Bureau 42 Summer School 2011: Assessment and Education

W. Blaine Dowler

May 28, 2011

Contents

Author's Preface

The author would like to thank his proofreaders: Rob MacDonald and Dorothy Wassenaar.

Chapter 1

What Is Assessment, and Who Is It For?

1.1 Introduction to Assessment

Assessment is one of the most important aspects of an education system. However, many of the world's education systems do not have the funding or equipment to perform assessments the way they should be done. Even when new and preferable systems are developed, they are not always implemented due to the time and financial costs of educating parents and students about the new methodologies. Many of these methods will be detailed in this nine lesson course, so that parents, students, and educators reading it will better understand the options and the need for changes to be made worldwide.

1.1.1 What is Assessment?

Most commonly, assessment is the measuring and reporting of student achievement or ability. Assessment techniques can be used for many other purposes, such as employee performance evaluations and more, but the most common uses are in the field of education, and that is the series of applications this course will focus on.

1.1.2 Why Do Assessments?

Assessments are done for a few broad purposes. The main purposes are as follows:

- To give a general picture of how successful a student was in a particular course.
- To inform students of their strengths and weaknesses.
- To screen candidates for a particular placement at work or school.
- To evaluate the effectiveness of instruction.

There is no single assessment tool or reporting technique that does all these jobs well. The current public school systems in North America tend to focus on the general picture (grades) and the screening (university and job placement) while, unfortunately, leaving out the tools detailing a particular student's strengths and weaknesses. This highly important aspect of assessment is the one students need to improve their grades and skills in the future, and it is the goal least effectively served by traditional grade reporting methods in North American public schools. Little Johnny may come home with a 68% on his report card, but that report card probably doesn't say which skills are among the 32% of the course he can improve on.

1.1.3 Who Are Assessments For?

Assessments can be done for a number of individuals. The most common individuals are students, parents, teachers and school administrators, and prospective employers or educational institutions. Although there is certainly overlap in the goals of these groups, there are goals that are distinct to each of them.

- **Parents and Students:** Parents and students are the groups most likely to have overlapping goals. (In most cases, though certainly not all, differences arise when the parental goals are loftier than the student goals.) The primary goals of these groups are to make sure the student is equipped with all the skills required to succeed in the future. Thus, though the prevailing opinion among parents and students is that overall grades are the most important reports, research and logic dictate that a skill-based reporting scheme is, in fact, the most effective reporting methodology.
- **Teachers and School Administrators:** Many teachers look at grades the same way parents and students do, with the intent to help student

performance improve. However, there is an additional purpose for assessment, and that is the evaluation of instructional effectiveness. If a class usually averages 65% on unit exams in a particular course, and then averages 50% on a quiz based on a particular lesson, this is a strong indicator that the lesson delivered was not very effective, and needs to be revised and/or retaught. School administrators can also use overall class averages on assessments to determine the effectiveness of teachers as a whole, and determine which teachers are best suited to the course assignments. This can be difficult in situations in which individual teachers produce their own assessment tools. If each teacher creates his or her own unit exam, then the marks may not be directly comparable, as the exams may not place the same emphasis on the same skills. Standardized assessments, be they standardized throughout the school, district, state or province, country or world tend to be the preferred assessments for this purpose.[1]

- **Prospective Employers and Educational Institutions:** This group has their own goal in mind: filtering out the "best of the best." This is a completely different type of evaluation. Rather than looking at individual skill performance, they are more concerned with how prospective employees or students performed relative to the class average.

Each of these goals is best served by a different reporting scheme.

1.2 Types of Assessment Items

An assessment item is an individual task or question used to evaluate student performance on a skill or group of skills. By "item," we typically refer to a single question rather than a complete test. Assessment items can be grouped into two broad categories of objective and subjective items.

1.2.1 Objective Assessment Items

Objective items are the questions which have a definitive right or wrong answer, which means these are the questions which are eligible for machine-scored assessments. The most common questions of this type are now listed, in order of difficulty levels for assessment item creation.

