bins as we are in the end bins, corresponding to the much higher probability that a sum score will be in the central zone.

We now see easily the differences between the two surprisals, and especially for the bottom and top bins. The first two bins, containing sum scores 0 to 8, indicate that the surprisal of a test taker being in those bins is about the same, namely about eight bits. We think that this is probably because the small number of test takers in this zone are guessing or otherwise choosing answers with no reference to knowledge of the respective subjects. For bins 3 to 9 (sum scores 9 to 36), we see verbal surprisals that are rather higher than their quantitative counterparts, suggesting that the verbal subtest is quite a bit easier than the quantitative subtest because we are more surprised by a verbal sum score in this low-scoring zone than we are by a comparable quantitative sum score.

For the middle bins 10 and 11 (sum scores 37 to 44), the two surprisals are about the same. For the remaining bins, however, the surprisal that a quantitative sum score is in that bin is much higher, which again indicates that the quantitative test is substantially more difficult and therefore has fewer sum scores in high sum score zone. Moreover, when we compare Figure 2.1 to Figure 6.2, we see more clearly in the surprisal version the differences associated with the more extreme sum scores. These stand out due partly to the differences between surprisals in these end zones being much larger than the differences between the corresponding probabilities, and partly due to these differences being higher in the plot.

Indeed, although we can see differences in proportion/probability plots, these difference have no fixed meaning. A small difference of, say 0.01, when the probabilities involved are near 0.5 is far less interesting than the same difference for probabilities around 0.05, but to our eye they appear equivalent.

Surprisal is also a convenient way to compare two tests with different numbers of questions. Figure 6.3 shows the surprisal values within 20 bins for the SweSAT-Q and the National Test. Over the remainder of the lower half of the scores, the surprisals for the SweSAT-Q are higher indicating that low scores are more prevalent in the National Test. This is easy to understand when we remember that a test taker cannot benefit from guessing on the constructed response National Test. The larger National Test surprisals in the central bins numbered 7 to 11 indicate that a larger proportion of SweSAT-Q scores are within this zone. At the upper end, however, National Test scores are more likely to be found than SweSAT-Q scores, so that the difficult National Test questions are not as difficult as the comparable SweSAT-Q questions.

[5] Zero sum scores are added to the first bin.

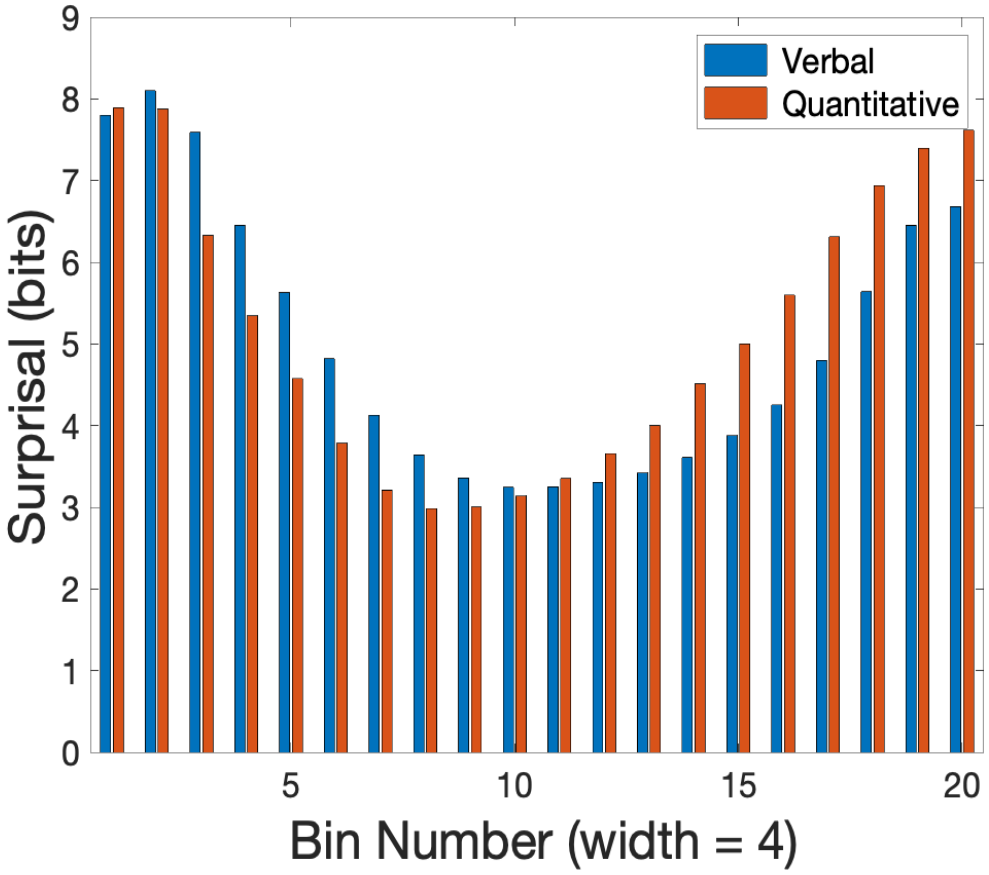


Figure 6.2: The bars in each bin show the two surprisal values for the proportions of test takers in the SweSAT-V and SweSAT-Q subtests within that bin. The proportions in the bins are for four consecutive sum score values, except for the first bin which also contains sum score 0.

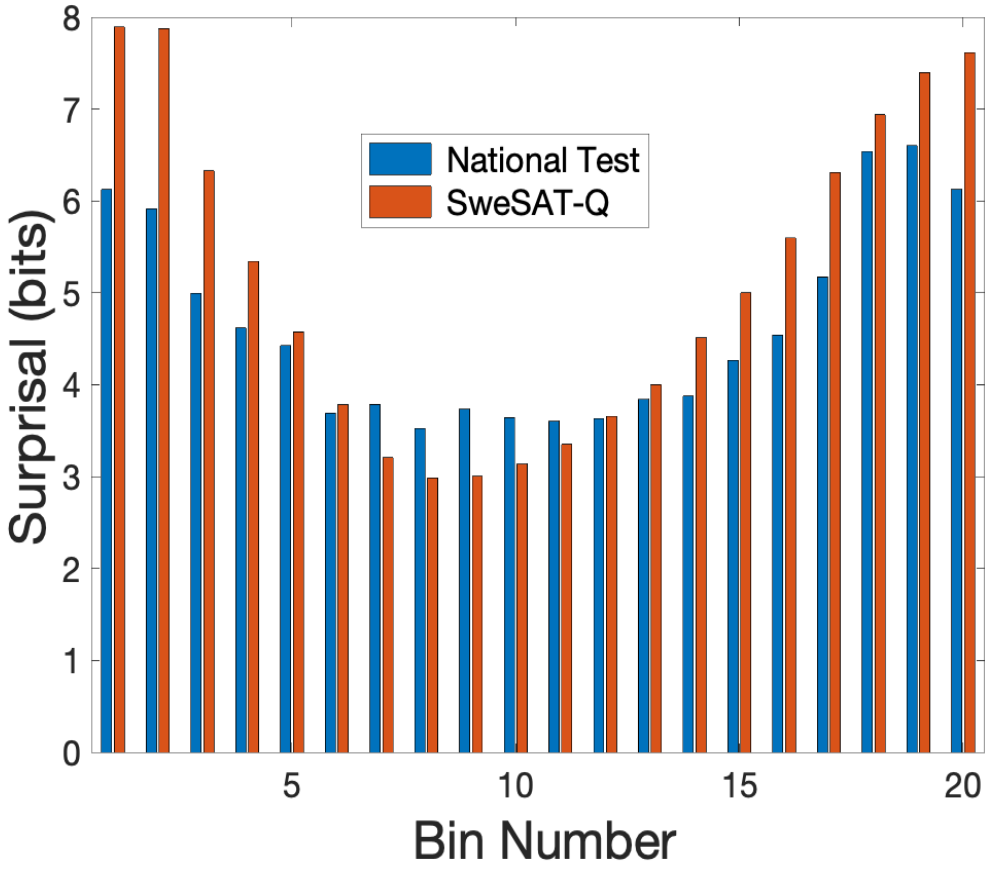


Figure 6.3: The bars in each bin show the two surprisal values for the proportions of test takers in the SweSAT-Q subtest and National Test within that bin. The proportions in the bins are for four consecutive sum score values, except for the first bin which also contains sum score 0.

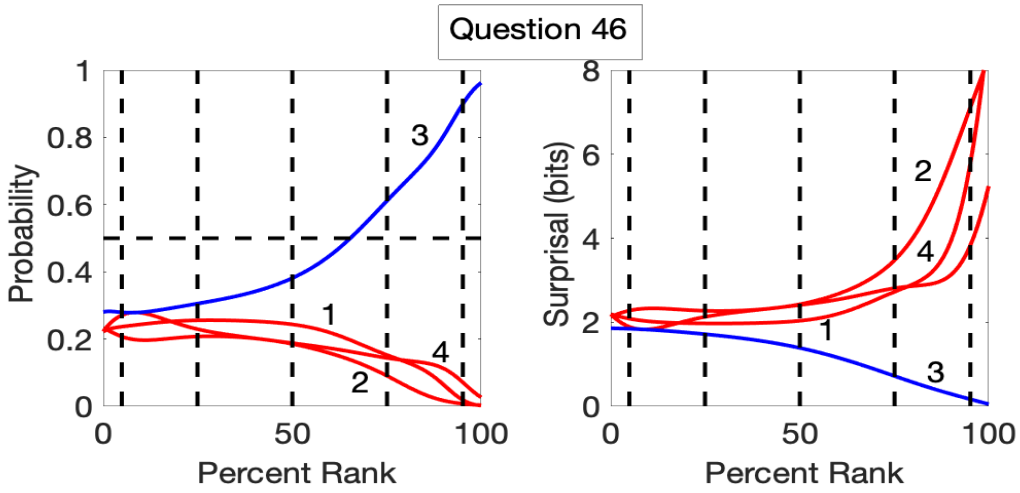


Figure 6.4: Question 46 in SweSAT-Q, the quantitative subtest of the SweSAT. The top panel displays the probability curves for the correct answer (blue) and incorrect answers (red), and the bottom curve shows the corresponding surprisal curves.

6.6 Surprisal Curves for Answers

We now use surprisal in order to assess question performance. First we will look at a direction comparison of answer surprisal curves and their probability counterparts. Figure 6.4 shows both the probability curve for the correct answer to question 46 of the SweSAT-Q and its corresponding surprisal curve. We see, as in Figure 6.1, that when the probability increases to near one, the surprisal decreases to near zero. This says that we aren't much surprised when a really smart test taker gets this question right. We are mildly surprised, by just over two bits, when a poor soul on the left of the plot answers the question correctly, remembering that one can get a correct answer just by guessing with probability 0.25, which corresponds to two heads in a row.

Figure 6.5 shows the five surprisal curves in the bottom panel for question 55. This time the surprisal of a top student getting the right answer is not quite as close to zero, and there is a two-bit level of surprise if wrong answer number one is chosen instead. The surprisals of choices among the other two wrong answers could be as

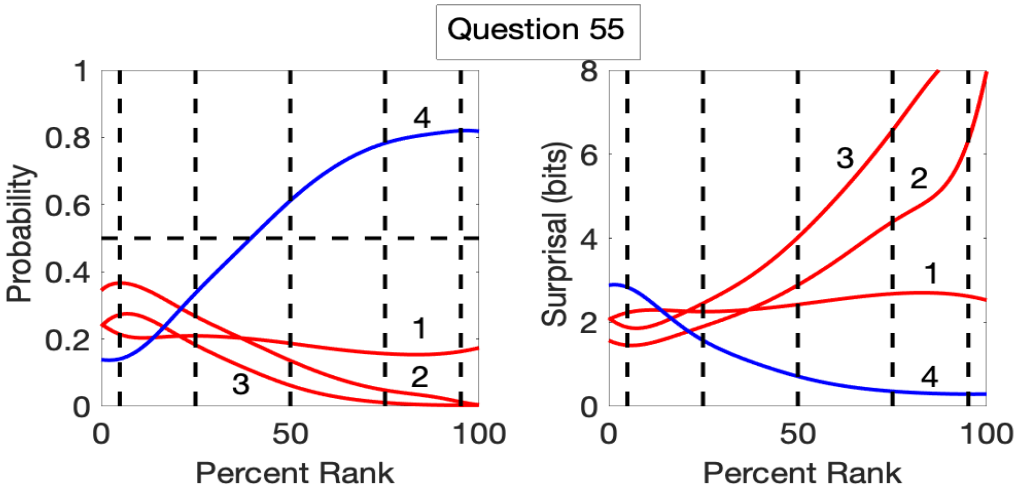


Figure 6.5: Question 55 in SweSAT-Q, the quantitative subtest of the SweSAT. The top panel displays the proportion data and the smooth probability curves for each answer, and the bottom panel shows the corresponding surprisal values.

high as that of tossing nine heads in a row.

Figure 6.6 displays all of the correct answer surprisal curves for the SweSAT-Q. One is struck by the surprisal of nearly two for the purple curve at score index 76 indicating the beginning of the top 25% of the test takers.

6.7 Surprisal-Slope

Each answer for each question at any performance level is now equipped with a value, which we call surprisal, which can be added and subtracted in a meaningful way. This value is the counterpart of the weight of a dumbbell in a weight room, the height of the crossbar, or the resistance to the rotation of a driveshaft in a transmission. Our discussions of these settings suggest that if we want to construct an indication that a test taker belongs either further up the performance scale or further below, we should consider the size a difference between surprisal values at the test taker's current location. This difference is constructed by looking at how surprisal rapidly surprisal

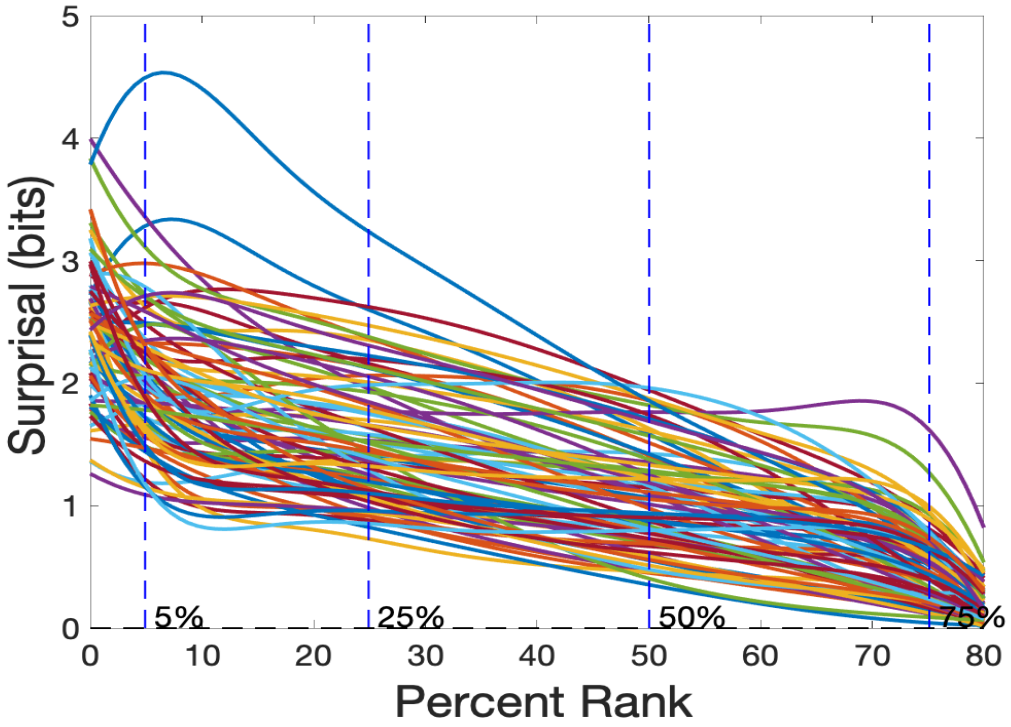


Figure 6.6: The surprisal curves for all of the correct answers for the SweSAT-Q.

is changing at any particular performance level. We now construct such a difference, which we will refer to as *surprisal slope.*

We already have an intuitive feel for how rapidly a curve is either increasing or decreasing at a particular point on the curve that interests us. We usually use the word *slope* for this property.[6] How do we move from intuition to quantification? First of all, zero slope applies naturally to a flat spot on the curve where it is neither rising or descending. A rise is naturally expressed by a positive number and a drop by a negative number. The closer a curve comes to increasing vertically, the greater the slope, and without limit; and similarly for a near vertical decrease indicated by an unbounded negative value.

But how do we calculate the size of the slope? Our answer to this question must deal with the fact that the slope is constantly changing for most curves, and certainly for the surprisal curves that we have been inspecting in the previous chapter. The slope of a curve at a particular point is a ratio. The numerator of the ratio is how much it rises over a small interval and the denominator is how large the small interval is. This quotient is often called the "rise over the run" at that point. Slope has a unit since when the rise equals the run, the slope is at an angle of 45 degrees from the horizontal and therefore has the value one.

But what is "small" in this context? By "small" here, we mean so small that making it even smaller would only change the slope ratio in, for example, the fifth decimal point, or, to put it another way, change the quotient by too small an amount to matter. A practical definition of "small" is an interval so tiny that the slope is for all useful purposes constant over the interval.[7]

In Figure 6.7 the longer straight red line that passes through each point has the same slope as the curve at that point. The slope of this line is computed by the change in the height of the straight line (the "rise") divided by an increase of one in the horizontal position (the "run")[8].

We can approximate well the slope at a point by making the change in horizontal position really small, and, consequently making the change in height proportionately small. How small? Well, those of us who convert mathematics into computer code lead high-risk lives, and routinely need to check for errors in either the mathematics or the code. We do this by computing the height of the curve a specified position, and then at a point 0.01% alway from the point. If the ratio of the height difference

[6]Since we assume in this book that you have made it through secondary school, you most likely have already studied slope. It was called the *tangent* of an angle θ from the horizontal, and is expressed in math notation as $\tan(\theta)$.

[7]Humans are flat-landers, and are therefore really bad at estimating slope. A slope ratio of 0.08 or 8% doesn't look like much in a graph, but if you're a cyclist it looks like a wall. The routes chosen for the Tour de France do contain such slopes, but seldom.

[8]The mathematical calculation of slope is taken up in calculus courses. The mathematical term for slope is the *derivative*. The term was introduced by the French mathematician Lagrange at about the same time that "probability" was defined, so we may not be surprised that the term never took off on the street.

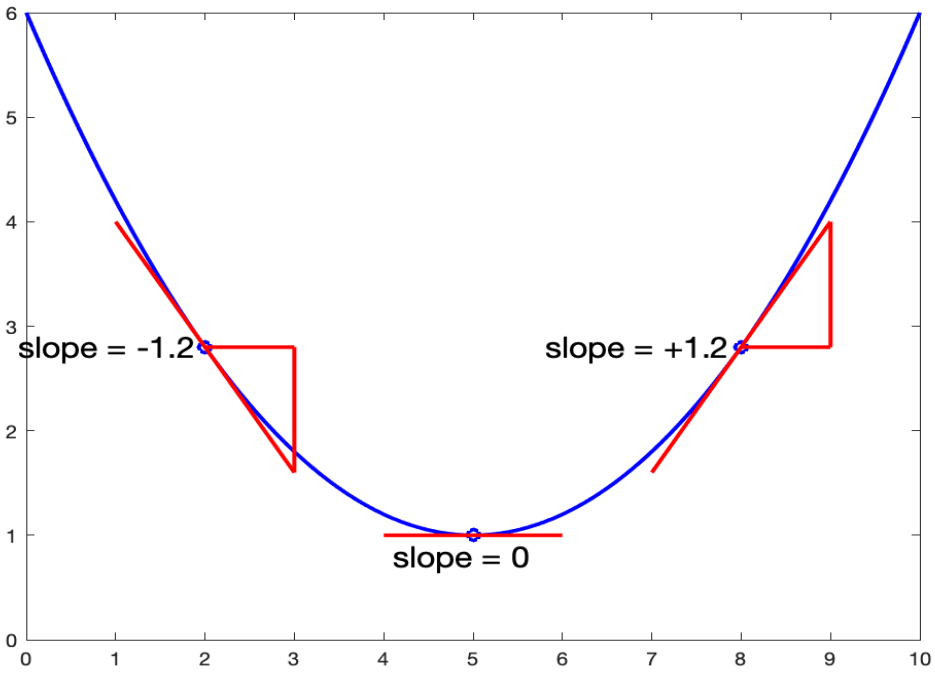


Figure 6.7: The slope of a curve at three points. The slope of the red line that just grazes the curve at a point is indicated in text, and is the rise of the line divided by the change in position of the point.

divided to the position difference isn't correct to about three significant digits, we know that we are in trouble.

We now need the concept of a curve whose value, at any point, is the slope of a given curve at that point. Figure 6.8 allows us to examine what the slope curve of the correct answer surprisal curve in Figure 6.4 looks like. The surprisal curve for question 46 is shown in the top panel and the surprisal-slope curve is in the bottom panel computed in two ways. The solid blue line is the completely accurate version computed using calculus. The red circles are the result of computing the differences between 81 consecutive equally spaced values, and then dividing these by the intervals between the points, namely one. We see that the two slope curves line up perfectly at the level of visual inspection. The largest difference between the true and approximate values was 0.007, or 0.7% of the true value. We just don't notice differences this small.[9]

[9]You might wonder how big a difference in magnitude has to be before we can notice it. Research the area of psychology called *psychophysics* reveals that the difference has to be around 5% of the magnitude, and this holds over a wide range of magnitudes and a wide range of magnitude values.

The correct answer surprisal in the top panel of Figure 6.4 decreases between 0% and 15% and again between 90% and 100%, and its slope curve in the bottom panel is correspondingly negative. In between, the surprisal curve is quite flat and and surprisal-slope near zero.

6.8 Answer Sensitivity

This section is about the easiest of all transformations. We define the *answer sensitivity* as the negative of the surprisal-slope function:

$$sensitivity(score) = -surprisal\,slope(score).$$

You might wonder, "Why bother with the negative sign?" The reason is purely cosmetic. The curve for probability for the right answer usually increases, which seems to feel good. Surprisal for the right answer decreases, and consequently surprisal-slope for the right answer is mostly negative. Which seems not quite right. So we cure the problem by just flipping the sign. When we get to Chapter 7, where we see how we define an optimal score index value, we will see that this sign change has no effect on the result. But we will see that a larger positive sensitivity value for a chosen answer suggests that the estimated optimal test score should be increased from its current value. Negative values of course have the opposite effect, and near zero values imply that knowing the choice is neither here nor there as far as the test score is concerned.

Figure 6.9 shows the probability and sensitivity curves for each of the four answers for question 55, presented in Section 4.6. The right answer curve in blue in the bottom plot shows that this question is most informative for test takers near the 25% dashed line, where it is strongly positive and the corresponding probability is rapidly increasing. At that point, a choice of the right answer would be a performance score of 0.1. But a choice any of the three wrong answers would hardly change the score at all.

However, knowing that the answer is correct is nearly worthless for test takers at the top 75% dashed line, where the sensitivity curve is virtually zero. This is because the corresponding probability values for answers 1 and 4 are hardly changing at all over the top interval. But we also note that a choice of wrong answers 2 and 3 would, at that position, apply a strong penalty to a score.

Why, we might say, would choosing wrong answer 3, $f(x) = 2$, actually benefit a test taker at the topmost score level? We can imagine that really bright persons resolve the ambiguity of this question by assuming that the question calls for a single numerical answer rather than the behaviour of the function all the way across the range of x.

Note, too, that wrong answer 1 prevents many top level test takers from obtaining perfect test scores. Its sensitivity is nearly zero everywhere. As a consequence, the choice of answer 1 will be virtually ignored by the scoring process that we describe in the next chapter.

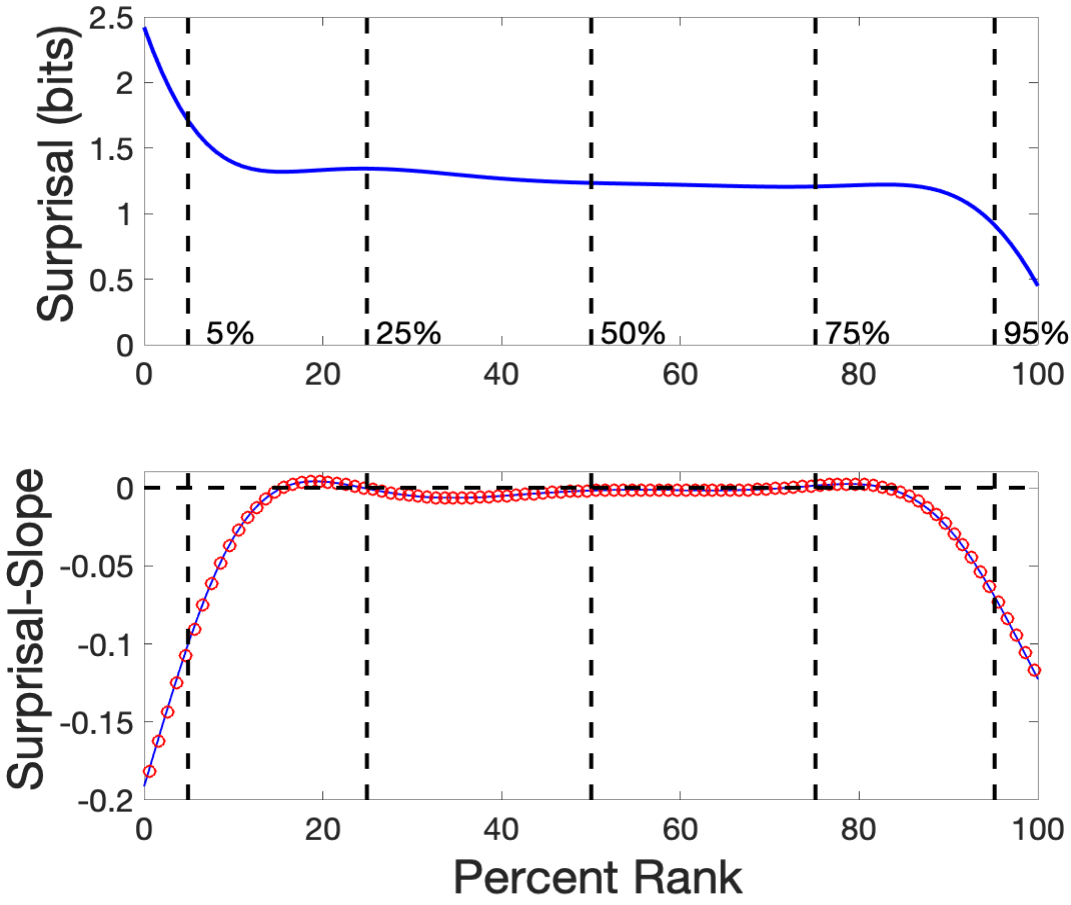


Figure 6.8: The surprisal (upper panel) and the surprisal-slope curve (lower panel) for the correct answer for question 46, shown in Figure 6.4), in the SweSAT-Q. The thin solid line in the lower panel is the highly accurate curve computed by calculus, and the red circles are the approximate curve values by computing rises over runs.

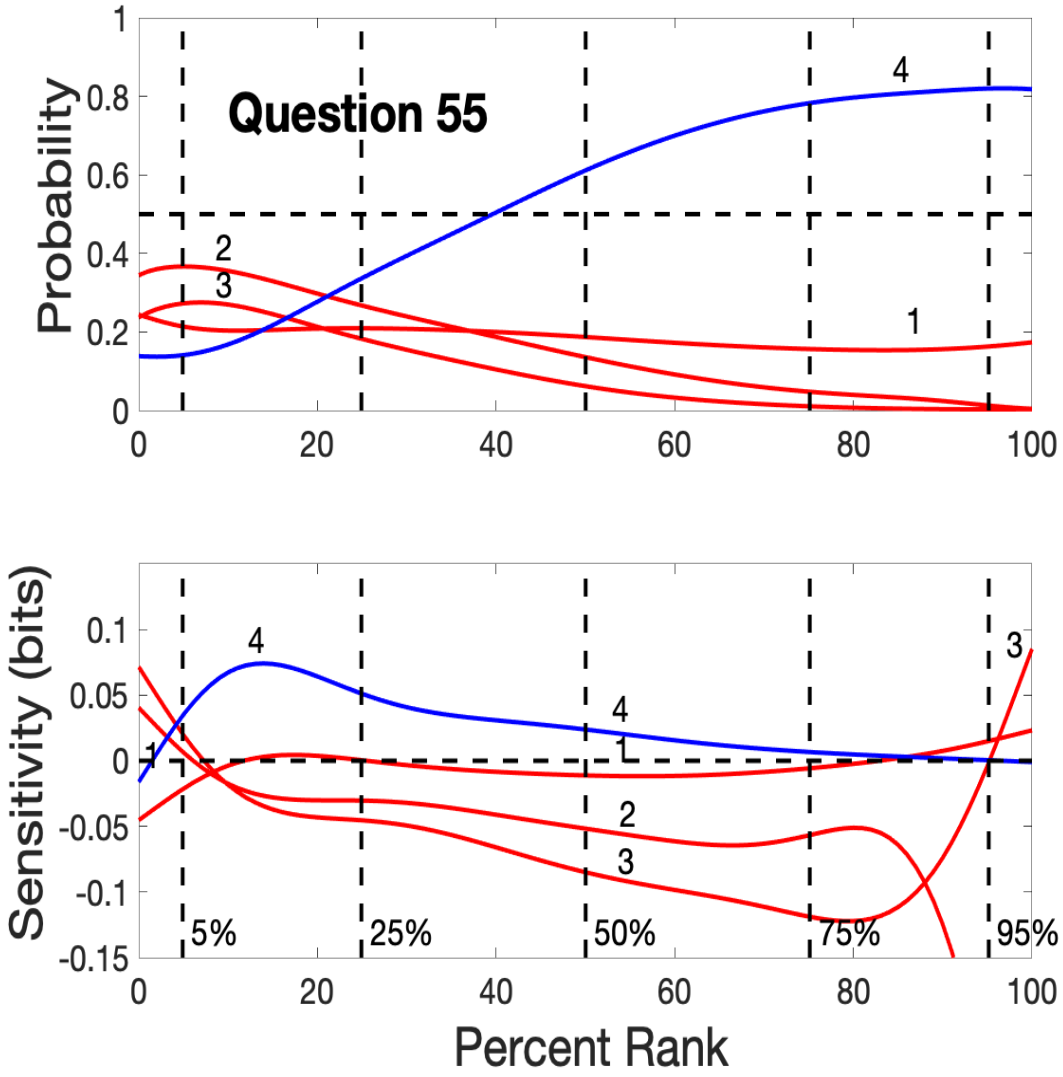


Figure 6.9: The full data probability (upper panel) and sensitivity curves (lower panel) for question 55 in the SweSAT-Q.

We now possess a way of constructing scores for all answers in the test for test takers that are the counterparts of the weight increases in the weight lifting example or the gear shifting in the transmission example. You might complain that we only had one weight lifter, but if a bus had pulled up to the gym and provided forty or so potential weight lifters with widely varying proficiency and physical fitness, the trainer would have varied his weight ranges, and we could have produced plots that were roughly like Figure 6.9.

Chapter 7

Test Surprisal and Sensitivity

7.1 Introduction

We now have a gold plated shovel for mining test data! We can add surprisal values and entire surprisal curves to combine data across test questions and even examinees. We can also take differences secure in the knowledge that a difference means the same thing everywhere on the surprisal continuum, which starts at zero, has a unit quantity and is unbounded above. Think of surprisal as the heat of knowledge. If you know something already, it won't change anything, but once in a while you'll be blasted out of your seat.

Now we put surprisal (or information) to work in order to combine surprisal curves across questions in order to define a surprisal curve that characterizes a test taker's performance on the entire test. And we use the combined surprisal slope or sensitivity curves, these being effectively differences, to find that performance level that best describes the choices that a test taker has made.

We'll stick with the bit as a unit, but it can be useful to use something else. The bit describes choices between two alternatives; but, for instruments like the Symptom Distress Scale where questions are like throws of five-sided dice, the unit could logically be the surprisal associated with probability 1/5, which is about 2 and 1/3 bits.

In this chapter we revert to using a score index distribution that ranges from 0 to 80 and roughly is that of the sum score distribution. But since test scores remain the same if we change score index systems, no harm is done. In the next chapter we will use a score index system that has itself the unit of a bit, and will re-display some of the plots that we display here.

7.2 Surprisal Curves for Test Takers

Now we can use the surprisal curves for the answers that a test taker has chosen in order to assess the total surprisal of the test taker's answer choices. We do this by

adding up the surprisal curves for the test taker's chosen answers. Here is the word formula, defined in terms of the total surprisal at a specified score index value.

$$test\ taker's\ surprisal\ curve(score\ index) =$$

$$sum\ of\ surprisals\ of chosen\ answers(score\ index)$$

Since are allowed to add surprisal values, the test surprisal curve has the same unit, the bit, as the curves which are summed.

In order to illustrate test taker surprisal curves, we chose five test takers at random who all have a sum score of 35, the median score on the SweSAT-Q subtest. Of course we don't believe that, just because they have the same sum score, their true performance levels are the same. But their performance similarity will be sufficient to allow to see what a variety of test taker curves look like.

Figure 7.1 shows all five test taker curves. For each curve, the location of the minimum surprisal value is the least surprising for that test taker's data. This minimum score index value is called the *least surprisal score index*, and is indicated in the plot as a vertical dashed line of the same colour as the corresponding curve. At that value, the test taker's choices for the whole test is as consistent with the test model as possible. The "least surprise" principle for defining the best estimate for a model is considered something of a gold standard in statistics. Score indices below the least surprise locations are considered to be under-estimates of performance, and those above over-estimates. The rapid rise on the right side of each curve provides unambiguous evidence that none of the test takers deserves a score index value above 50 or so. The shallow slopes on the left side of the minimum surprisal value are a little less clear cut, but nevertheless suggest that all minimal surprisal score values are not too far from the sum score of 35.

The values of the test surprisal curves at their respective minimum locations also vary. The yellow curve with a value of about 125 bits suggest that this test taker's model fits the data better than the others. The blue curve, on the other hand, with a minimum value of about 175 bits, also suggests that there is another much lower minimum location. We can interpret this as indicating that the test taker has a much lower skill level for certain kinds of questions. For example, perhaps this person finds algebra and symbolic expressions impossible to work with, and is essentially guessing for questions like these.

Figure 7.2 shows the surprisal curves for five randomly selected test takers with sum scores of only 18, the marker score for the bottom 5%. Now we see least surprisal locations mostly below that of the sum score. This happens because at that poor performance level the data suggest that many of the 18 right answers were just lucky guesses, and probably because they were for questions that were clearly well beyond their performance range. We shall see that the sum score is a *biassed* estimate of poor performances in the sense of systematically being above what it should be. Note also that the surprisal values at the minimal surprisal locations are substantially higher

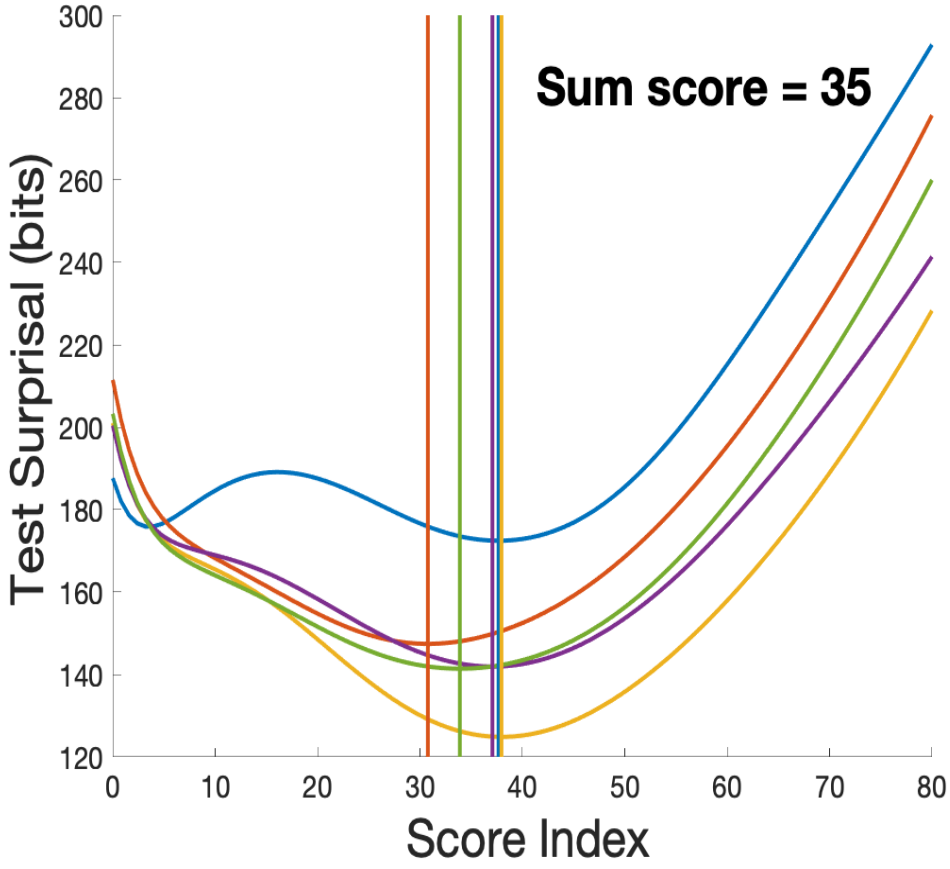


Figure 7.1: The test taker surprisal curves five test takers with sum scores of 35 on the SweSAT-Q. Each vertical line is positioned at the minimum value of the surprisal curve of the same colour.

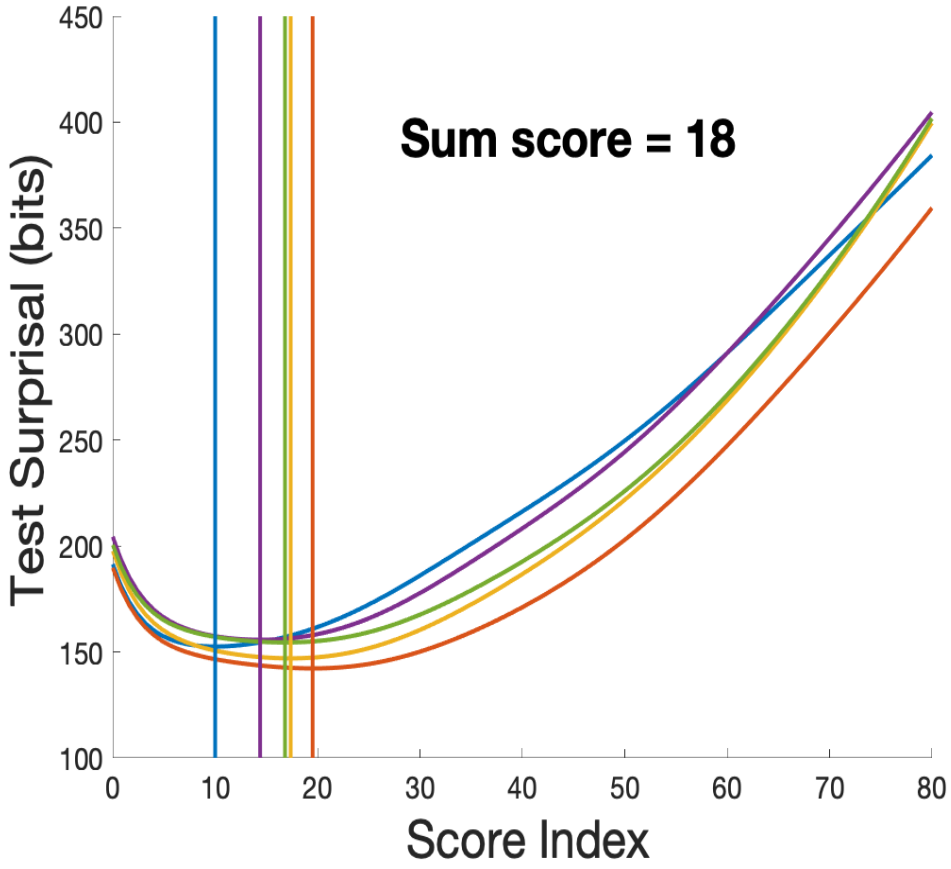


Figure 7.2: The test taker surprisal curves five test takers with sum scores of 18 on the SweSAT-Q.

than those for sum scores of 35. This is because the less the knowledge level, the more the impact of guessing, which is a chaotic process. At these low levels all curves are surprising with respect to the model.

What about five high flyers? Figure 7.3 shows surprisal curves for five test takers with sum scores of 72. Four of the least surprise locations are well above the sum score and the other is close to it. The score estimation considers these test takers to be possible victims of questions like 39 and 55, and views the sum score of 72 as tending to be an under-estimate of their performances. That is, it appears that the sum score tends to be biassed *against* high performance test takers. The minimum surprisal values are also small because at this high performance level most questions are answered correctly and there is relatively little variation in answer choices as a consequence. The model predicts that they will get the questions right, and they do so.

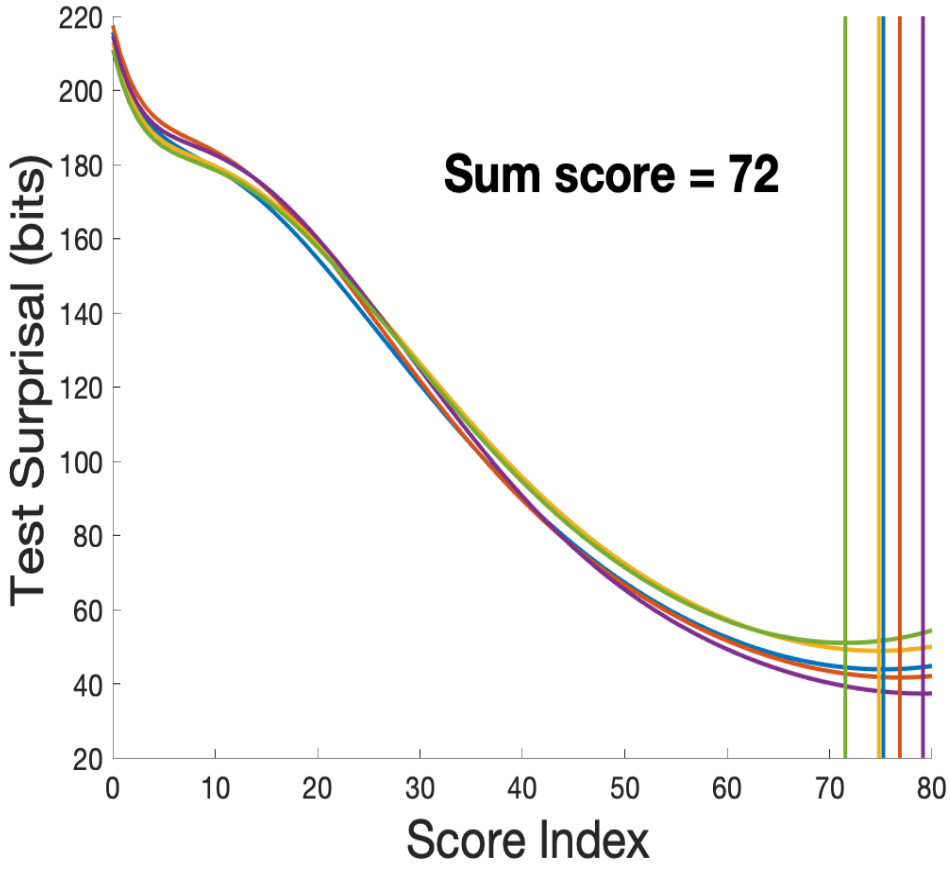


Figure 7.3: The test taker surprisal curves five test takers with sum scores of 72 on the SweSAT-Q.

7.3 Sensitivity Curves for Test Takers

We now combine the sensitivity curves for the chosen answers in the same way that we combined surprisal curves in Section 7.2. A test taker's *test sensitivity curve* is sum of the answer sensitivity curves for the answers chosen by the test taker. In a word equation,

$$test\ sensitivity\ curve(score\ index) = sum\ of\ chosen\ sensitivities(score\ index)$$

This curve is designed to show the total of the chosen answer sensitivity scores for the entire test at any possible score index value, or, if divided by the number of questions, the average chosen-answer sensitivity.

The sensitivity curve is constructed from the surprisal slope curve, and therefore the slope of the test taker's test surprisal curve at its minimum is zero. Thus, we now want to find the point on the score index continuum where the value of the test sensitivity curve crosses zero. At that point, the test designing team and the test taker are evenly matched and the test taker is losing as much as winning. The total value of the positive sensitivities is equal to the total value of the negative sensitivities. Any further increase in the score index is going to over-estimate the test taker's performance level, and any decrease will be an under-estimate. Our best indicator of the test taker's performance is therefore that score index value at which

$$test\ taker's\ score\ index = where(test\ sensitivity\ curve = 0).^{1}$$

Figure 7.4 displays the test sensitivity curves for the same randomly selected test takers that we selected in Section 7.2, all having sum scores of 35. The locations of the points at which these curves cross the horizontal axis at 0 bits are the same as we saw for the minimum surprisal values above.

7.4 The weight lifting and cycling equilibrium points

In the weight lifting example the sensitivity value for a weight lifting task is the difference between current and previous weights multiplied by one if the new weight was successfully lifted, and by minus one if not. But this only applies to the weights within the range selected by the trainer. Weights either above or below this range

[1] In statistics, this is an application of the *maximum likelihood principle*, but which can be transferred into our own jargon as "minimizing the surprisal of the data." Statistical theory shows that these principles define what has come to be the gold standard for estimation of unknown parameter values. Maximum likelihood or minimal surprisal estimation is usually the statistician's first choice when confronted with a new data structure. Here's how it works. Let θ denote performance level. Let U_{ji} be the index of the answer to question i by test taker j, and let $W_{i,U_{ji}}(\theta)$ denote surprisal curve values, for question i at performance level θ. The negative log likelihood or data surprisal is $\sum_i W_{i,U_{ji}}(\theta)$. Consequently the minimum of data surprisal occurs when $\sum_i dW_{i,U_{ji}}/d\theta = 0$.