[1] No system is perfect, of course. A teacher who is aware of this evaluation scheme may be tempted to "teach to the test" in this situation if the test can be accessed in advance, thus preparing students for that assessment tool rather than for the general course outcomes. More details on this will follow in lesson three.

- **True or False Questions:** These questions are amongst the easiest to create. The possible responses are standardized for the entire section: it is either true or false. Students also rarely struggle with the instructions, and their responses are easy to mark. Unfortunately, because there are only two options, it is relatively easy for students to guess correctly. This is further compounded when the item is not well written, and specific determiners such as "never," "every," and "all" are used. When those words are in place, the correct response is often "false," simply because there are exceptions to almost every rule.[2]

- **Fill in the Blank Questions:** These are only slightly more difficult to write than true or false questions. One simply thinks of a true statement, and them omits a word or term. The student must then complete the sentence. Marking is again efficient, and it is far less likely that the student answers the question correctly through a random and uninformed guess. However, as responses to these questions are typically hand written, they are difficult to machine score in a school not equipped with OCR technology, and need to be marked by humans. The human marking is very efficient, so it is typically not a problem when dealing with a single teacher's course load, but it gets dramatically more cumbersome with assessments at the school or district scale, such as year end finals.

- **Matching Questions:** These questions allow for a move diverse set of true/false question structure. The student is faced with a list of prompts as well as a list of possible responses, and needs to match one to the other. Again, marking is absolute and efficient, but there is a chance of guessing correctly. The odds of guessing correctly decrease as the number of options increases, and the odds increase when the matching is one-to-one, meaning every prompt matches a response (as they all should) and every response matches exactly one prompt. In those situations, a student who can correctly identify the responses to, say, nine out of ten prompts will likely match the remaining prompt to the remaining response, scoring 100% with an understanding of only 90%. This is further compounded when responses are not of the same type. For example, imagine a history assessment which attempts to evaluate a student's ability to match ten treaties to the years they were signed. If both the prompts and responses include years and treaties, then the odds of guessing correctly increase. In this situation, the best manner in which to measure this outcome with minimal odds of guessing in multiple choice would be to use the treaties themselves, sorted alphabetically,[3] as the prompts. The possible responses would then include all ten correct years, as well as a number off additional

[2] The content of that sentence is the reason I included the word "almost" before the word "every" in that sentence.

[3] Human attempts to randomly order the treaties would result in subconscious associations, or in deliberate attempts to distance related treaties, so they wouldn't be random. An alphabetical sorting choice would randomize them effectively without providing any hints as to associations.

years from similar points in history.

- **Multiple Choice Questions:** These questions are the hardest objective test questions to create effectively, but they are amongst the easiest to mark, as most schools have the technology required to machine score the options. Again, the odds of guessing correctly are not zero, but four or five response options make this less likely. Also, if options "all of the above" and "none of the above" appear on every single question in which they might apply, odds of guessing are reduced. (Most teacher created tests using these options tend to use them only when they are the correct responses.) Another tendency is to lack detail in the incorrect responses. As a student, if one is unsure about the correct response on a multiple choice assessment, selecting the longest answer tends to work out well. As a teacher, the greatest difficulty in creating multiple choice items is creating plausible distracters, meaning incorrect answers that are likely to appear correct to unsure students. (A well written distracter will tend to draw the attention of the weaker students, meaning the lowest performing students in a group have a strong tendency to choose that response instead of the correct one.)

Virtually all objective tests share a drawback: students are more often required to *recognize* responses than create them. These item types tend to be effective for many science and math assessments, but they are not very effective for language arts and social studies, where opinion based responses are more common, and must be generated by the students rather than the teachers.

1.2.2 Subjective Assessment Items

Subjective assessment items tend to be easier to create than objective assessment items, but can be far more difficult to mark. These include short answer and essay questions, in which the students must present ideas of their own. The prompts tend to be short, but effective scoring is far more difficult to achieve. For example, if the curricular outcome is "write a short story," and the teacher marking the results tends to prefer romance to science fiction, then two well made examples from the two different genres may get two different marks as a result of the teacher's subconscious bias rather than the product on the page.[4] However, these are the only assessment items which allow students to create and present their own ideas, so they are the only assessment items which can measure some curricular outcomes.