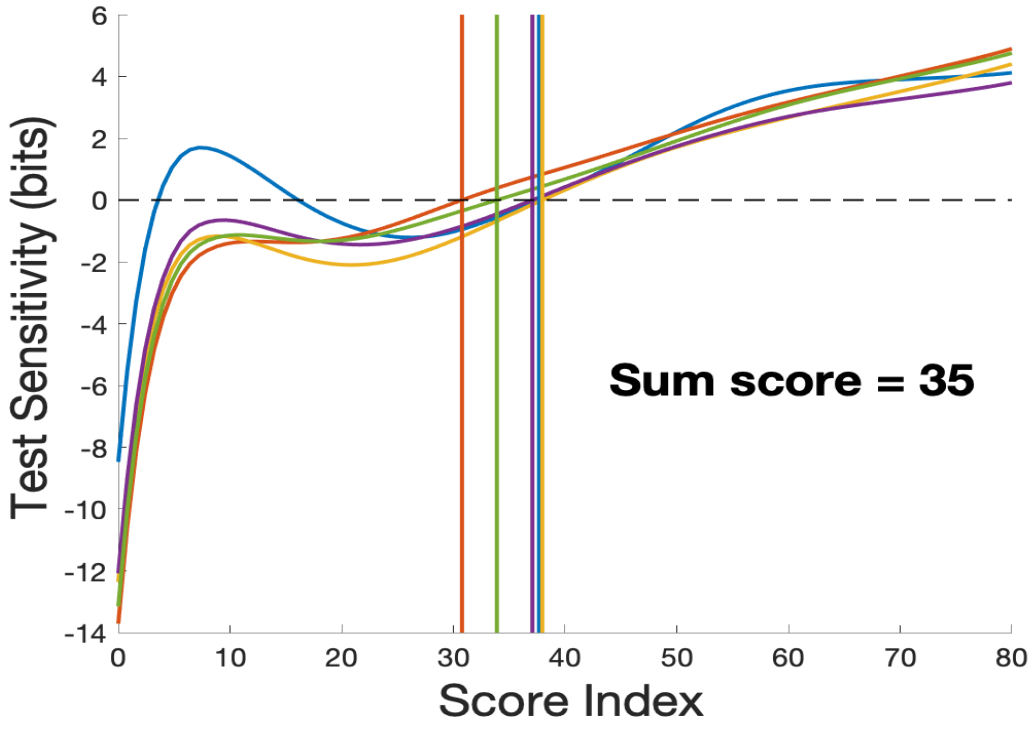


Figure 7.4: The five curves are the test sensitivity curves for five randomly chosen test takers with SweSAT-Q sum scores of 35. The solid vertical lines locate the points where the curves cross zero and define these test takers' score index values.

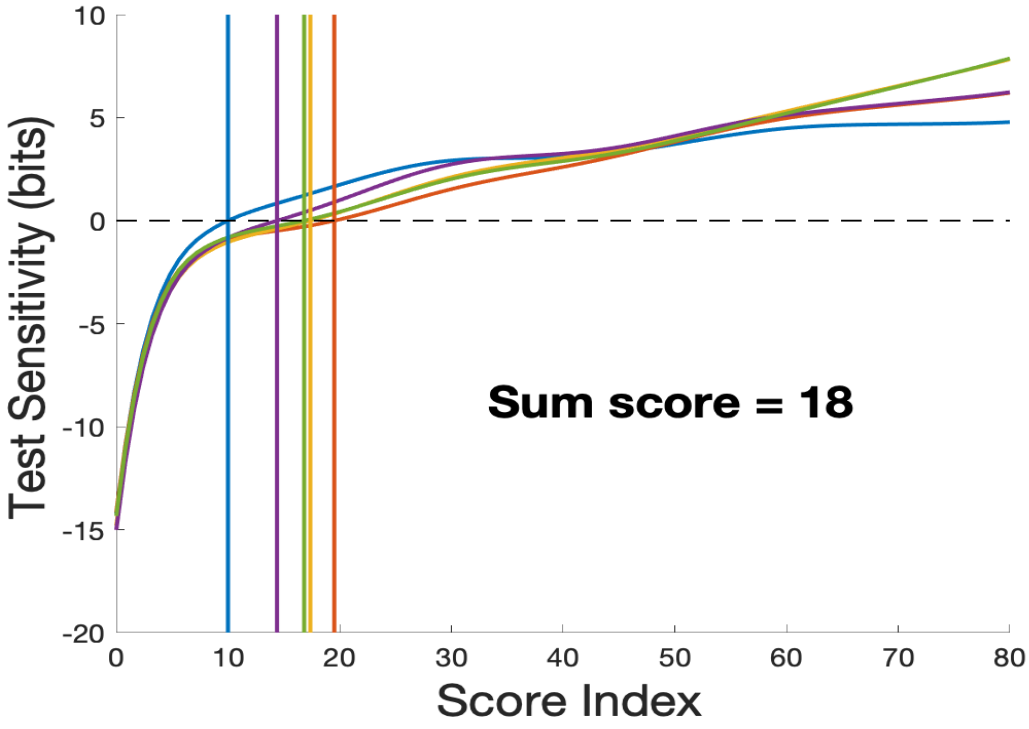


Figure 7.5: The five curves are the test sensitivity curves for five randomly chosen test takers with SweSAT-Q sum scores of 18.

were in effect given zero sensitivity values because the trainer judged them to convey no useful information about the weight lifter's performance.

The same may be said of the gear selections of a cyclist, who only chooses gear ratios within that prove to be useful. Outside of that range, gear ratios are ignored, and by implication are zero.

What we have achieved at this point are specification of the weights and gear ratios for the possible answers to each question. We've used the data provided by the whole cohort of test takers to do this, and an important step was to turn probability into surprisal, which has the essential property of having meaningful sums and differences. The slope of a surprisal curve at a particular point is essentially a difference, except that the actual difference is divided by a small constant that does not change over either test takers or score index values.

There is, however, a feature separating weight lifters and cyclists from test answers. The former knew the sensitivity values for their problems ahead of time, but we used the power of a statistical model combined with the speed of a computer to work out sensitivities using the data from a test administration. If this sounds like weight lifting by using your own bootstraps, you are close to the truth, except that the bootstraps of 53,000 Swedes were required.

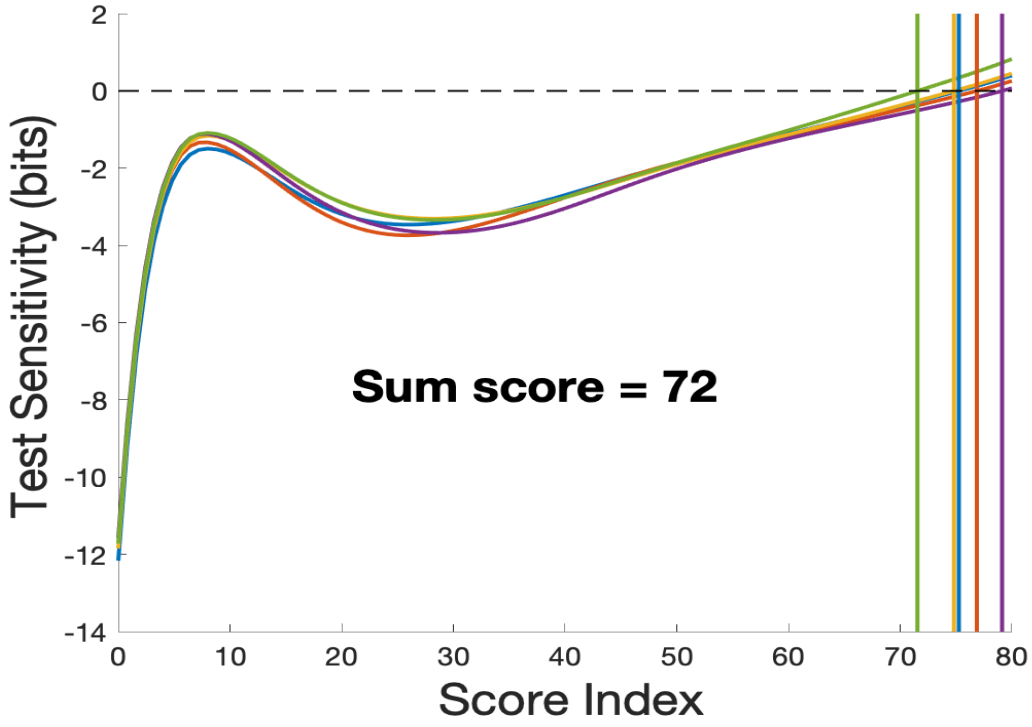


Figure 7.6: The five curves are the test sensitivity curves for five randomly chosen test takers with SweSAT-Q sum scores of 72.

Moreover, the weight lifting and cycling examples concerned performance assessment for a single person. In our context, we have to estimate the performances of over 53,000 takers of the SweSAT simultaneously. This is the reason why we need test sensitivity curves instead of test sensitivity numbers. We have to be able, for each test taker, to search along that person's test sensitivity curve to find the score index at which the test sensitivity crosses the zero line for that person. Happily, mathematicians and computer scientists have evolved really fast and highly reliable methods for locating such a point, and our computer program for this task can handle the entire set of SweSAT takers in a bit over a minute.

Chapter 8

Score Indices, Test Scores and Test Effort

8.1 Introduction

In Chapter 4 we detailed how we assign a score index value to test taker. This involved either finding the test taker's minimum test surprisal value; or, equivalently, the score index value at which the test sensitivity curves is equal to zero. Now we can display the distribution of the score index values for any score indexing system that we choose to use.

Let's first review how we designed a score indexing system back in Chapter 4. There we started off using sum score values in Figure 4.2 to display the proportions of test takers getting question 46 on the SweSAT-Q. In this plot, we effectively used the 81 possible sum score values as a score indexing system. Then we sorted the actual sum scores into bins, each containing roughly 1000 test takers, and computed the proportions of test takers in the 80 bins who chose the correct answers for question 46. In Figure 4.3 we plotted these proportions, and in this plot we switched to using bin centres as a score indexing system. Next, in Figure **??**, we passed a nice smooth curve through the proportions. This was an important step since we both possible sum score values and bin centres values are discrete values, but with the curve in place, we could replace these by a *continuum* for score values from which we could select any value between 0 and 80, 80 being the number of SweSAT-Q questions. This was, finally, our first operational score indexing system. But we pointed out that any smooth transformation of this system that preserved the order of the index values was also legitimate as an alternative score indexing system. We proposed, for example, that would turn score values running from 0 to 80 to percent ranks running from 0 to 100. We found this useful in Chapter 5 for exploring the performance of a wide variety of questions drawn from three different tests and a scale.

With a test taker's score index in hand, we can compute the probabilities that this test taker will choose an answer. We saw in Chapter 2 how these probabilities are

used to define the *average test score*. In this chapter show that test scores can fail to reflect the true performances of test takers in both the highest and the lowest ranges. In the next chapter, we will also show that the average test score is far more accurate than the sum score itself.

Test scores are, we recall, defined by what numerical weights test designers attach to each answer for each question. For multiple choice style tests, these weights are merely one for a correct answer and 0 for the test, and this is what defines the sum score that we are trying to improve upon. We saw in Chapters 4 and 5 and will see again in this chapter that these weights can seriously bias either a straightforward sum score or an average sum score with respect the actual performance level a test taker, especially for the two performance extremes. Can we do something about this bias?

There is a unique score index system that is free of test designer impacts, and we call it the *test effort* index, that combines the desirable properties of both test scores and score indices. We can add and subtract designer-defined test scores simply because they are already sums. It turns out that we can also add and subtract at will the test effort scores, since they are magnitudes. These scores may be expressed in bits, or percents, or any numerical system produced by multiplying a test effort score by a positive constant.

8.2 Score Index and Test Score Behaviour

We've learned some essential things at this point. First, that there is an essential difference between a score index and a test score. The score index is not in general a measure of performance on the test, but rather a system for selecting test scores. But a score index system does imply something important about test scores, namely that they evolve smoothly. The score index is also a line segment having lower and upper limits. But we have huge latitude in specifying a score index, since any smooth transformation of a score index that preserves the order if any pair of points is also a valid score index. Three examples of such transformations from score index x to score index y are:

linear transform: $y = ax + b$ provided $a > 0$.

power transform: $y = x^p$ provided that $x \geq 0$ and power $p > 0$.

exponential transform: $y = C^x$ provided that base $C > 0$.

More complex transformations are used in the test equating process that we described in Chapter **??**.

A test score, on the other hand, usually involves a specification by a test designer that it will be a sum of scores that the designer has assigned to the chosen answers, and therefore is entirely defined by the these answer scores. The test score does

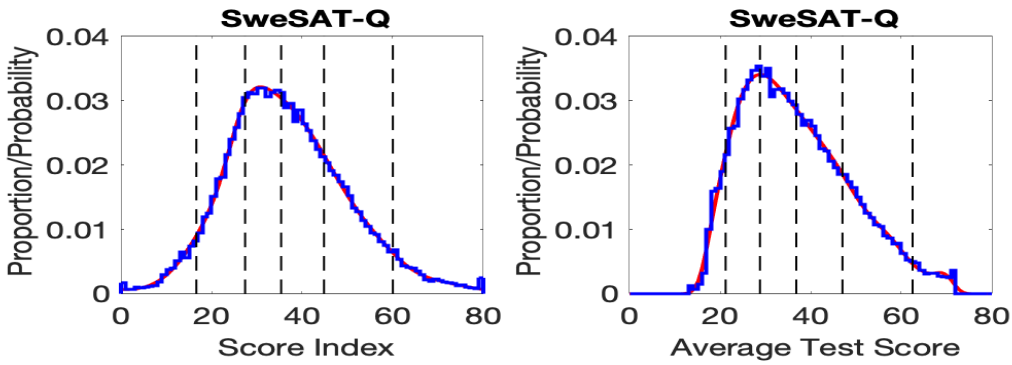


Figure 8.1: The left panel shows the distribution of the score index values for the SweSAT-Q, and the right panel shows the corresponding average test scores.

not change if we change the score index to a new score index. This is a valuable feature because it permits the test designer to revise answer weights for questions like SweSAT-Q 39 and 55 after the test has been administered and the data analyzed and it has been discovered that these questions do not work as intended. Alternatively, it may be that the test administrators like to use more than one score indexing system, and they can happily do so because no matter which system is used, the test scores will not change.

These two mathematical structures, the score index and the test score, can behave quite differently. Figure 8.1 displays the distribution that values of our first score index running from 0 to 80 in the left panel, and the distribution of the values of the of the test scores that these score indices select in the right panel for the SweSAT-Q. What stands out is that the score index values have pile-ups of value at both extremes, indicating that there are a noticeable number of test takers who are either beneath or above the test in terms of performance. The average test score distributions shows exactly the opposite, namely that are no test takers scoring below about 18 or above 70. We think that the score index tells the right story and is more plausible, and we think that that test score heavily penalizes high performance test takers because of questions like 39 and 55, and that the impact of guessing on test scores conveys a too-benign view of the abilities of low-performance test takers.

In Figure 8.2 we switch from probability to surprisal in order to see more clearly how the simple sum score, the average test score and a score index behave for the SweSAT-Q. The differences between surprisal values in the middle are minor, but not

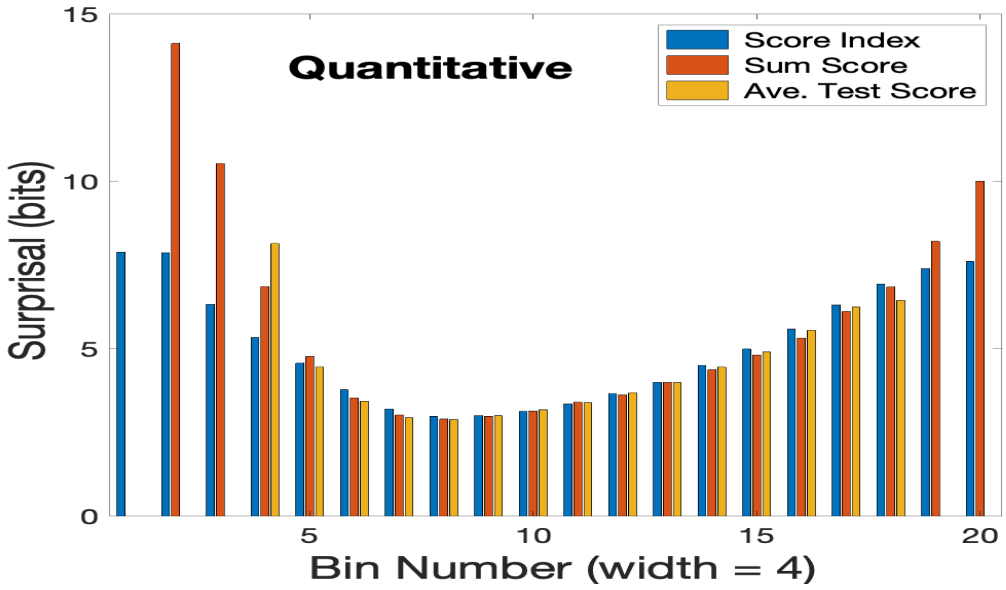


Figure 8.2: Each of the three bars corresponds to the surprisal of either a test score or a score index value falling into one of 20 equal sized bins for the SweSAT-Q. The bars that are not displayed in bins 1, 2, 3, 19 and 20 correspond to proportions of zero that correspond to infinitely large bars.

so at the extremes. We see at the highest bin 20 that sum scores are about two bits more surprising than score index values, suggesting that sum test scores in that bin are much more rare than score index values. The situation is even more dramatic for average sum test scores, where the sum is over the products of the chosen answer scores and the probabilities of choice. We see that this type of score never appears in the top two bins.

If you were a top performing test taker in bin 20, you would very much prefer to be assessed in terms of the score index rather than either test score. If we were you, we would claim that your less-than-perfect sum or average sum test score reflects the harm done by questions like 39 and 55, and that in fact, if these questions not in the test, your score would indicate that you know enough to be classed as outside of the range of the test. We may say the same even more forcefully at the bottom bin 1, where the score index effectively corrects for the benefit of guessing and assigns many test takers to being below the test, even though their sum test scores can be nonzero. We even see this down-grading by the score index in bins 2 and 3. Of course a test taker in this zone would prefer a test score since, for the average sum test score for example, this will be at least 13. But someone evaluating a test taker for a job or admission to further education would rather look at more clear-cut evidence virtually no knowledge of the topic provided by the score index.

In other words, we would argue that the multiple choice format has the disadvan-

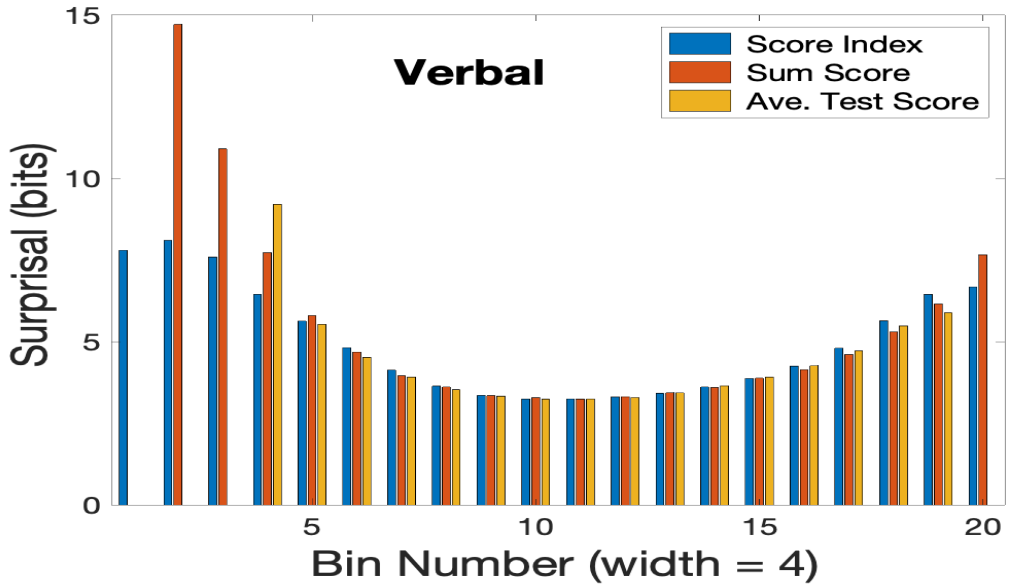


Figure 8.3: Each of the three bars corresponds to the surprisal of either a test score or a score index value falling into one of 20 equal sized bins for the SweSAT-V. The bars that are not displayed in bins 1, 2, 3, 19 and 20 correspond to proportions of zero that correspond to infinitely large bars.

tages of (1) contaminating the data by the possibility of guessing the correct answer and (2) assigning the overly simple scores of 1 and 0 to correct and incorrect answers, respectively. We find that the results for SweSAT-V in Figure 8.3 confirm these conclusions. Moreover, Figure 8.4 shows, for the National Mathematics test where guessing is much less of a factor, the test scores tell much the same story as the score index, but that at the highest performance range the score index remains the test taker's preferred option.

In the remaining sections we show that there does exist a score index that is unique and that can also be used as a measure of test performance.

8.3 Test Effort: The Test as a Ruler

Now we introduce a special score index that measures the amount of information tested in bits, and thereby allows us to place each test taker along a line that defines how much of the information in a test that person commands. This, you might say, is what you expect the test score to do. But the test score does not do this because a fixed increase in a test score does not mean the same thing at all test score values. In fact, we have seen that an improvement of a test score of, say, one, when the test taker is in the middle of the distribution means far less than an improvement

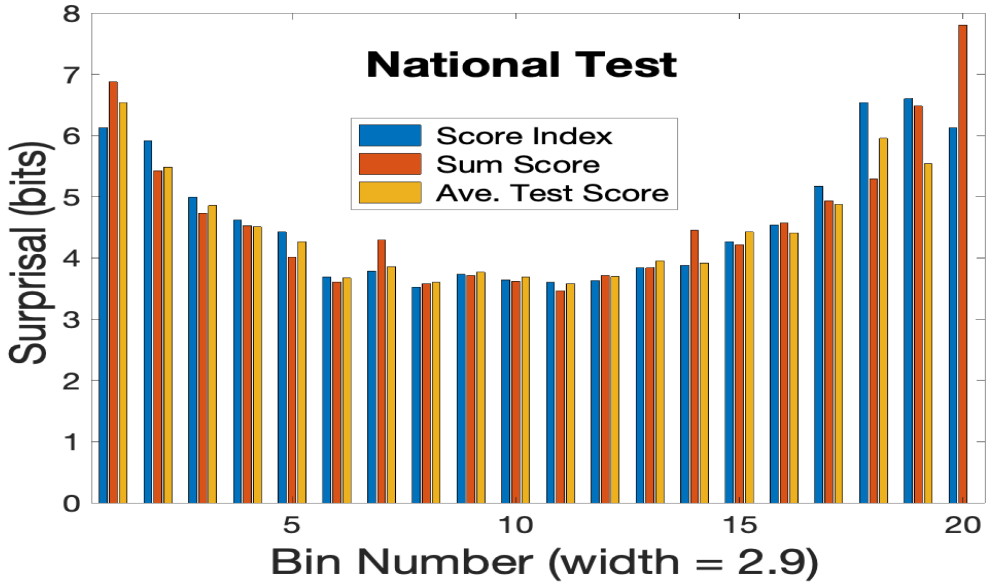


Figure 8.4: Each of the three bars corresponds to the surprisal of either a test score or a score index value falling into one of 20 equal sized bins for the National Mathematics test. The bar that are not displayed in bins 20 corresponds to proportions of zero that correspond to infinitely large bars.

by one at the 95% marker score. The latter requires, as we will show, vastly more studying, practicing, dedication, money, self-denial and other aspects of effort. Not to mention the complete impossibility of obtaining an average sum score, for the SweSAT subtests, equal to the number of test items. Moreover, the test score is defined by the pre-allocated weights placed on answers, whereas this uniquely defined score index that we now take up is unaffected by any change in the test scoring protocol.

8.3.1 A 3D Probability Plot of a Three-question Binary Test

We researchers in the mathematical and information processing sciences know that optimism is one of our worst enemies, often leading us into tackling a large-scale complex problem before we've completely understood the basics. Sound familiar?

So let's design a little SweSAT-Q test with only three questions. After a careful review of the 80 questions in the SweSAT-Q, we chose questions 43, 46 and 71 as good representatives of three respective difficulties. These questions are easy, mid-difficulty and hard, respectively. In this way we span the performance range, and we can then see something in a three-dimensional plot that has escaped us in the preceding two-dimensional plots.

We've already looked at the question profile for question 46 in Figure 6.4, so we first display the profiles for the other two questions. Question 43 displays these two

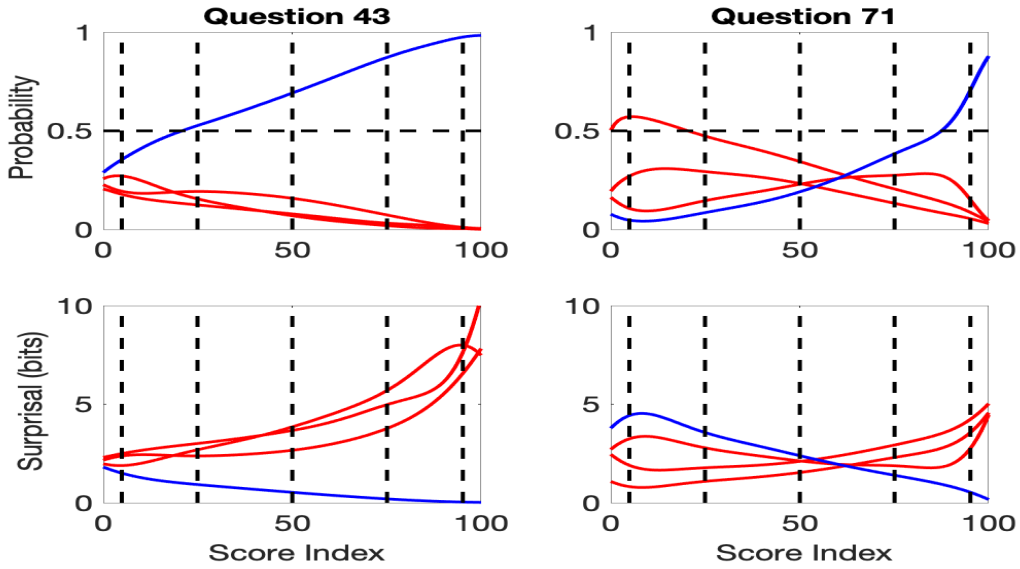


Figure 8.5: The probability and surprisal curves for the answers to questions 43 and 71 from SweSAT-Q. The correct answers are in blue and the incorrect in red.

mathematical expressions: I $x(y+z)+x^2+yz$ and II $(x+z)(y+x)$. The answers are: *(1) I is larger than II, (2) II is larger than I, (3) I is equal to II and (4) Information is insufficient.* The first answer is correct. Question 71 requires extracting information from charts that we don't bother reproducing here. The four panels of Figure 8.5 present the probability and surprisal functions for questions 43 and 71, and those for question 46 are shown in Figure 4.3.

What we estimate are the three probabilities at any specified test performance level of getting the questions right. We can plot these three probabilities as a single point in a graph with three axes, each ranging from 0 to 1. In Figure 8.6 we plot these probability triples as they vary over the common test score range of 0 to 3. Now you can't see the score index variable explicitly, but you might imagine that it is spread out along the curve starting from near the lower corner (0,0,0) and ending up close to the perfect score corner (1,1,1). The progress of the score index is marked out by the points on the curve that are positioned at the five marker percentages. The total length of the curve is 1.2 if we were to pull it out to a straight line. We call this the length of the curve *taken along the curve* or, in math speak, *arc length.*

The lowest level test takers make forward progress along curve mostly with respect to question 43 while maintaining a low performance on the other two. Then, between the 5% and 50% markers, the curve turns left and the performance on question 46 improves. Beyond 50% the curve rises off of the lower plane as test takers reach a level that permits them to handle question 71. The distance along the curve between the 5% level and the 25% level is small compared with the effort required to travel from the 75% level to the high-flying 95% mark. That is, learning how to answer

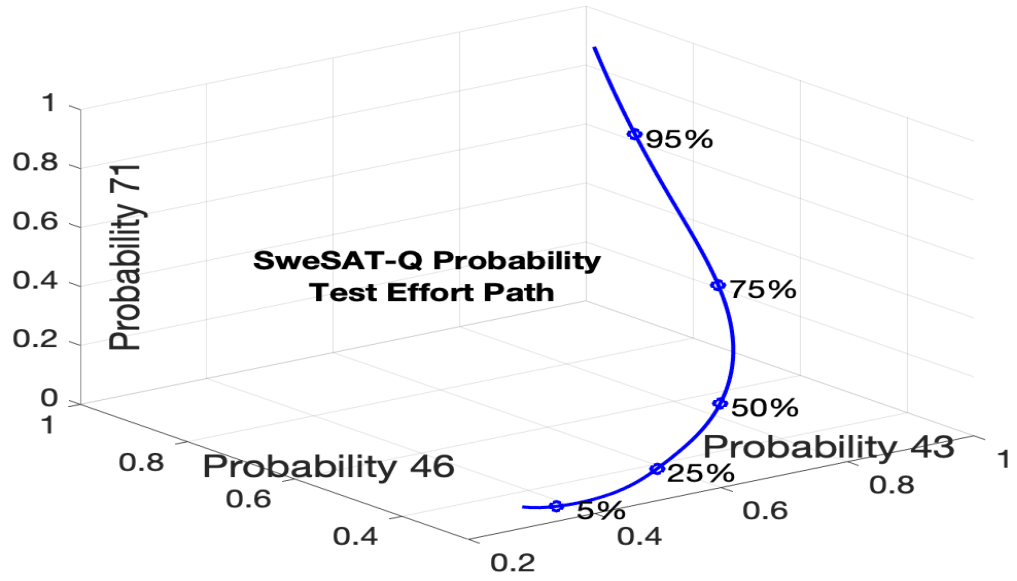


Figure 8.6: The solid curve shows the joint variation in probability of getting the three SweSAT-Q test questions 43, 46 and 71 correct as the score index passes from its minimum to its maximum. The dots show the score values corresponding to the five marker percentages. The questions are easy, medium and hard, respectively.

easy questions is easy, but getting hard questions right is much harder. At the 50% marker, the median test taker score index, one notes that there is a fairly solid success with questions at the level of question 43, but progress on questions at the level of 46 has not yet been realized.

A metaphor comes to mind. We fancy an executive jet gathering speed along the runway on the lower plane, rising and banking left to clear the clutter on the ground and soaring toward mastery of high school math, leaving even question 71 behind. The jet's destination? Well, perhaps a university degree in mathematical statistics.

Because a curve position depends only on three probabilities, and probabilities do not change if we change the score indexing system, this shape of this curve does not depend on which score index system we use. Points on the curve therefore behave like a test scores, which also do not change with changes in score index systems. But, in contrast to the test scores, this curve does not depend on what some test designer thinks a right answer is worth.

8.3.2 A 3D Surprisal Plot of the Three-question Binary Test

We can also view in Figure 8.7 the curved path in three dimensional space as defined by the evolution of the surprisal curves. Since surprisal decreases as probability

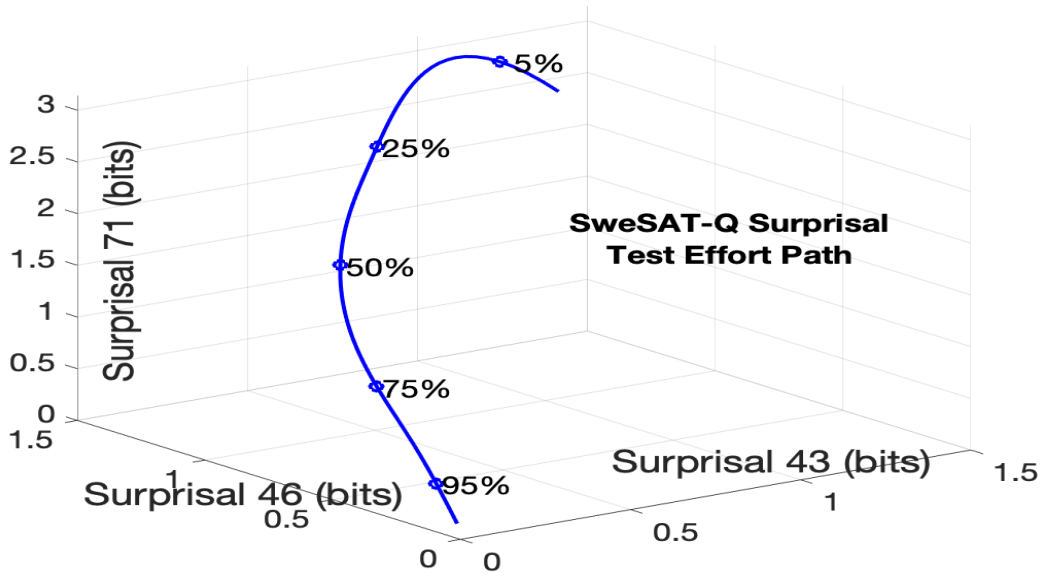


Figure 8.7: The solid curve shows the joint variation in surprisal of getting SweSAT-Q questions 43, 46 and 71 correct as the score index passes from its minimum to its maximum.

increases, we see the progress along the curve running in the reverse direction. Noting that the vertical axis varies from 0 to 3.5 while the other axes vary only from 0 to 1, we still see that there is a longer trajectory from 75% to 95% than for its lower end counterpart.

But now, and since surprisal has a unit, the bit, and can be added and subtracted, we can actually *measure* the length of the curve in the same way that we can measure distance along a railway track or the trajectory of a rocket. And so, this curve is 3.5 *bits* in length, and we can measure in bits the distance between, say points for two marker percentages. We can, for example, say that motion along the curve is at the average speed of 1 1/6 bits per question. Or we can ask questions like, "How much better is my performance if I get one bit further along the curve?" "How many more bits do I need to pass this course?" "How many bits away from me is that test taker over there that I'd like to get to know?" We can answer questions like these because distances can be added and subtracted, just like those for any other magnitude.

8.4 Test Effort: Curve Features

The three-dimensional curves in Figures 8.6 and 8.7 offer a unique and privileged indexing system, that we call the *test effort index*. *Test effort* is the distance travelled along the curves in these figures from their beginning to a particular point on the

curve. Since every test taker is located somewhere along this curve, we can think of this curve as representing whatever the test taker must invest to reach the test taker's position. Perhaps it is number of hours spent in the classroom and studying. Perhaps it might reflect the money invested in acquiring this level of knowledge. Or the amount of pleasure foregone. Whatever it means, the model says that everyone has to go along the same path. We can debate whether this is so, but the data that we work with in this provide pretty impressive support for that proposition.

How can we work with this curve when 80 questions are involved, and for each question there are multiple answers? It seems impossible to visual such a thing. We especially want to study the very high dimensional surprisal version of the curve in Figure 8.7. Distance along that curve will be measured in bits and this will open a rich range of possibilities for measuring aspects of the test and what happens to test takers as they progress along the test effort curve.

8.4.1 Measuring Distance along the Test Effort Curve

Measuring distance along the curve in bits turns out to be quite easy. For any point on the curve, and for each answer, the sensitivity curve tells us how fast the surprisal associated with that answer is changing. Suppose we take a teeny step along the score index. Imagine doing this on a bicycle going up a road snaking up a mountain pass, for example. The steepness of the road multiplied the length of the step, or in the bicycle case the result of a single revolution of the pedals, tells us how much elevation we have gained. Or lost if we are going down the hill. We can break the elevation gained into two spatial directions, of course. In the same way, a small step, say 0.1 along the score index ranging from 0 to 80 for the SweSAT-Q, for a single question implies changes in surprisal for each of the question's answer curves. Each of these changes in surprisal value is computed by multiplying the curve sensitivity value for that answer at that point on the test effort curve by the size of the step along the score index.

If you're still hanging in here at this point, let's press on; but if not, just agree with us that we can measure the distance along the curve for a small step, and skip to the visualization of the curve in the next subsection.

Next we reach back into our high school geometry experience and recall that the distance between two points in any direction is equal to the square root of the sum of the squares of the changes in position in each coordinate direction. This is an application of Pythagorus' Theorem if that helps. Well, this *root-sum-of-squares* operation works no matter how many directions are involved. For example, there are no less than 412 answers in the SweSAT-Q subtest, each with its sensitivity value, and the computation of the distance along the curve will involve summing all the squared products of the sensitivity values and the step size, followed by taking the square root of the resulting sum.

The last step is, if we want to measure the length of the entire curve from start

to finish, adding up all the distances associated with the short steps. This is, for the SweSAT-Q, about 159.6789 bits or almost exactly two bits per question. Now if we want to know how many bits along the curve a specific test taker has gone, we just add the steps that take us to the test taker's score index value.

Figure 8.8 shows the relation between distance along the SweSAT-Q test effort curve and the score index ranging from 0 to 80. The relationship is not that far off being linear, but notice that the sharper slope at the beginning and the end of the curve means that the efforts required to get from 0 to the 25% point (about 58 bits) as well as from the 75% to the end (about 46 bits) are considerably longer than the score index indicates. And longer than the journey from 25% to 50% (25 bits) or from 50% to 75% (30bits). This is what we mean when we say that there is a "steep learning curve" at the beginning, as well as for reaching perfection at the end. Indeed, we are looking at exactly the learning curve! And progress is measured in bits. "That's the nature of math," as they say.

8.4.2 Test Effort: Visualizing the Test Effort Curve

Now let's "see" what this means for the SweSAT-Q test. We can't even imagine four dimensions, let alone this many. But we like the following visual image. Imagine following a footpath through a dense forest, a climax forest that has never been logged or burned. We are accompanied by a botanist, who informs us that the forest has, at last count, 412 different species of plant life. He would love to tell us about each one, but we say that we are pressed for time and intend to walk all the way along this path to the other side. The path twists and turns, but is nevertheless is a one-dimension thing. Sure, the forest is complex, but we ignore its complexity. Moreover, the path contains markers along the way indicating how many metres have been walked to that point, so that we can leave the forest knowing how long the walk was.

A favourite tool in a statistician's toolbox is a technique for looking at high dimensional objects in one, two or more dimensions. The technique involves rotating the curve within its high dimensional space so that as much of its shape as possible is shown within a preset set of dimensions.[1] We can then at least view its three-dimensional image. The rotation technique also yields a percentage measure of how much of the twists and turns in the curve actually lie within this lower-dimensional diagram. It turns out, amazingly, that 99.6% of the shape of the 412-dimensional SweSAT-Q surprisal answer curve can be seen in this way! The curve is not, after all, that complicated.

Figure 8.9 shows this path. There is no particular meaning to the three dimensions themselves, so we have labelled them "West", "North" and "Above." The displayed shape also does not depend on the ranges of the three plotting axes, so we have set all three to from 0 to 100. The 3D plotting software allows us to rotate the path as

[1]This is called *principal components analysis*, and only our readers who have taken advanced courses in statistics will know how this works.

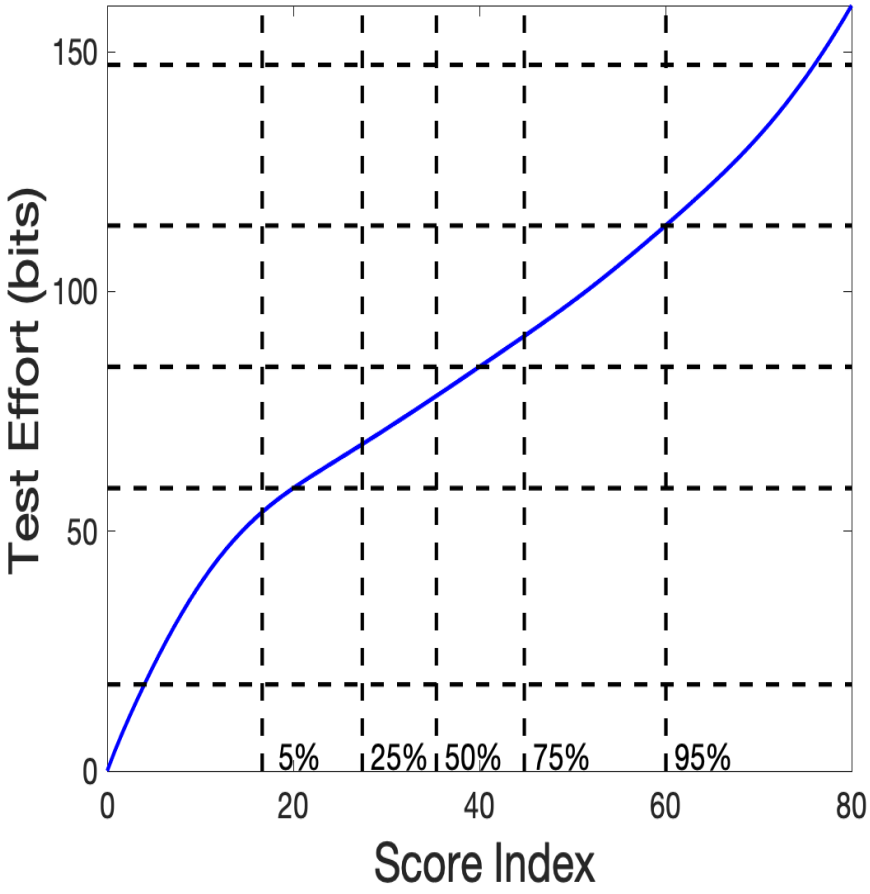


Figure 8.8: The relationship between the test effort curve and the score index for the SweSAT-Q. The dashed lines indicate the marker percentages in the respective directions.

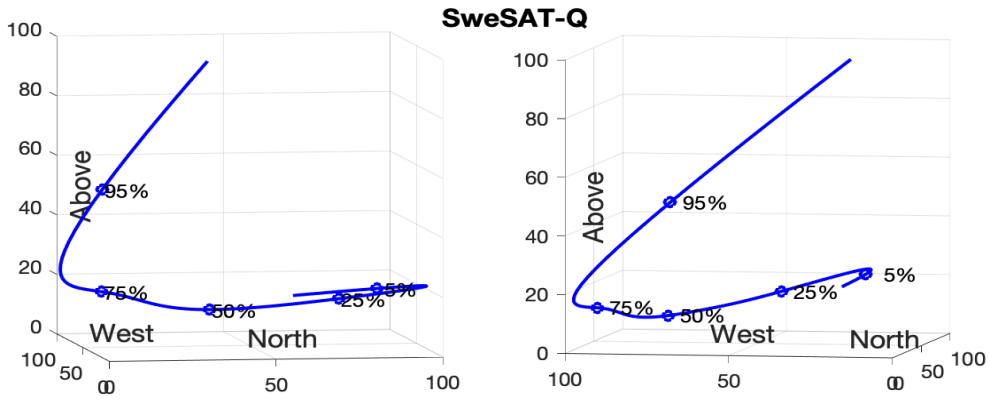


Figure 8.9: The SweSAT-Q test effort path displayed in the three dimensions in 99.6% of its shape is situated. Also displayed on this curve are the points corresponding to our five marker percentages.

we please, and so we have tried to help the viewer to see the 3D structure in a static 2D image by supplying two viewing panels with different orientations.

We see three bends in the test effort curve in Figure 8.9. The first is at the 5% point, the weakest performance end. Here we see the point where the most limited test takers switch from just guessing to being able to get at least a few of the questions right on their own. Beyond this point, we see what looks like steady progress in a relatively straight line to the 50% marker point. That portion of the curve passes close to the relatively easy questions. Then there is a slight bend, followed by a transition to beyond the 75% marker, which we can take as including the test takers who are reasonably competent for the medium difficulty questions. But the top 20% of the test takers display brilliance by a dramatic and lengthy flight in a new direction toward perfect performance.

How does the 3-D test effort curve look for our other data sets? The test effort path in Figure 8.10 for the SweSAT-V generates almost exactly the same shape, and has a length of about 165 bits. The only shape difference is minor; the hook below the 5% point is much smaller, which is consistent with the fact that this subtest is easier and therefore has more able test takers, even in the bottom performance level.

The National Math test effort path in Figure 8.11 exhibits a smoother path without sharp changes in direction, but nevertheless a similar overall shape where first 50% of the test takers are close low on the vertical plane and the remainder rise

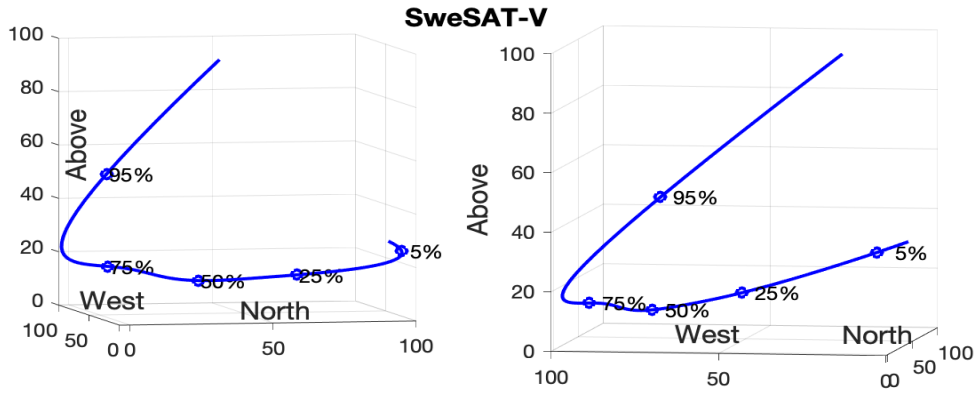


Figure 8.10: The SweSAT-V test effort path displayed in the three dimensions in 99.5% of its shape is situated. Also displayed on this curve are the points corresponding to our five marker percentages.

upward. The test effort curve is nearly 96 bits long, which amounts to an average of 3 bits per question. This is interesting! Constructed response questions seem to be about 1 1/2 times as informative as multiple choice questions. The shape of the curve encourages us to suspect that there is an overall two-phase process in acquiring knowledge, whether quantitative or verbal. In the first phase, we master the basic technology of either solving equations or writing prose. In the second phase, we learn how to put this technology to work for real-world problems.

The Symptom Distress Scale test effort path shape is also remarkably like those for the performance-oriented SweSAT and National Math tests, but the 25% and 50% marker locations come somewhat earlier along the path. Its length is just over 75, amounting to nearly six bits per question. That fact that the answers are ordered from least to worst causes the answers to each question to be more informative. We can conclude that the SDS is "easier" than the tests in the sense that rather more of its test takers can be found in the lower portion of the curve. It also appears that the test effort curve is roughly the same two-phase shape for both achievement tests and self-report scales.

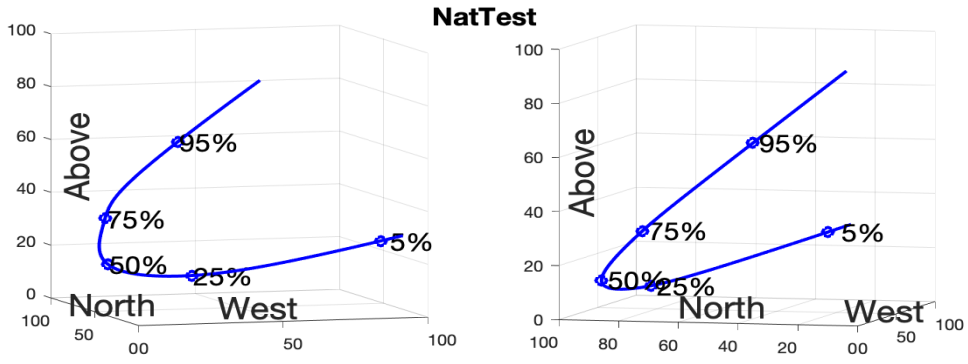


Figure 8.11: The National Mathematics test effort path displayed in the three dimensions in 98.2% of its shape is situated. Also displayed on this curve are the points corresponding to our five marker percentages.