[4]This is why so many state, province or country wide standardized high stakes tests are marked by multiple individuals whose scores are averaged together.

1.3 How Students Learn

Students learn in stages. Some outcomes and skills are easier to learn than others. These were detailed most effectively by Benjamin Bloom in what is called Bloom's Taxonomy. As students learn, they progress through six stages, assuming the skill is eventually mastered. If a student only progresses through three or four of those stages, then that student is more likely to regress in that skill area over time if that skill is not applied. Once the sixth level of the taxonomy is reached, the skill is truly mastered, and can be far more effectively retained over time despite disuse.

The six levels of Bloom's Taxonomy are as follows:

1. **Knowledge:** This is the lowest and easiest level to obtain, and the one which is the easiest to reach with objective assessment items. It only requires fact recognition and recall. For example, a student who does not understand that multiplication is the repeated act of addition may still memorize that $3 \times 4 = 12$, and answer such a question correctly with no true understanding of why 12 is the correct answer.

2. **Comprehension:** This is the next level, in which the student understands why the knowledge is true. This is the level often reached by students, particularly working at home with their parents. As the mass populace is not typically instructed in Bloom's Taxonomy, parents are often confused by the struggles students have when they reach this level but do not perform in school, or do not retain that information for tests. They begin to question student effort or physiology, as those are the causes for the struggles they are already familiar with, and may not identify the real cause of the problems: a lack of progression through the complete taxonomy. This is more pronounced when the parents make up questions to quiz students with while studying, as these questions are typically presented at this level.

3. **Application:** This is the next most effective level to reach. Students can then apply the knowledge to unfamiliar situations. For example, if a student has learned about multiplication as a mathematical abstract, and is then asked how many apples there are in the back of a truck with 30 crates of 144 apples each, then the student answers correctly by moving through the application level.

4. **Analysis:** This level is one in which the student examines the underlying processes of a skill and completely understands how they work. This is typically demonstrated by comparing multiple correct answers and choosing the best answer, or developing the best answer. This is the highest level of Bloom's Taxonomy which can be effectively evaluated with a machine scored or objective test, as all higher levels require responses to be

created by the student rather than the test creator. Students typically complain about the "choose the best answer" questions because the students have not been taught about Bloom's Taxonomy, and their self study habits have only taken them to the application level.

5. **Synthesis:** This requires a level of convergence. Students need to take multiple skills and combine them to produce a new concept or approach not previously seen in class. Students tend to complain about these questions, too, with the argument "we were being tested on stuff the teacher never even taught us!" If the curricular outcome requires this stage of understanding, questions that challenging will appear.

6. **Evaluation:** This is the highest level of understanding. At this level, students are ready to evaluate their own ideas, and the ideas of others, to determine which are the most effective. If a student understands a skill at this level, the student will likely retain the skill for life.[5] As such, this is the level both students and teachers should strive for, particularly in the critical areas of reading and arithmetic.

1.4 Upcoming Lessons

Through the rest of this course, we will analyze the assessment tools needed for each of the primary purposes, with particular emphasis on how teachers and assessment administrators can ensure that the assessment items they use are appropriate to the goals at hand.

[5] Retention is absolutely critical for student success. If a skill is not retained, then it will not be present when a subsequent skill is taught, and that subsequent skill cannot be retained. For a rather obvious example, imagine trying to teach multiplication to a student who doesn't understand addition. If the foundation skill of addition is not present, the student cannot learn multiplication. This is the inherent cause for students to continue struggling after hiring tutors or getting extra help from the teacher outside of class: if a student is struggling in grade six, the problem is usually because that student has not retained a prerequisite skill from grade five, or four, or three, and so forth. When a typical tutor or teacher gives additional help on the grade level material, the prerequisite skills are not being rebuilt in the student's foundation, and the new skill will not be retained in the long term either. The student may be able to learn how to go through the motions and answer the homework, but that's operating at the first level of Bloom's Taxonomy only. The student will probably struggle with the same questions when the unit exam or year end final comes around.