8.4.3 Test Effort as a Score Index

Figure 8.13 shows how test effort pays off in terms of the average test score for the SweSAT-Q, which we recall is defined by the multiple choice answer scoring system. We are struck by the fact that the bottom 25% of the students can expect to have about the same test score of about 18 or so, which again is consistent with the fact that they are essentially guessing at the right answer. Can say, as a consequence, that the average test score tells us about the performance level of the bottom quarter of the test takers. They might as well have stayed home. We also note that, among the top 25% of the test takers, test effort brings less and less in terms of improvement in test score. This is another indication of the mis-behaviour of some of the questions.

Finally, Figure 8.14 displays the distribution of the test effort scores. Here we use bits as the unit, but there would be no harm in re-scaling the scores by dividing them by 160 and multiplying them by 100, so that they would indicate the percent of perfect performance. Or, indeed, using any positive constant since re-scaling magnitudes only changes their unit of measurement and leaves the appropriate of adding and subtracting intact.

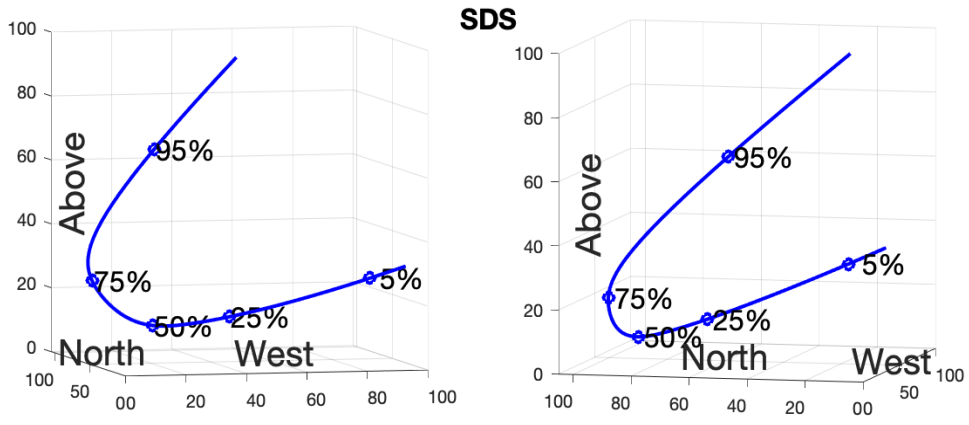


Figure 8.12: The Symptom Distress Scale test effort path displayed in the three dimensions in 98.8% of its shape is situated. Also displayed on this curve are the points corresponding to our five marker percentages.

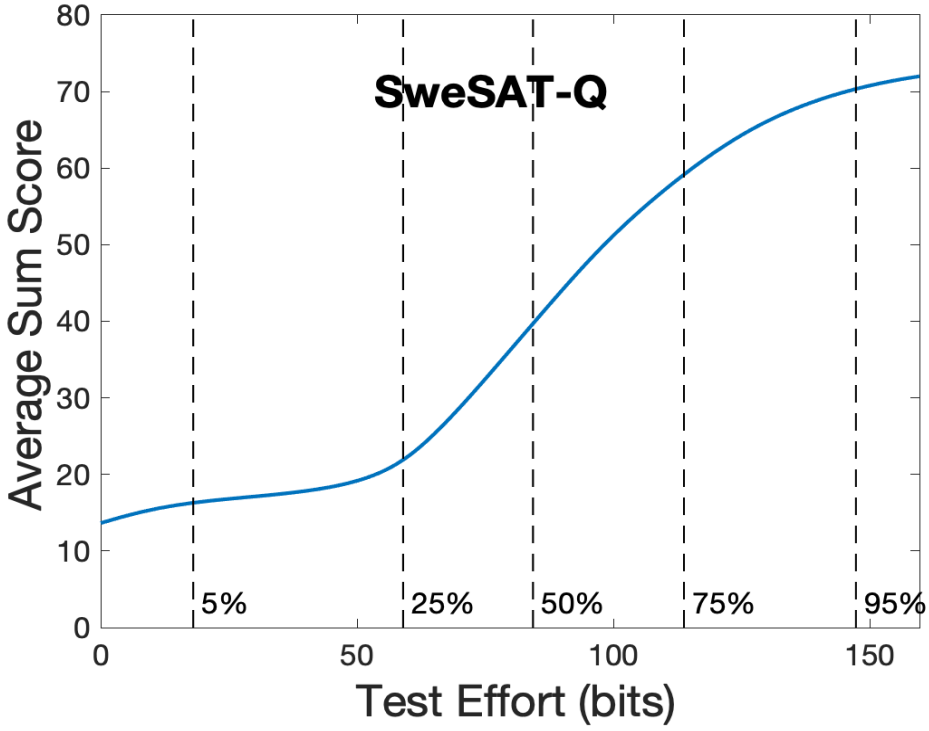


Figure 8.13: The average test score displayed as a function of test effort.

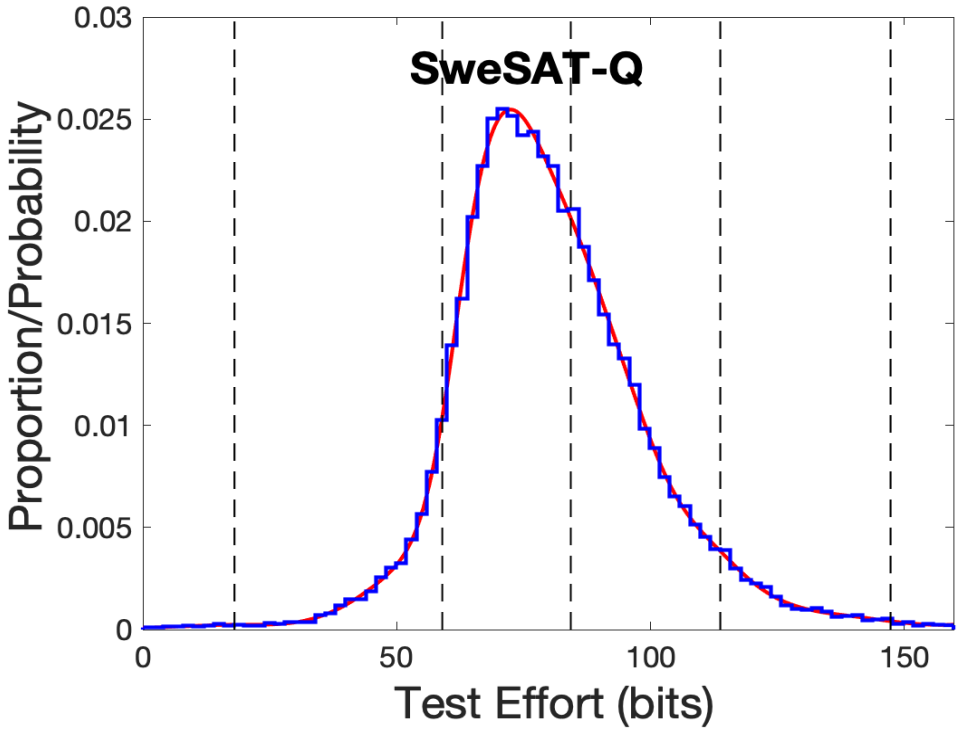


Figure 8.14: The distribution of test effort scores

Chapter 9

Score Performances

Whether you came to this chapter directly from the introduction, or you took the longer route of reading the intervening chapters and now understand how better scoring works, it is time to see just by how much the optimal score outperforms the sum score.

But before we can look at the evidence, we need to look carefully at how we describe the quality of a performance estimate. In particular, we need to consider both the *fixed error* and the *variability* of a performance estimate, and so recognize that the quality of an estimate is a matter of two properties, not one.

The quality of an estimate depends on the performance level, and therefore will displayed as a curve. The two properties of fixed error and variability add together in a particular way that we will describe below, and as a result we will want to look the *fixed error performance curve*, at the *variability performance curve*, and finally at the *total performance curve*.

We do assume here that you know what an *average* value is, namely

$$average\ value = sum\ of\ many\ values\ divided\ by\ number\ of\ values\ summed.$$

We have already used the median to indicate the centre of a distribution of quantities, but the average, or in statistical jargon the *mean*, has some special properties that we need now, and a mean value is for various reasons often different from the corresponding median value. But we'll stick to the term "average" instead of "mean" so as to avoid statistical jargon.

9.1 Two Types of Error

The term *error* can be defined in many ways, but almost always it is used to indicate by how much a desired outcome fails to materialize. Most of the time we use differences to represent errors, as in

$$error = observation - truth,$$

and we will do so in this chapter.

We now discuss two types of error: *fixed* and *random*.

9.1.1 A Look at Fixed Error

Golf is a game capable of generating almost every type of error one can imagine. One thing it sure has for some of us is fixed error. This happens when we fire a drive down the fairway and the ball consistently heads to the left (hook) or right (slice). You'd think that this would be easy to fix, but it isn't. Fixed error happens when the average result is off the mark. We can detect fixed error by taking the average of a fairly large number of ball positions and compare this average with where we want the golf ball to go. Specifically, fixed error is the difference between the target and the average of the data.

$$fixed\,error = average - truth$$

We can think of fixed error as the amount of systemic error in the process. Once we detect fixed error, we want to do our best to fix it. There is usually some process that will get us back on target overall, such as signing up with a golf pro. The key insight here is that fixed error is the part of error that does not change from one shot or one measurement to another. That is, fixed error is constant error.[1] As such, it is predictable. And because it is predictable, we would rather not have any excuse to making this type of error, or at least hope to make it small.

9.1.2 A Look at Random Error

Even after we have managed to reign in the fixed error, experience tells us that the golf ball will be more or less randomly off the target. This type of random error is much more difficult to fix since it is usually caused to forces over which we have no control, such as in wind, grass quality, inability to estimate distance exactly and so on. Because random error is intrinsically unpredictable, it is the gambling part of performance and therefore, for some of us, the main reason why we spend so much money on games like golf. While we can't eliminate random error in golf, we can do things like take lessons or training to reduce its typical size.

We used the average to define fixed error, and we also use the average to assess this event-to-event random variation. Since random variations are sometimes positive and sometimes negative, and therefore would tend to cancel out in the long run if directly averaged, the usual procedure in data analysis is to square the variations around the average, and then average these squared differences. The result, an averaged squared deviation from the mean, is called *variance*.

$variance = average\ of\ (observation - average)^2$.

[1] Statisticians refer to fixed error as *bias*, one of the few happy choices of terminology in the field.

Variance quantifies what we want to reduce since the smaller the variance, the less uncontrollable error we will see. Variance is almost always positive, and will only be equal to zero if every observation is equal to the average observation. Of course, this still leaves fixed error error in the picture.

But variance has a cosmetic problem. Most of us don't want to think in terms of squared things, so that in practice we take the square root of the variance, and this is called in statistics the *standard deviation*.[2]

standard deviation = square root of variance.

Standard deviation is a direct estimate of random error since it is in the same scale or unit of measurement as the observations themselves. We sometimes call it a "root–mean–square" quantity. We will, however, continue to refer to what it measures as "random error."[3]

Using more observations to calculate an estimate of something does reduce the estimate's random variation. But not as quickly as you might think.

Let's use N to indicate the number of observations. We expect that the bigger N, the better an estimate of an average is as an estimate of the true average. But the quality of an estimate of an average, measured by how small its standard deviation over a large number of equal-sized sets of data is, is proportional to $1/\sqrt{N}$, and not to $1/N$ as you might suppose. This means that if you want to halve the size of the random error of an estimate, you will have to *quadruple* your number of observations.[4]

This principle is so important in statistics and every day life, that it deserves its own word equation. In statistics we use the Greek letters μ ("mew") to indicate the average or mean and σ to indication the true standard deviation of a set of data. What we compute for a specific set of N observations is only an estimate of μ, and we stats folks indicate this by $\bar{x}$. Our word equation provides a single formula for what the standard deviation of $\bar{x}$ would be if a large number of laboratories all, quite independently, collected N observations.

[2]Standard deviation is a type of average that belongs to the more extended family of "transform – average – back transform" operations. Here the transform is the squaring operation. These more general types of average are used widely in statistics.

[3]There is a lot more to say about both the mean and the standard deviation, but if you are still waiting to take your first statistics course, we can pass on a tip or two. If your target is a number, and you add and subtract the standard deviation from a specific score that you have obtained, the interval between these two values will contain the true average about 2/3 of the time. If the fixed error is negligible, this also implies that your interval will contain the target with this probability. If, instead, you add and subtract two standard deviations to your observed value, it will contain the the true average and possibly the target a satisfying 95% of the time. These two values are called *confidence limits* in the statistical literature.

[4]The accuracy of the variance of your estimate, on the other hand, really does decrease in proportion to N, but most of us prefer the standard deviation as a measure of error, and it improves in proportion to $\sqrt{N}$.

$$\textit{Standard deviation of } \bar{x} = \sigma/\sqrt{N}.$$

Of course, we don't often have a lot of independent samples, each of size N, but replacing σ in this equation by the standard deviation of our own sample provides a useful estimate of $\bar{x}$'s true standard deviation.

cite Howard here

9.1.3 Combining fixed error and random error to get total error

The error that we see is the error that we want to fix, and is a compound of both random error and fixed error. But how, exactly, are these two types of error combined together? We define the *total error* as

$$\textit{Square of total error} = \textit{average of } (\textit{observation} - \textit{truth})^2.$$

That is, we substitute *truth* for *average* in the equation for *variance*, and then, as for the standard deviation, take the square root of the result to get total error itself. That is, total error is another "root–mean–square" quantity.[5]

Average squared error has a simple relationship to variance and fixed error:

$$\textit{squared total error} = \textit{variance} + fixederror^2.$$

This implies that *average squared error* = *variance* if there is the fixed error is zero. We are especially interested in reducing fixed error down to a point where it is negligible relative to random error.

9.2 Measuring Sources of Error by Computer Simulation and Mathematics.

We have some tools at hand that are neither difficult or expensive, and that can inform us about fixed, random and total error.

The results concerning the quality of our estimates of test score and score index that we report here are based on computer simulation. We choose a test, such as one of the SweSAT-Q, SweSAT-V, National Math Test and Symptom Distress Scale, and we estimate both the probability and sensitivity curves for each answer within each question, and the score for each test taker.

Then we pretend that these are true values. This is not unreasonable if the tests provide a large amount of data and our estimates of the curves and the distribution of test scores are really rather accurate. That is, whatever the truth is, it is surely not very different from our results, and therefore we can consider it useful to see how close we can come to these "true" results.

[5]Statisticians refer to squared total error as *mean squared error* and abbreviate it to "MSE".

We then simulate some data. This is easy to do because the probability curves have a direct connection to the data that we simulate. It isn't necessary to use the score estimates that we computed. We can, as we did, set up a fine sequence of fixed scores, and then simulate large amounts of data for each of these fixed score values. We simulated 1000 sets of data for each test score value that we used, and this took only a minute or so on a desktop computer. The results for each of possible sum score value with are correct about 4 in the second decimal place.

It is natural to report our results using graphs rather than tables, since the error levels change smoothly as we move through the sequence of score values that we used. We developed a three-panel plot that shows, from top to bottom, in red, the fixed, random and total error levels for a particular type of score. For comparison purposes, in each plot and within each of its three panels, we also show the error levels for the sum scores in blue.[6]

9.3 Sources of Error for the Test Scores

Since the score index values determine the corresponding test score values, we first look at each of the plots of the score index error curves for each of the four sets of test data.

9.3.1 Error Levels for the SweSAT-Q and SweSAT-V Test Scores.

The three error levels for the SweSAT-Q are displayed in the panels of Figure 9.1. In each panel the error level for the test score is in red and that of the sum score in blue. We've used percent rank for the horizontal axis in order to put all four tests on the same scale.

Beginning with the top panel displaying fixed error or systematic bias, we note that the sum score has hardly any fixed error. In fact, the sum score can be shown to have theoretically zero bias for all score values, this being about its only good quality. The test score $\mu(\theta)$ does have some positive fixed error over must of the percent rank range, and some negative fixed error for the top test takers. But the fixed error, when combined with the random error using the word equation in Section 9.1 to obtain the total error, turns out to be negligible. That is, we see the bottom two panels that the shape of the red curves for random and total error are almost identical, implying that most of the error is in fact random. This is good news.

[6]The the error level results in the following sections do not depend on which score index is used, and we want to minimize the effort in comparing results. As a consequence, we defined the score index as extending from 0 to the largest possible score value, and with roughly the same distribution. If a different score index is used, such as percent rank or test effort path length, the plots will change their shape because the horizontal axis changes, but will not change in terms of their levels at marker percentages, boundaries or at the locations of features.

The middle panel shows the level of random error, which is a type of error that is unrelated to the fixed error. The bottom panel displays the corresponding total error, which we have noted is almost identical to the random error, and so effectively tells the same story as the middle panel.

The test score for test takers near the middle 50% is about one score point better than that for the sum score, which in turn is about 4.3 test score points. Moving to the left, we find that the 25% level random errors for the test score and the sum score are 2.2 and 3.9, respectively; which is a considerably larger discrepancy. But at the 75% level the random errors are 3.4 and 3.8, respectively. The discrepancies at the score extremes 5% and 95% are on the other hand much larger than at either the median level of 50% or the two on either side of it. The 0.5 versus the 3.6 discrepancy at 5% is due to the fact that the optimal score removes most of the inflation in the sum score variation due to guessing. At the high-end 95% level the test and sum score random errors are 1.9 and 2.9, respectively, so that the optimal test scores are about 50% more accurate than the sum scores. We will see in the next Section that all of these discrepancies are much more serious than they appear in this figure when we consider the relative costs of obtaining the two accuracy levels.

Figure 9.2 displays remarkably similar results for the SweSAT-V verbal subtest, except that the discrepancies between the two types of scores are a bit smaller. Perhaps this small difference in error patterns is due to the fact that the verbal subtest is on the whole rather easier.

9.3.2 Error Levels for the National Mathematics Score Indices.

Figure 9.3 reveals somewhat similar shape features in the error curves for the constructed response National Mathematics test. However, we see here that the test score and sum score random errors are in general much smaller than those for the SweSAT-Q and SweSAT-V over the central 50% of the test takers between the marker percents of 25% and 75%. We think that the better performance of the sum score here is due to the small role that guessing plays in a constructed response test. That is, the optimal score gains more accuracy for multiple choice test because it strips off most of the guessing effect.

Of course, the SweSAT subtests have more questions than the National Math test, so it is interesting to look at the random error for, say, the 50% performance level, as fractions of the number of questions. For the SweSAT-Q subtest, the random error per question is $3.38/80 = 0.042$, and for the National Mathematics test this ratio is $3.05/57 = 0.054$. It helps to have more questions, in other words.

On the other hand, at the high performance 95% level, the ratio of test score to the sum score error random error is 0.65 for the SweSAT-Q and 0.58 for the National Math test. That is, the optimal score reduces the random error more for the constructed response test than it does for the multiple choice test.

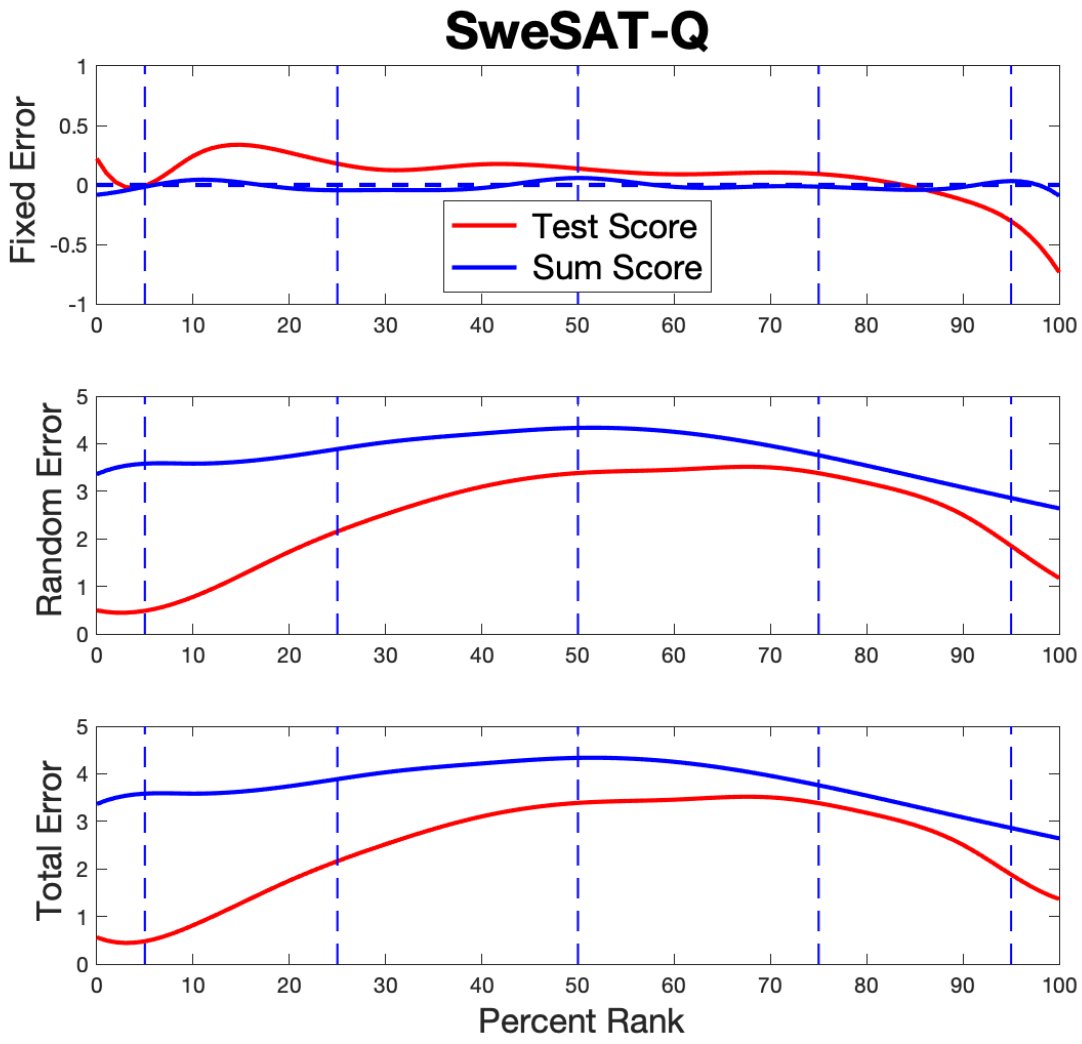


Figure 9.1: The levels of error for the three types of error for the 80-question SweSAT quantitative multiple choice subtest. Those for the test scores $\mu(\theta)$ are shown in red, and those for the sum score in blue.

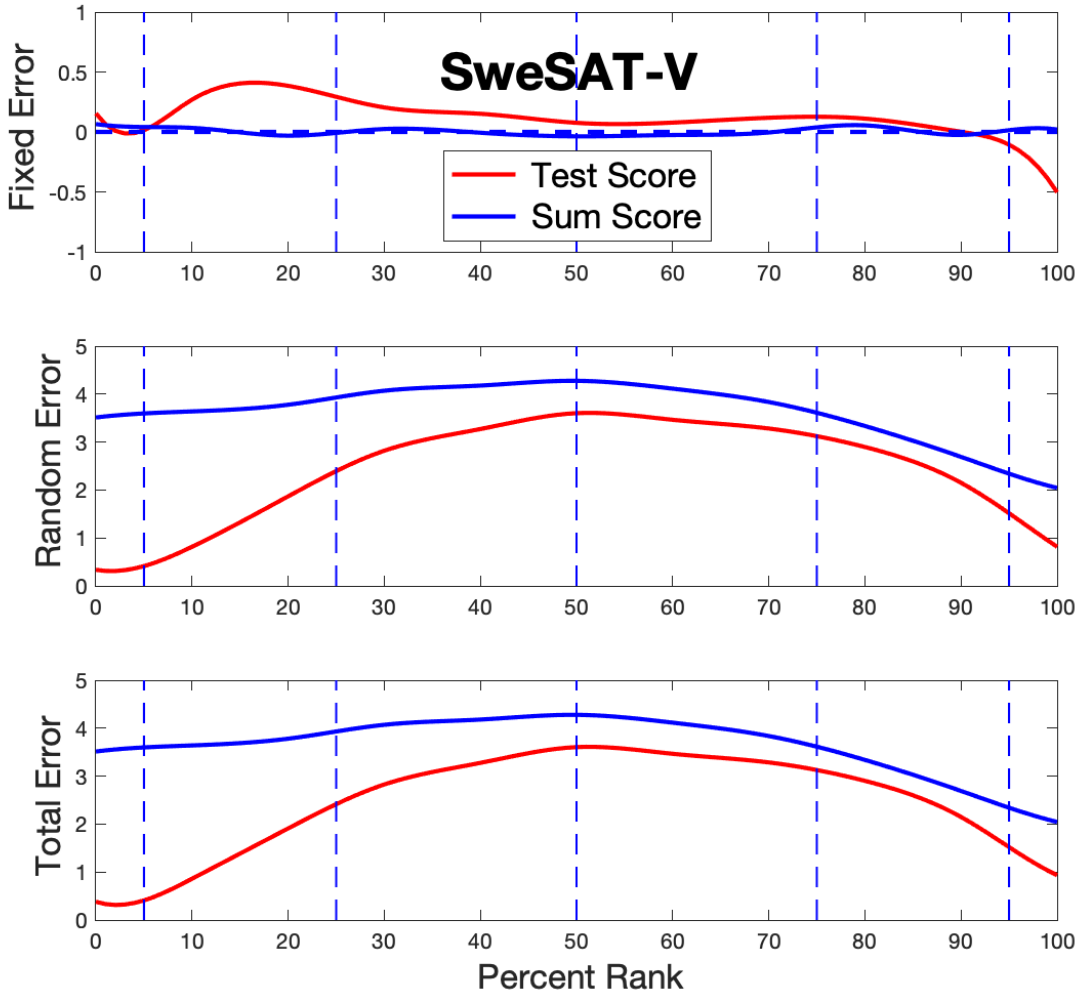


Figure 9.2: The levels of error for the three types of error for the 80-question SweSAT verbal multiple choice subtest. Those for the test scores $\mu(\theta)$ are shown in red, and those for the sum score in blue.

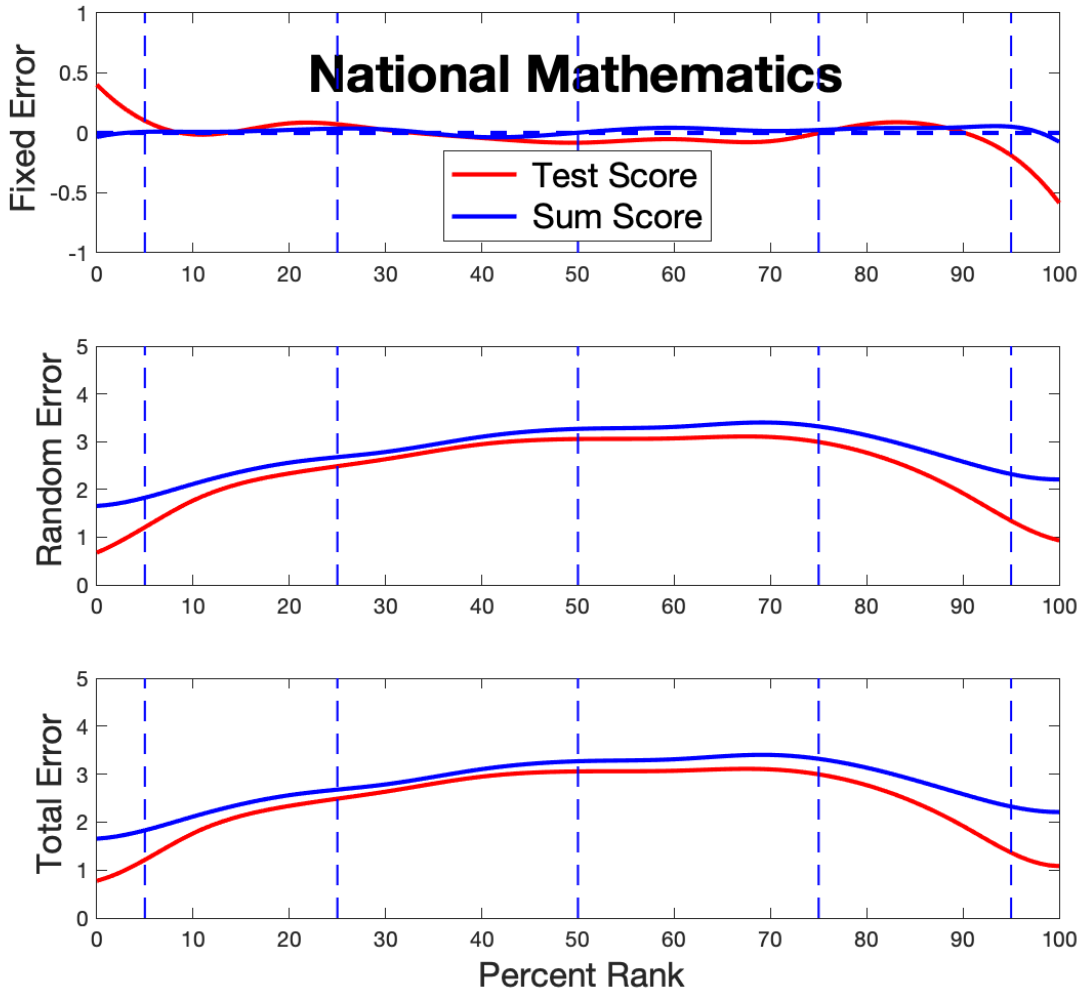


Figure 9.3: The error levels for the score index for the National Mathematics test are shown in red, and those for the sum score in blue.

9.3.3 Error Levels for the Symptom Distress Scale Score Indices.

The Symptom Distress Scale, whose results are in Figure 9.4, has a very different distribution of score values, with many more test takers at the bottom end of the scale than at the top. The nursing profession will be more interested in the higher end of the score range, where their interventions can provide greater benefit. We noted that the considerable majority of the cancer patients experienced relatively manageable levels of distress. We see that the spread in random error between the test score increases steadily above the 50% level. At the highest level of distress the random error for the sum score is over twice as high as for the test score $\mu(\theta)$. For a small scale of this nature, using optimal test scoring really pays off where it counts.

9.3.4 Error Levels for the Arc Length Percent Score.

We have argued that the total effort or arc length score should be considered as an alternative score because it is not affected by the score values that test designers assign to answers. Figure 9.5 shows for each of our four tests the total error curve for this score. Only the test effort score is shown in each panel because the sum score depends on the answer score values, and therefore is not comparable. The percent version of the test effort score is used, where the unit is one percent, in order to facilitate comparing the curves over tests. We see that the test effort score has about the same accuracy as the test score.

9.4 The Cost View of Test Scores

Both test takers and test designers want to ask, "What's in better scoring of test scores for me?".

The test taker may be a bit disturbed to learn, after all day of completing a 80–question test, that the sum score will have a random error of as much as four. We have pointed out that a good rule of thumb is that 95% of the score estimates among those taking the test and all being at the same level of true performance will be within two standard deviations of the right value.

At a true performance level equal to the median score, about 36 for the SweSAT-Q, this implies that sum score values for those with a "true" score of 36 will range from 27.4 to 44.6 for 95% percent of this group. This pretty much coincides with 50% of the scores on the actual test. It's difficult to justify calling such a score "accurate." If a university decides to accept test takers with scores of 36 or above in the naive belief that these actually have a performance level of at least 36, about 25% of those admitted will fall short of the assumed performance level threshold.

The optimal score will do somewhat better, of course, and has a corresponding range from 29.4 to 42.4. But we have to face the reality that multiple choice questions

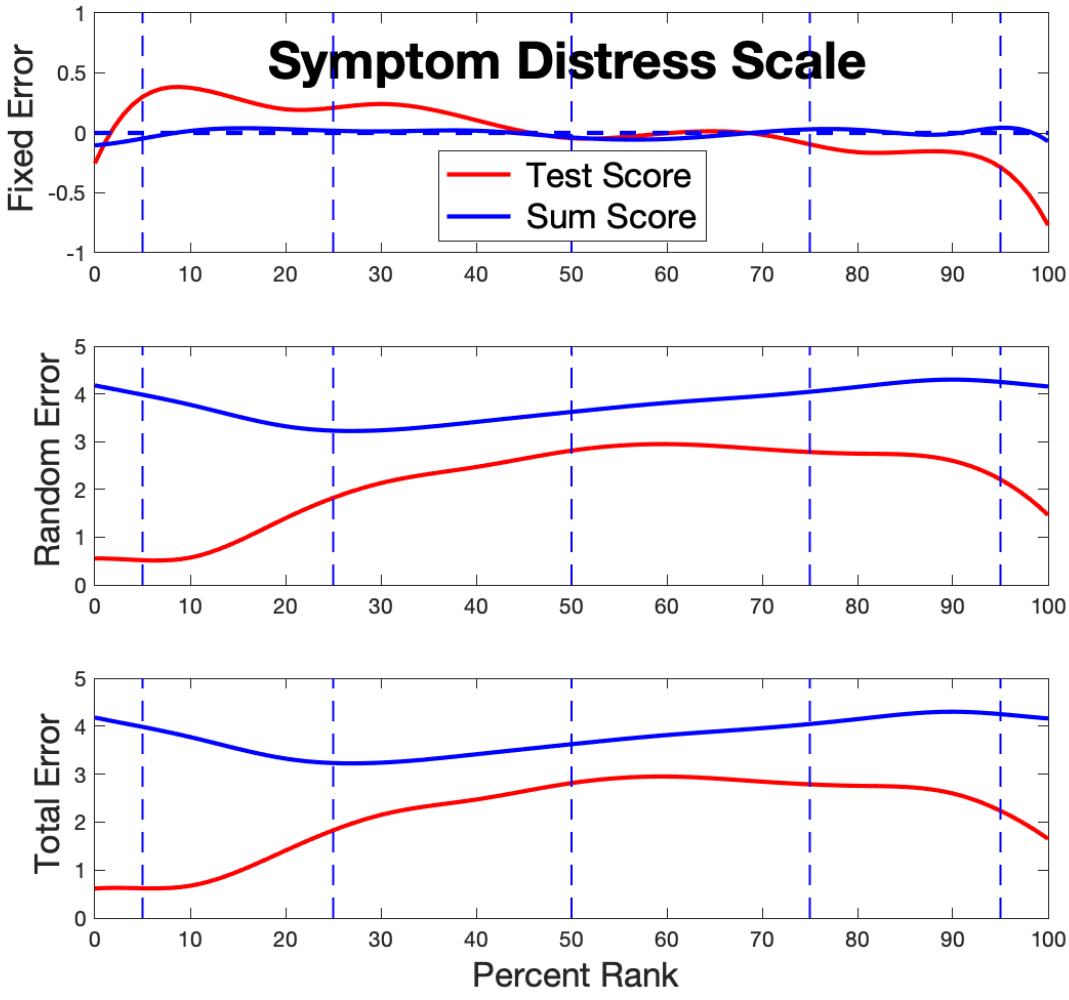


Figure 9.4: The error levels for the score index for the Symptom Distress Scale are shown in red, and those for the sum score in blue.

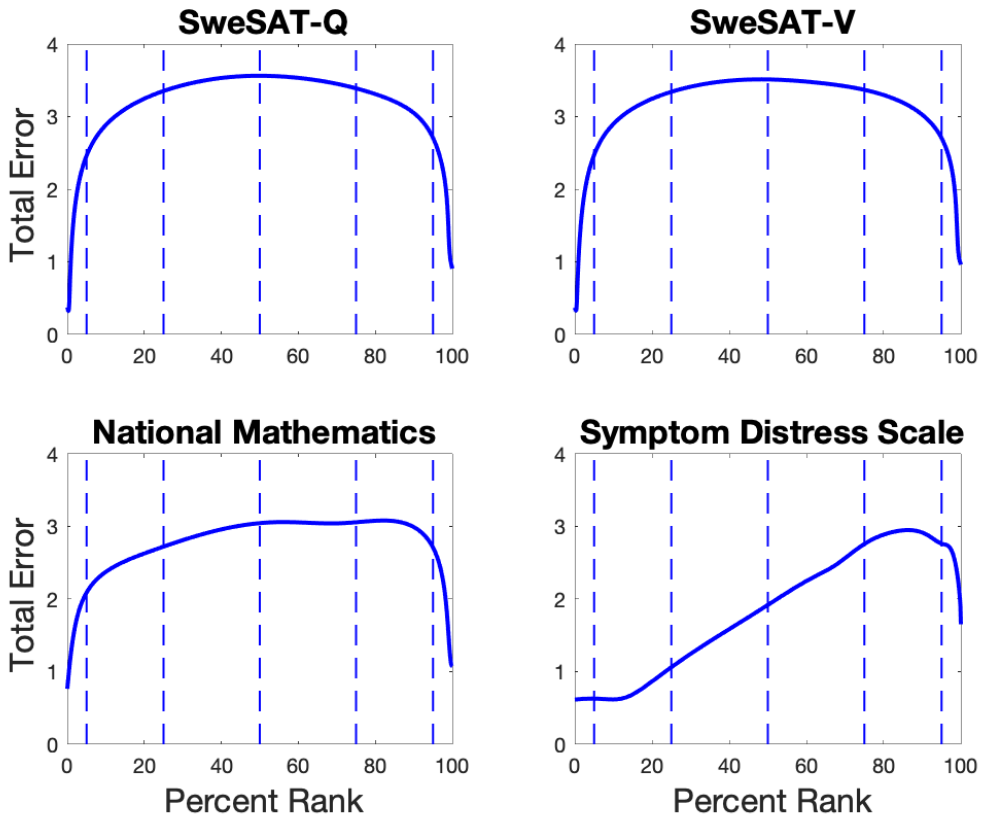


Figure 9.5: The total error at any test effort level (percent unit).

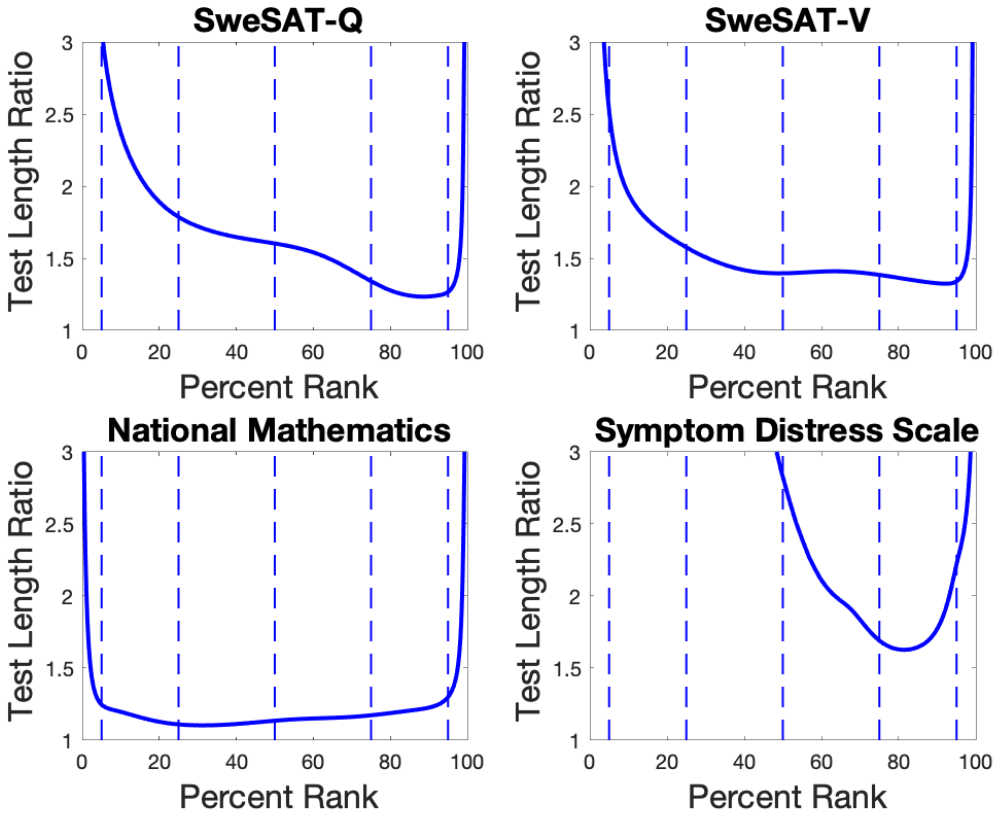


Figure 9.6: The length of a sum scored test that would have the same accuracy as as the current better scored test.

are just not that informative, no matter how they are scored. The test taker will be on solid ground in demanding that optimal scoring be used to improve the accuracy of the score, but will also face the statistical reality that a really substantial improvement would also require taking a longer test. A university will tend to agree, since it costs a lot to attempt to educate a student who does not have the capacity to cope with the curriculum.

Figure 9.6 shows for all four tests the length of a sum-scored test that would have the same accuracy as the current optimally scored test actually has. The word formula for the cost fraction is:

$$test\,length\,ratio = (total\,sum\,score\,error)^2/(total\,test\,score\,error)^2$$

The test designer, on the other hand, may be inclined to say that the country has done well in the past with its post-secondary education infrastructure in basing selection on scores as noisy as the sum scores. But a lot of money could be saved if the test could be shortened substantially, and about the same level of accuracy maintained by employing optimal scoring. Especially, as we intend to do, if the

required computer software is made available for free. State and federal governments, as well as tax payers, will add their support to this approach.

Figure 9.7 shows for all four tests the fraction of the length of a sum-scored test that an optimally scored test will have and still maintain the traditional sum-scoring accuracy level. This is the fraction of the sum scored test required to produce the optimally scored equivalent test. The word formula for the cost fraction is:

$$cost\ fraction = (total\ test\ score\ error)^2/(total\ sum\ score\ error)^2$$

If the error standard deviation at the median score value is to be maintained,

- the length of the optimally scored SweSAT-Q subtest would be 61% of the length of the current test,
- the length of the optimally scored SweSAT-Q subtest would be 71% of the length of the current test,
- the length of the optimally scored National Mathematics test would be 87% of the length of the current test, and
- the length of the optimally scored Symptom Distress Scale would be 38% of the length of the current test.

9.5 What Score Should be Reported?

We are now left with some conclusions, and some questions.

There can be no justification except nostalgia for bad ideas for reporting the simple test score. What we've called the average score is far superior in terms of the size of the random error, and especially for extreme scores. The superiority derives from how effective the score index is at finding the best position on the test effort path for fitting a test taker's data. And we remind ourselves again that the score index estimation pays no attention to what the test designer assigns as a score to any answer. Instead, the score index uses as information how high the sensitivity curve is over score index values, which in turn signals the strength of the total evidence contributed by the test taker's choices on questions. In effect, sensitivity curve values play the role of the test designer's set of assigned scores, but does so in an intelligent optimal manner.

Test designers who assign weights to answers in tests and scales like the National Mathematics Test and the Symptom Distress Scale will understandably want a test score that reflects their choices in weights. The multiple choice format has little to defend it except that no one has to do make any decisions beyond which answer is correct. But, as we have seen, it is not rare to have questions that have more than one right answer, no right answer or a right answer that is treated as wrong.

We therefore recommend reporting both average sum scores and a scoring index. The great advantages realized by being able to add and subtract scores would strongly recommend the test effort index, or a normalized version of it.

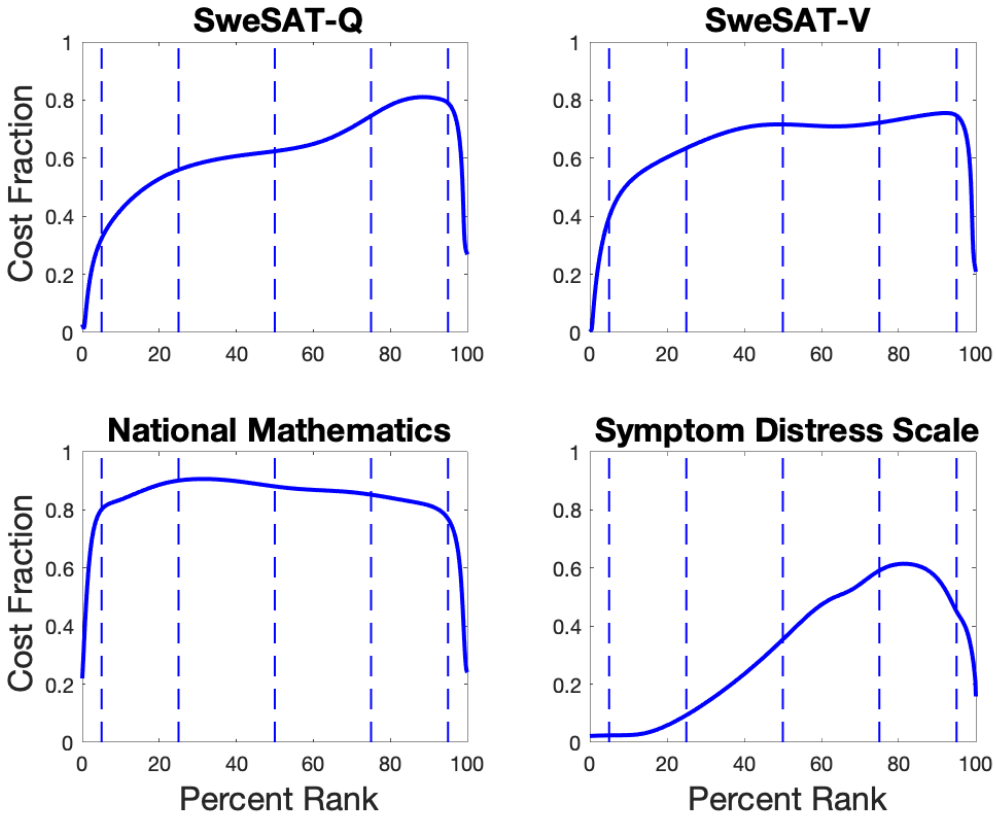


Figure 9.7: The fraction of the length of a sum scored that an optimally scored test will have if it has the sum score error level of accuracy.

Chapter 10

Test Analysis Cycle

this chapter is too short and would benefit from more details and some examples

10.1 Putting it all Together

At this point we want to bring the elements of an optimal scoring analysis together. Figure 10.1 shows the initial step followed by a cycle of four steps. This diagrammed as a loop because we recommend passing through these four steps twice.

Here is more information about each step.

10.1.1 Step 0:Sum Score Computation

This is a single step that we do only once to get the cycle started. But default, the score index continuum is set to the interval $[0, n]$ where n is the number of questions. Other intervals can be used, however, and the sum scores transformed to fit these intervals. The sum scores or their transformed counterparts are initial estimates for the score index values, and in the cycle itself these are changed to improved values, so that after one cycle the sum score no longer has a direct impact on any aspect of the analysis.

10.1.2 Step 1: Probability Density Estimation

The smooth curve in Figure **??** that approximates the proportions of test takers for the SweSAT-Q at each possible sum score level not only gives us an easier image to look at, but also a better framework for estimating the probability that a test taker will fall between any two score index values. In statistics jargon, this curve is called a *probability density function.*

In the earlier days of statistics, these curves were defined choosing within a relatively small library of simple mathematical curves one that fit the data reasonably well. This worked fine when the number of data values was small to medium. But

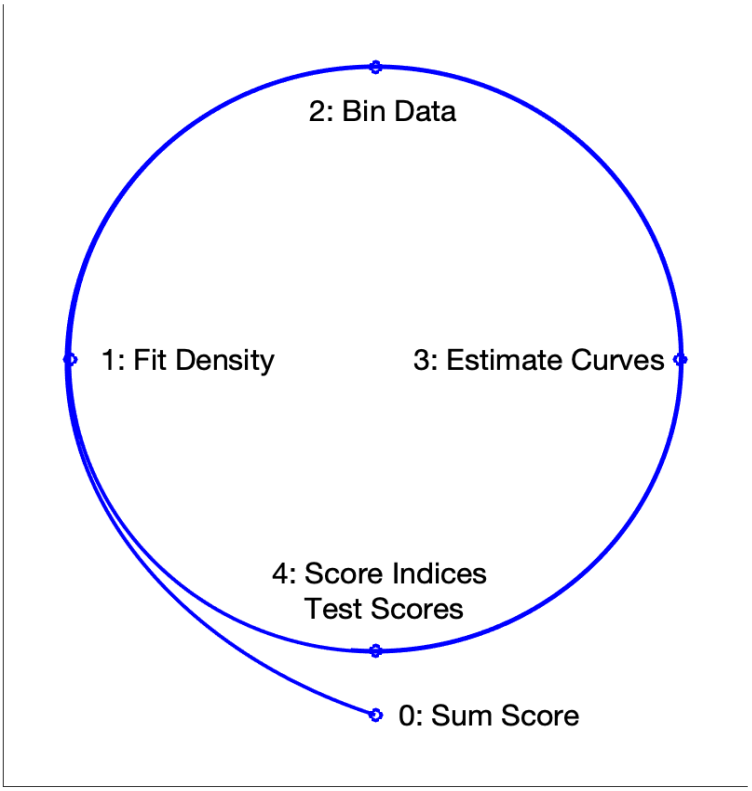


Figure 10.1: The test analysis cycle is initialized by computing sum scores, and then moves clockwise into the cycle where the first step is to estimate a probability density function, the second uses this density to define bin boundaries, the third estimates the probability and sensitive curves using the binned data, and the fourth computes optimal score index values and test scores. Two passes through the cycle are recommended.

within environments like testing, the shapes of these curves become more and more complex. Nowadays, we use methods that estimate curves up to any degree of complexity that is needed by the data. These are referred to as *nonparametric probability density functions.* The annotated reading list at the end of the book offers some places to go to learn more.

10.1.3 Step 2: Binning the data

We need to divide the test takers into a set of groups or bins of equal sizes but increasing score index values. For example, for the SweSAT subtests where we had over 55,000 test takers, we used 55 bins with about 1000 test takers per bin. The application TestGardener that we describe in the next chapter provides helpful information about how to do this in a range of situations.

To bin the data, we first sort the test takers in terms of their score values starting with the lowest scores and ending with the highest scores. The probability density function curve computed in Step 1 is used to compute the lower and upper boundaries of each bin so that about the same number of sorted test takers are within all of the bins. The center points of the bins are also calculated.

10.1.4 Step 3: Computing the surprisal, probability and sensitivity curves

We calculate for each answer within each question and for each bin the proportion of test takers within that bin that have chosen the answer. These proportions are then converted to the corresponding surprisal values. Each surprisal value is paired with the center of the the corresponding bin. This defines a set of points on a graph, and the surprisal curve is fit to the points using a curve-fitting techniques. Then the corresponding probability and sensitivity curves are computed from the surprisal curves. This stage only takes about a second for the SweSAT data.

10.1.5 Step 4: Computing the optimal score index and test score values.

This is the longest stage in the cycle. For each of the over 55,000 SweSAT test takers we use the 80 answer choices to find the score index value that defines the set of probability values that minimize the surprisal of the whole set of choices. This stage of the analysis required about a minute on a mid-speed desk top computer.

From this point, we can pass through the cycle again, so that each step has been executed twice. This puts enough distance between the initial sum score values and the optimized score values to ensure that going through further cycles will only produce relatively negligible improvements in the results.

Chapter 11

The TestGardener Application

11.1 Introduction to the TestGardener Application for the Analysis of Test Data

11.1.1 Who is TestGardener Designed For?

Using TestGardener does not require any formal statistical knowledge beyond what would be provided by a first course in statistics in a social science department. There the user would encounter the concept of probability and become familiar with basic statistical notation. TestGardener can also be used by test takers themselves if they are supplied with their own test data and the specifications of the model used to analyze these scores.

TestGardener uses graphical displays to communicate results, and provides help facilities for anyone having questions about what they are seeing. The essential aspects of each display were designed to be self-explanatory, although more statistically sophisticated users will also find information that they may find helpful. TestGardner is used interactively, but also stores the results of its calculations in files that may be processed later by either TestGardener itself or other programs such as Excel.

Instructors or questionnaire developers will find TestGardner helpful for diagnosing problems with items, and for deciding whether to rewrite items in order to clear up ambiguous wording or to offer wrong options that are more plausible.

Instructors of courses in the social sciences will find that using TestGardener along with its manual is an effective way to communicate the basic ideas of psychometrics and item response theory to students.

TestGardner makes use of modern statistical methods to produce accurate estimates of examinee or respondent characteristics. For example, for examination data, TestGardner enables better estimates of examinee proficiency or ability by making use of the information provided by which wrong options were chosen for incorrectly answered items. These estimates will be more precise than the conventional estimates based only on number correct, and especially for examinees of low to medium profi-

ciency. These more efficient estimates, which are still expressed in the familiar range as [0,n] or [0,100], can be used to either replace or modify the classical number correct scores reported for examinees. TestGardener is based on TestGraf, a program for the graphical analysis of multiple-choice test and questionnaire data that was published in 2000 by James Ramsay. The main differences between TestGraf and TestGardener are their smoothing and scoring algorithms - the former used kernel smoothing and sum score, which the later implements spline smoothing and optimal score.

Although by default TestGardner is used to study the internal structure of a test or scale, TestGardner can also be used to study how individual items relate to scores on some entirely separate set of scores of measures on the examinees or respondents. For example, TestGardner might be used by an instructor to see how well test items relate to the final grade of examinees, which might be a composite of other tests as well as this one. A clinician interested in developing a scale measuring, for example, level of distress could employ TestGardner to if ethnic group or language proficiency play a role in how patients respond to certain questions in the scale.

TestGardener has two versions: a stand-aloneapplication on Windows system and a web-based version that can be used on major browsers. For the simplicity of distribution, especially for users on other operating systems, this tutorial will focus on the web-based version. The TestGardener is still under development, as we are working on the Manual, Theory, and Resource materials and adding or modifying some of the displays. We will focus on the main functionalities and displays in the following tutorial.

11.1.2 TestGardener, Score Indices and Test Scores

We offer here a recapitulation of the material in Chapter 6 about the two types of scores that TestGardener produces for each test taker. The term "score" can refer to either of these score types and TestGardener will produce files that containing both of them.

The Score Index

The score index is a continuum over a closed interval. Two popular examples are the number correct continuum $[0, n]$ and the percent score continuum $[0, 100]$. By "closed interval" we mean that scores are possible at each of the boundaries of a continuum, and that such scores can be interpreted as "off the scale" in one direction or another.

The score index is the independent variable or horizontal axis along which probability values are displayed. It is typically used as the display variable or abscissa for plots of probability values, sensitivity values, distribution and so on that are generated by TestGardener. Each test taker is assigned a position or value on this continuum.

A key feature of the score index is that it can be chosen arbitrarily, so long as it is a single closed interval. It functions essentially as a continuous indexing system for test takers, very much like a rank order would. But, since it is continuous, no two test

takers are likely to have the same index value, with the exception of those assigned to either boundary or those pairs with identical answer choices. Another way to say this is that the score index can be used to uniquely rank, without ties, test takers.

Psychometricians often use the Greek symbol θ for the value of a score index.

The Test Score

The test score for a test taker depends on that person's score index, but may have values that are quite different from those used a score indices. The test score is a function of three quantities:

1. The data values, whether index values of sets of 0's and 1's indicating which answers have been chosen.

2. Or, instead, the probabilities that each answer will be chosen that are computed by TestGardner and which depend on the value of the score index estimated for a test taker.

3. And, for sure, the numerical scores that the test designer chooses to assign to each answer for each question. When the questions are in the multiple choice format, it is often implicitly assumed that these numbers are (1) 1 for the correct answer, and (2) zero for the others. But even in this format designers may, and certainly have, used different values.

When a test score involves replacing 0/1 choice indicator values by the corresponding probabilities, which in turn depend a test taker's estimated score index, psychometricians call the test score an *expected score* since the term "expected" in statistics refers to the average of data over an infinitely large sample.

The expected test score is the sum over both answers and questions of the designer-assigned scores times the probabilities that the answers will be chosen.

The "observed" test score or "raw test score" is the same thing except that the probabilities are replaced by the 0/1 choice indicator values. Naturally in this case the resulting sum is actually only the sum over the chosen answers of the designer-assigned score values.

The essential features of the test score are that the test score does not depend on what score index continuum is used, but it, unlike the score index, does depend on the designer-assigned scores. If the test analyst decides to change one or more answer scores, the test score will change but the score index will not.

Psychometricians and statisticians often use the Greek letter μ for the test score. Since the test score is indirectly a function of the score index, we can also write this as $\mu(\theta_j)$ meaning the expected test score for test taker number j having a score index value of θ_j.

11.2 The Structure of the Data that Test Gardener Analyzes

11.2.1 The Data that TestGardener is Designed to Analyze

TestGardner is designed to aid the development, evaluation, and use of multiple-choice examinations, psychological scales, questionnaires, and similar types of data. These data have the following features:

- Each test taker is presented with a set of questions.
- Each question is either:
  - accompanied by a small set of answers and each of these is assigned a grade by by the test designer, or
  - requires a task to be completed, and a scoring person assigns one among a small set of grades to the completed information provided by the test taker.
- For each test taker, the final set of graded answers is converted to a single number that is designed to summarize the overall performance of some other status of the test taker.

The data that we use in this tutorial were the complete option choice records for randomly selected 2000 examinees who took the quantitative sections of one administration of the Swedish Scholastic Assessment Test, abbreviated here as the SweSAT. The quantitative section was administered in two sessions with 40 items per session. There were five options for items 23 to 28 and 63 to 68; and four for the remainder. In the full-information data, we added to each item an additional option to represent items that were either not attempted or had spoiled responses. Full-information data can also be transformed into binary data, where 1 indicates the examinee chose the right option and 0 otherwise.

11.2.2 Preparing the Data

Current version of TestGardener can only read text files with a specific format, but in a format that is easily exportable from other programs such as Excel or Microsoft Word. Figure 11.1 shows the top portion of a text file using the .txt format extension. This file is a *full information* file in the sense that the actual choice made for each question is recorded, rather than simply whether the answer was right or wrong. Figure 11.2 displays a the top portion of a file where only whether the answers are right or wrong. We call this format *binary*.

The number on the first line indicates the number of lines per examinee, which in this case is one. The second line is the key line that specifies for each test question

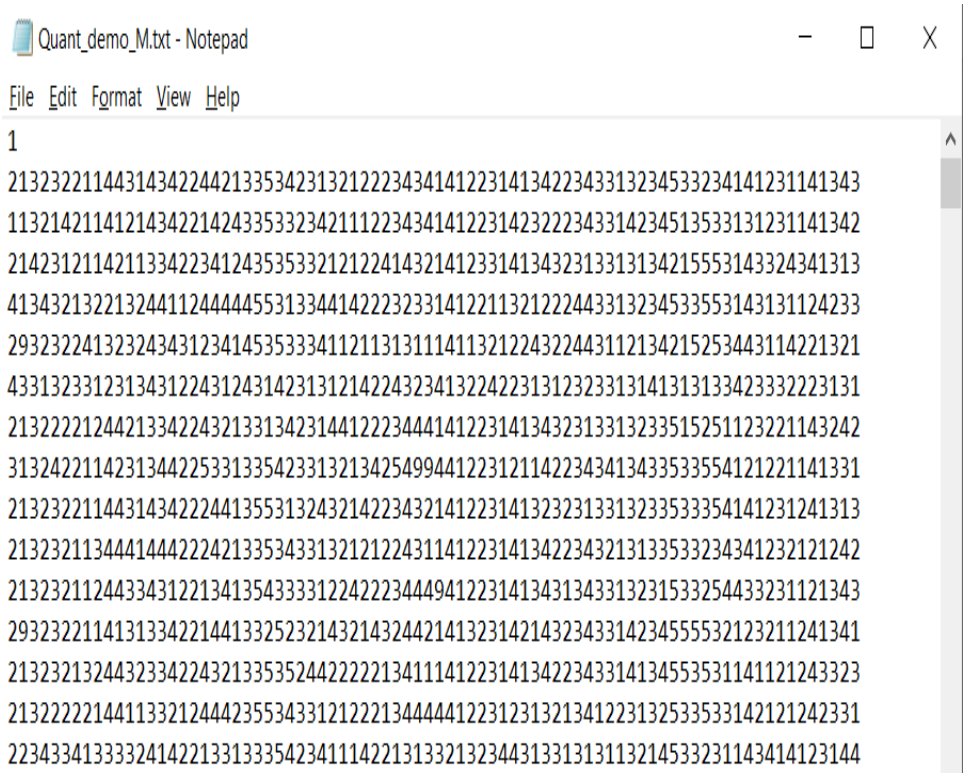
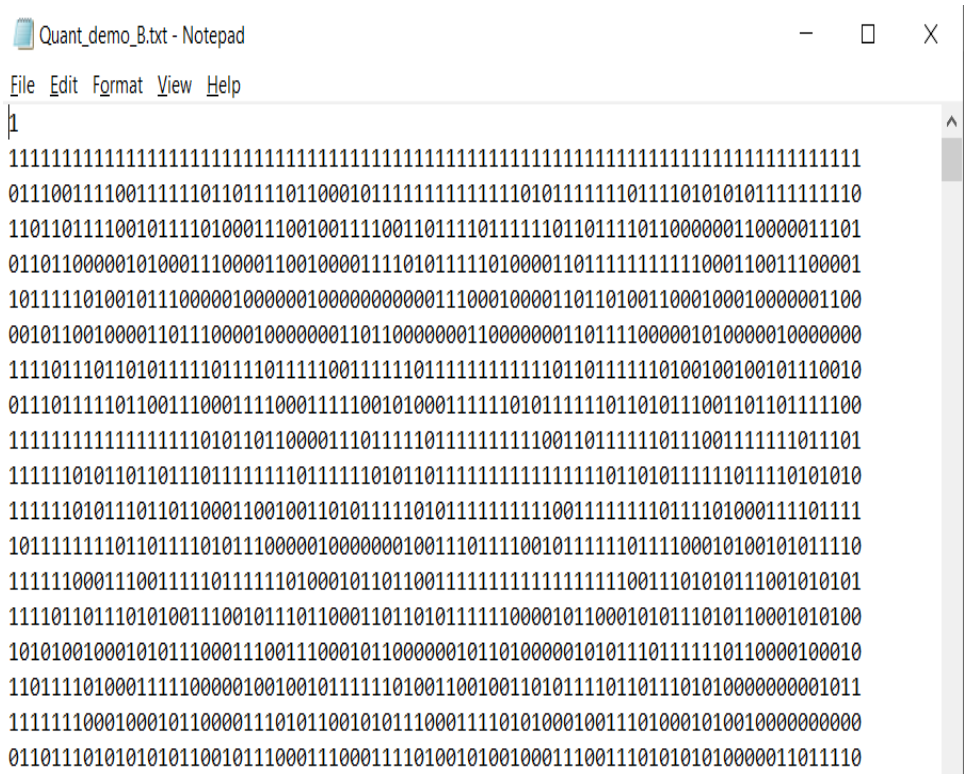


Figure 11.1: Date set-up full information data: The first line of the data contains the number of lines per examinee, and the second line is the key line. The remaining lines indicate the actual choices made for a test taker.

which answer is correct. The following lines contain for each test taker the index of the answer that was chosen. These index values are the integers from 1 to the total number of answers. One of the "'answers" may indicate that the question no answer was chosen or that the choice was in some other way not able to be identified.

We have to input the the number of lines and the key data in separate windows or files, 24 April 19

The stand-alone version of TestGardener can transform the full-information data into binary data based on the users choice. But the current version of web-based app does not yet have this function. You need to transfer the data into binary form (see Figure 11.2) using Excel, R, Matlab, or other software before to run the web-based TestGardener.


Quant_demo_B.txt - Notepad
File Edit Format View Help
1
11111111111111111111111111111111111111111111111111111111111111111111111111111111
01110011110011111101101111011000101111111111111110101111111011110101010111111110
11011011110010111101000111001001111001101111011111101101111011000000110000011101
01101100000101000111000011001000011110101111101000011011111111111000110011100001
10111110100101110000010000001000000000001110001000011011010011000100010000001100
00101100100001101110000100000001101100000001100000001101111000001010000010000000
11110111011010111101111011111001111101111111111110110111111010010010010111001 0
01110111110110011100011110001111100101000111111010111111011010111001101101111100
11111111111111111110101101100001110111101111111111001101111110111001111111011101
11111101011011011101111111101111101011011111111111111110110101111110111101010 10
11111101011101101100011001001101011111010111111111100111111110111101000111101111
10111111110110111101011100000100000001001110111100101111110111100010100101011110
11111100011001111101111101000101101100111111111111111111001110101011100101010 1
11110110111010100111001011101100011011010111111000010110001010111010110001010100
10101001000101011100011100111000101100000010110100000101011011111101100001000 10
11011110100011111000001001001011111110100110010011010111101101110101000000001011
11111110001000101100001110101100101011100011110101000100111010001010010000000000
01101110101010101100101110001110001111010010100100011100111010101010000011011110

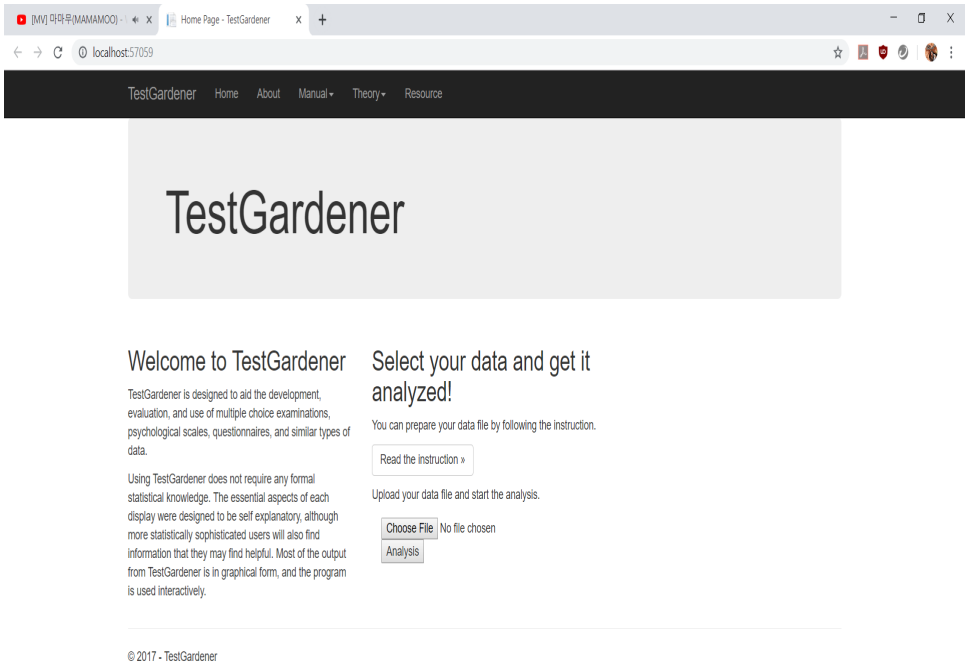


Figure 11.3: Home Page of TestGardener. If you wish to launch the analysis if a new set of data, click "Choose File".

11.3 A Page by Page Description of Test Gardener

11.3.1 The TestGardener Home Page

Once you have your data ready, launch TestGardener by going to `http://testgardener2019.azurewebsites.net/`. Figure 11.3 shows the Home Page.

What do the choices mean? What are the consequences of each choice? Another couple of words in each button?

You can click the Choose File button to choose your data file, and then click the Analysis to have your data analyzed.

How does this work? Do you have to visit the home page twice?

11.3.2 Data Analysis and the Display Choice Page

Once the analysis is finished, the site will jump to the display page as seen in Figure 11.4 with a list of options. Below, we will show you an example of each plot and briefly explain each plot. And we will show the plots of binary data and full information data side by side.

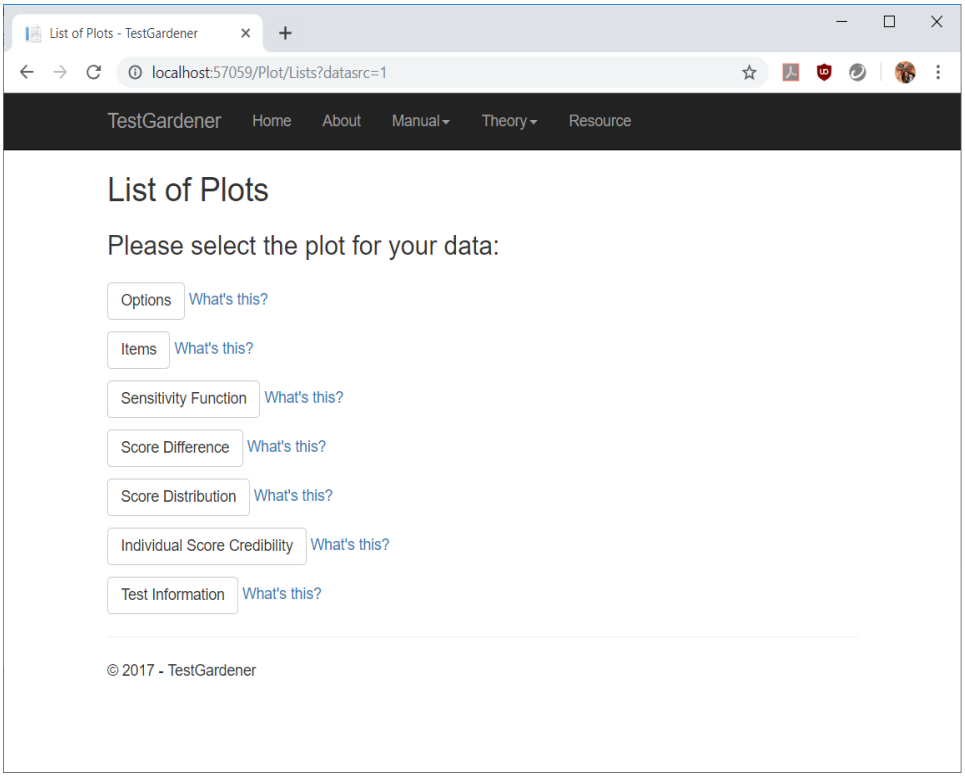


Figure 11.4: TestGardener Display Page List of Plots

Before this subsection, we have to talk about which display or score index variable will be used in the plots, including the default choice.

11.3.3 Plotting the Answer Choice Probabilities

TestGardener uses the the term *item* to. refer to a question and the term *options* to refer to the answers available for each question. The Options Plot displays the evolution of the probabilities of choosing the options over the display variable, referred to in this book as the "score index. These curves are referred to by psychometricians as *item characteristic curves*, abbreviated *ICC's.*

For binary data, the plot shows two curves: the right and wrong response respectively; and in full information data, it shows curves of all the options, including the extra option indicating non-response or spoiled response. It displays the probability of examinee at a certain ability level choosing the corresponding option. In a well-constructed test, the probability of choosing the right option should increase along the score index axis, while the probability of choosing other options should go down, as we see in Figures 11.5 and 11.6. The red vertical lines in Figure 11.5 and Figure 11.6 indicate the score index values below which 5%, 25%, 50%, 75%, and 95% of the test takers fall. Statisticians refer to these score index values as *score index quantitles.*

You can review the ICCs of all the items by clicking Previous and Next, or type in the item number and click OK. The current plot can be saved by clicking Save Image. And when you finish reviewing all the ICCs, you can go back to the list by clicking Back to List.

Can we put numbers at both sides of the graph to indicate the option index?

Plotting the Item Probabilities

The Items Plot are also ICCs, but just for the right option. When you click Items in the list, probability curves of all the items will be plotted, as shown in Figure 11.7. And you can also review the curve of a particular item by inputting the corresponding item number, see Figure 11.8. In binary data and full-information data, at a particular score level, the probabilities of choosing the right options are the same. Therefore, you will find the same Items plot in both cases.

Item Sensitivity Function

Item Sensitivity Function is used in the calculation of optimal score. The value of sensitivity function indicates the amount and direction of the information about the optimal score index provided by this item. As for the Items plot, the sensitivity functions for all the items will be shown first, and then the user can go to any item of their interest, see Figure 11.9 and 11.10. The sensitivity plots for binary data and full-information data, of the right options, should also be identical.

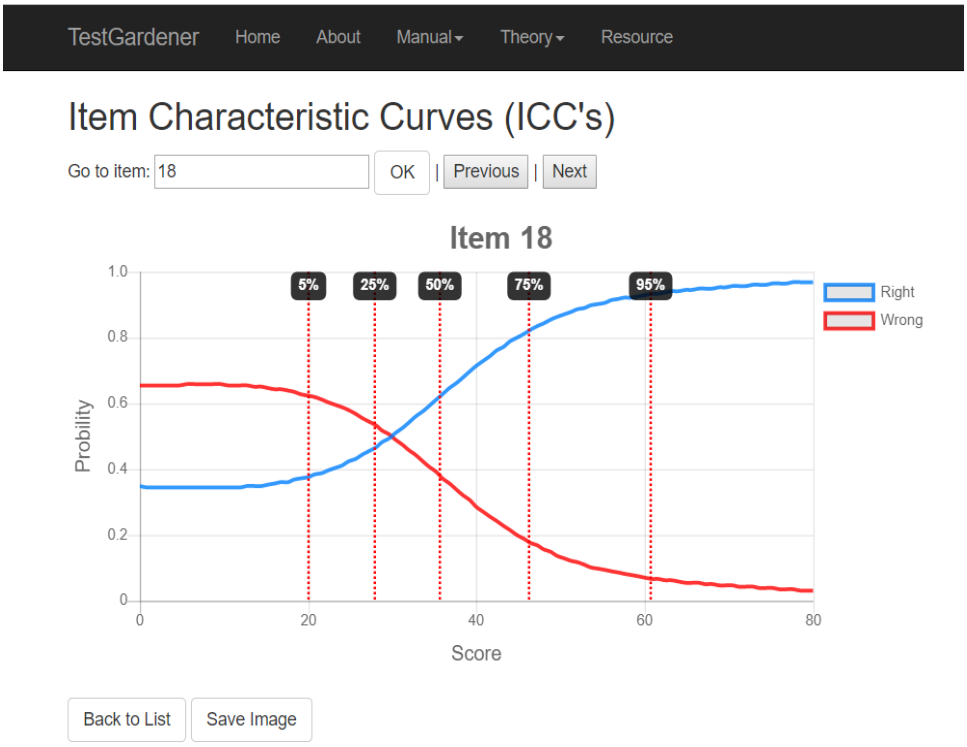


Figure 11.5: TestGardener Display the Option Plot for binary data. The blue curve displays the probability of choosing the correct answer and the red curve the probability of choosing the wrong answer.

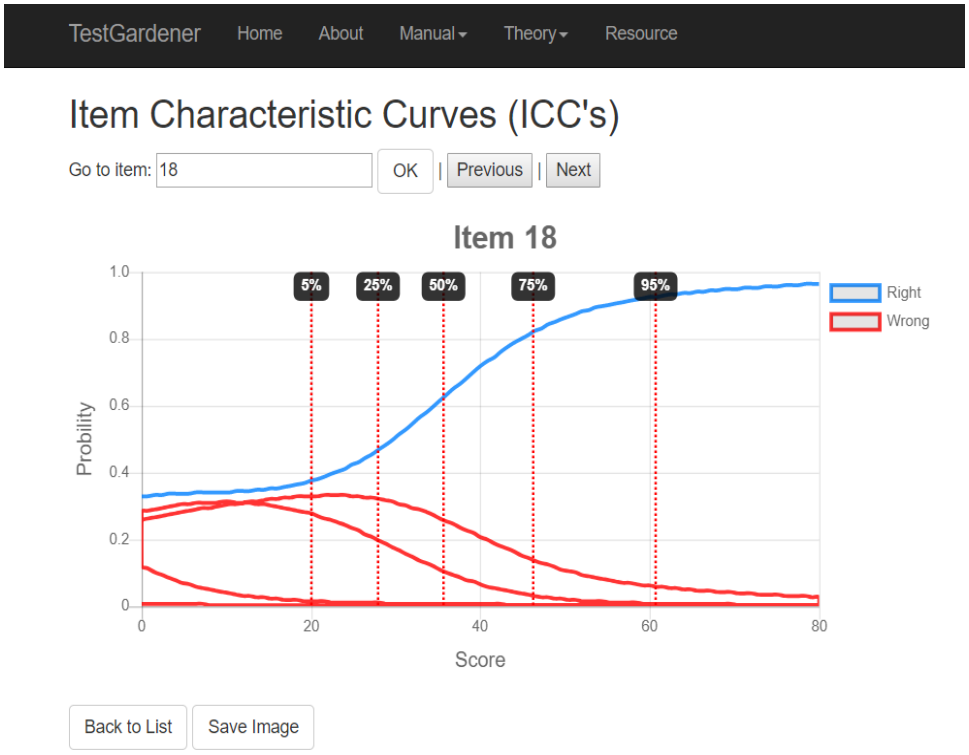


Figure 11.6: TestGardener Display The Option Plot for full data. The blue curve shows the probability of choosing the correct answer and the the corresponding probabilities for the wrong answers are shown in red.

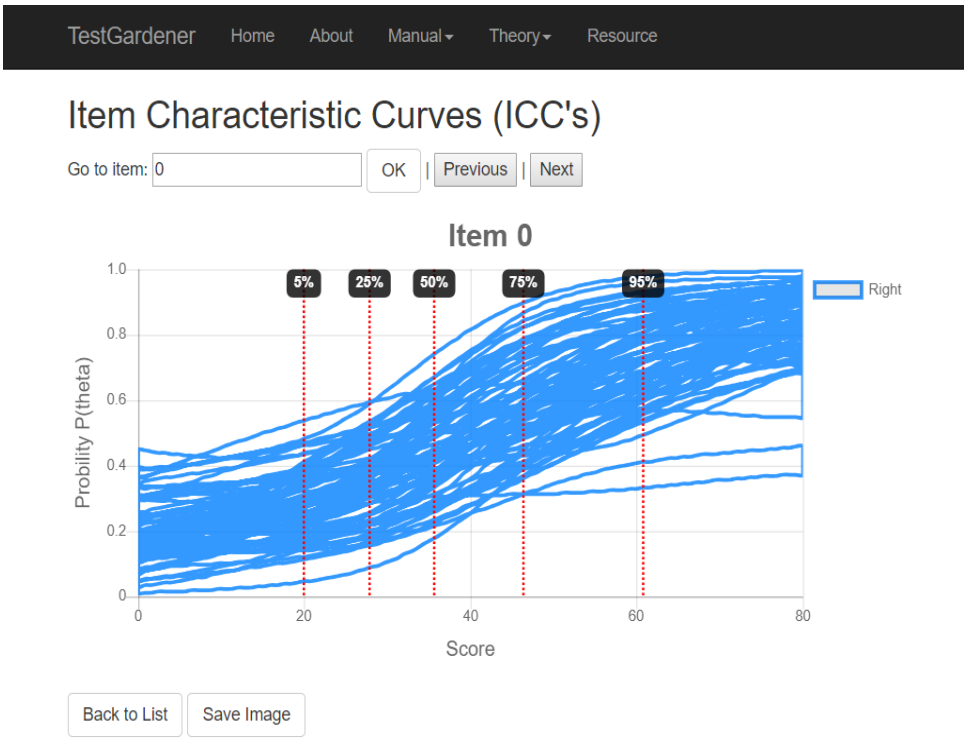


Figure 11.7: TestGardener Display Plot of the item characteristic curves for all the items. The ICC curves are for the correct answers only.

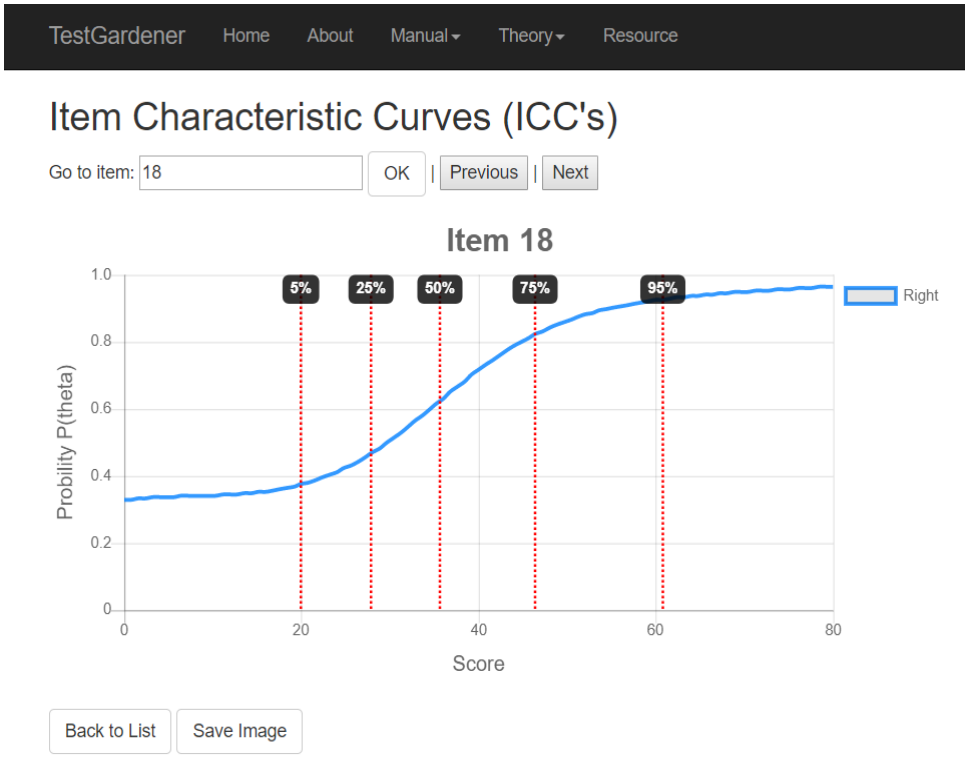


Figure 11.8: TestGardener Display Item Characterisrtic Plot of item 18.

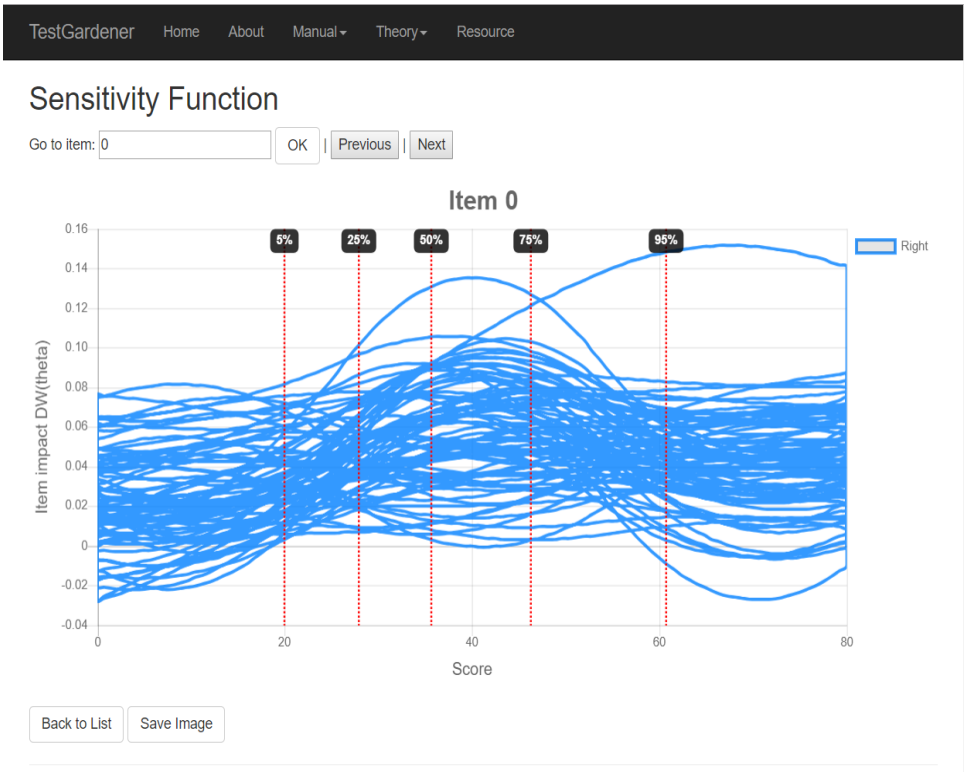


Figure 11.9: TestGardener Display Sensitivity Function Plot for all the items.

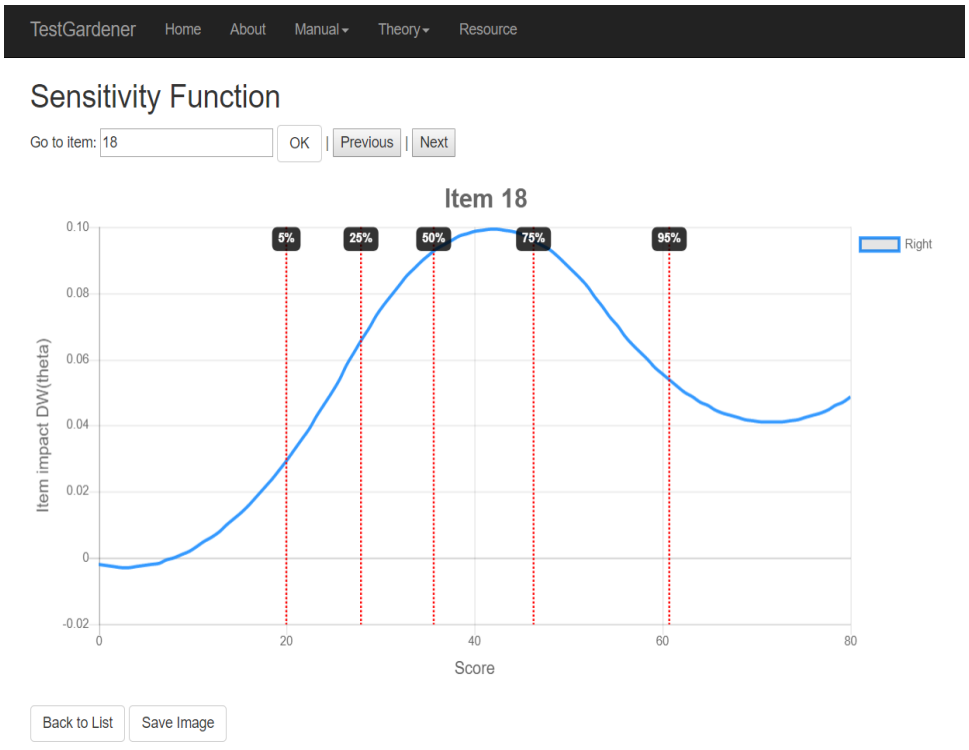


Figure 11.10: TestGardener Display Sensitivity Function Plot for the correct answer for item 18.

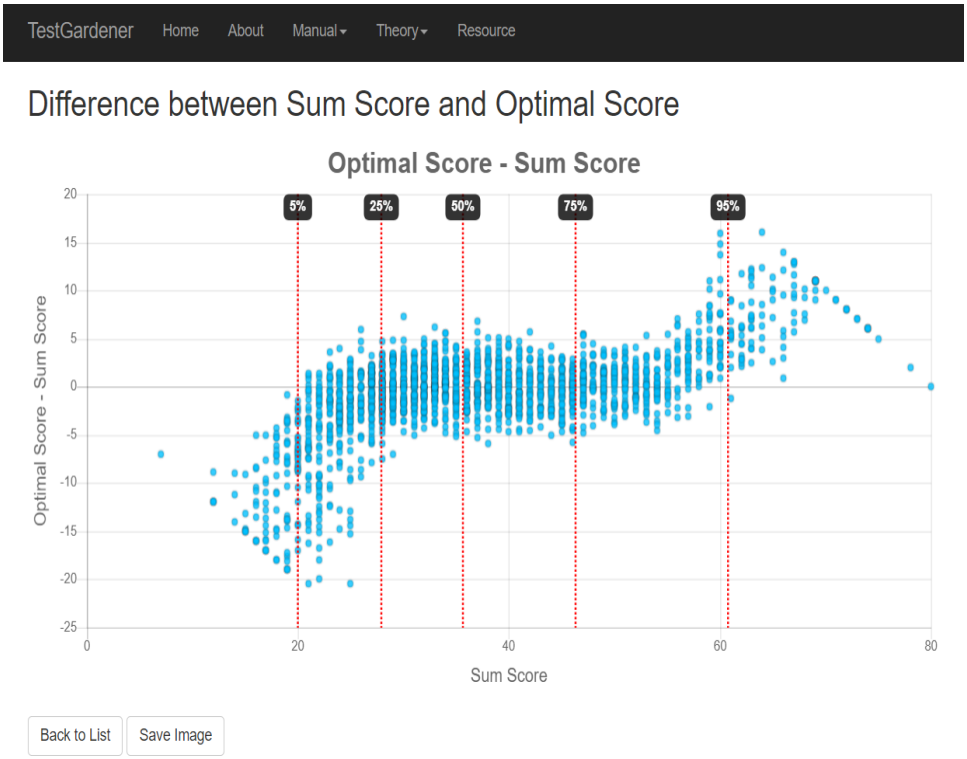


Figure 11.11: Binary data

Add plots of wrong answer sensitivity curves as well. Or perhaps display these by default.

11.4 Score Difference

The scatter plot in Figure 11.11 shows the difference between optimal score and sum score on the vertical axis and the sum score value on the horizontal axis. As shown below, optimal score corrects the positive bias of sum score at the lower end and the negative bias at the upper end.

11.5 Score Distribution

You can also see the comparison between sum score and optimal score using their distribution plot, as seen in Figure 11.12. The frequency of each score range is illustrated by the blue bars, and the distribution of scores in terms of a smooth curve called a probability density function indicating the relative probability that various score values will occur.

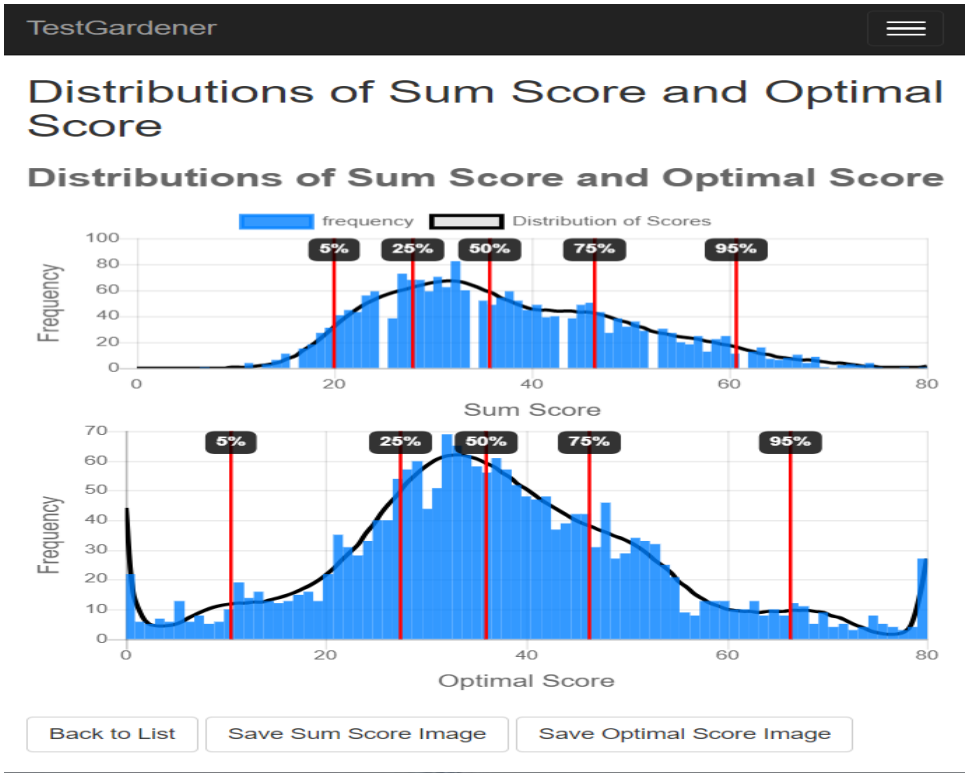


Figure 11.12: Binary data

Figure 11.13: TestGardener Display Individual Score Credibility Plot, binary data.

11.6 Individual Score Credibility

We shall now have a look at some examinee displays. That for examinee 1 is shown in Figure 11.13.

- The solid blue curve shows the relative likelihood or probability of this examinee's true proficiency level being at various values. It can be seen that, on the basis of the examinees option choices, wrong as well as right, it is very unlikely that his/her true proficiency is outside the range of 56 to 76 (the 95% confidence interval indicated by the two vertical black lines). We can also note that the most likely value, where the curve reaches the peak, is about 66, i.e. the optimal score. This is called the maximum likelihood estimate of proficiency.

- The vertical red line indicates the examinee's sum number of correct items. Note, however, that the maximum likelihood estimate also takes account of whether the wrong answer options chosen are typical of more proficient examinees or not. In the case of this examinee, his/her wrong option choices suggested that the true proficiency is about 6 points higher than the observed number correct.

Figure 11.14: TestGardener Display Test Information Plot.

11.7 Test Information

Figure 11.14 is the Test Information Plot. The test information function indicates the amount of information in the test about proficiency at various proficiency levels. It is like the sensitivity curve for a specific answer, but in this case applies to the entire test itself. We see in this plot that the test provides more information for the test takers between the 50% and 75% quantiles than for those at the extremes of the score index continuum.

References

Bloom, B. S. (1956). Taxonomy of educational objectives. Vol. 1: Cognitive domain. New York: McKay, 20-24.

Braun, H. I. and Holland, P. W. (1982). Observed-score test equating: a mathematical analysis of some ETS procedures. In P.W. Holland and D.B. Rubin. *Test equating*, volume 1, New York: Academic Press.

Crocker, L. and Algina, J. (1986) *Introduction to Classical and Modern Test Theory.* Harcourt Brace Jovanovich College Publishers: Fort Worth.

González, J., and Wiberg, M. (2017). *Applying test equating methods using R.* Cham, Switzerland: Springer.

Kolen, M. J., and Brennan, R. L. (2014). *Test equating, scaling and linking: methods and practices.* (3rd ed.). New York: Springer.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into practice, 41(4), 212-218.

Lord, F.M. (1980) *Applications of item response theory to practical testing problems.* New York: Routledge.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.*

von Davier, A. A., Holland, P.W., and Thayer, D.T. (2004). *The kernel method of test equating.* New York: Springer

Wiberg, M., Ramsay, J. O., and Li, J. (2019). Equating test scores with optimal scores. *submitted manuscript.*