Chapter 2

Broad Types of Assessment

2.1 Classifying Assessments

In the previous lesson, we established that assessments are performed for different people and with different goals. The specifics of some of these differences are detailed in this lesson.

2.2 Education Stage Progressions

One way to divide assessments is by the stage in the education process at which they are administered. These are broadly grouped into *formative* assessments, in which students are still forming their understanding of the skill in question, and *summative* assessments, in which students are expected to have already learned the material, and are simply being assessed on where they stand with it.

2.2.1 Formative Assessment

Formative assessments can be marked in a variety of ways, and their use is a point of debate. The most common formative assessment is homework.

When a new topic is taught, the knowledge and comprehension levels of Bloom's Taxonomy are expected to be reached from the lesson alone. These

levels are verified, and the application level is reached, through daily homework assignments.[1] Homework is a hot topic of debate right now, with a few questions surrounding it.

1. *Should homework contribute to the report card grade?* In an ideal world, homework would not contribute to final report card grades. A final report card grade should reflect the student's ability to apply and demonstrate understanding of the skills and curricular outcomes at the end of the course. Homework is a formative assessment; this is the opportunity students have to test their initial understanding, gain feedback from the teacher, and identify and correct their misunderstandings. As such, the understanding students have when the homework is first completed is expected to be less than the understanding they have when the course ends.

   Sadly, we do not live in an ideal world. If homework does not contribute explicitly to the report card grades, there is a natural human tendency among the students not to do the homework. Without doing the homework, the typical student will not pass the second level of Bloom's Taxonomy, and the skill will not be retained. The practice required to succeed at the end of the course is not obtained. Using homework to contribute to the report card grades reduces this problem to a degree, but then the report card grade will not necessarily reflect the student's understanding at the end of the course.

   There is no clear cut answer to this question, though the current trend is to record homework results on a completion basis rather than marks, which may or may not be worth a small percentage of the final report card grade.

2. *How much homework is the right amount?* Traditionally, the value of practice has been well known, but the application has been through assigning large amounts of homework on a particular topic the day that topic is taught in class, and then never assigning homework on that topic again.

   This doesn't work very well.

   Research has shown that doing more than 5-10 problems of a particular type in a 24 hour period is not beneficial to the student. Thus, giving out 10 math questions today will be about as helpful to the students as giving 30 or 50 today. Now, that's not to say doing 30 questions total isn't useful,

[1] For reasons the author has never understood, homework rarely exceeds the first four levels of Bloom's Taxonomy, despite the fact that unit and year end exams typically do. It is the author's belief that all six levels should be challenged on the homework as well, so that students are better prepared when they are challenged with these levels on exams. This is particularly confusing in cases in which the textbooks provide questions at all six levels of the taxonomy, and homework assigned from the textbooks is deliberately chosen to omit the higher level questions because "they are long."

but doing, say, six questions for homework every night *for five consecutive nights* will be far more helpful when trying to ensure retention. Spreading out the topics like this not only encourages retention, it provides review questions which students will find less time consuming, and it helps them make connections between consecutively taught and related topics.

3. *Should homework be taken home?* If a student does have a misconception or misunderstanding after instruction, doing homework can turn that misconception into a habit. Once a habit is formed, it can be very difficult to break. Furthermore, with a dishonest student, there is no longer any guarantee that the student taking credit for the homework was the individual who actually did the homework. This is particularly true with students who are struggling to levels at which they are frustrated. Most parents don't like to see their children in emotionally stressful situations, and some will do the homework for the child instead of watching the student suffer.[2]

Research is indicating that the answers to these questions require breaks from tradition. This can cause a lot of friction between teachers and parents, particularly when the parents in question are those who were successful in the traditional school system. They need to be convinced that the changes are truly in the best interests of their children. Nothing will raise a parent's ire more than endangering his or her child's future, so changes in this area have been slow to come to avoid these confrontations.

2.2.2 Summative Assessment

Summative assessments are the ones used at the end of a unit or course, when the instruction on skills assessed has been completed. In an ideal world, in which emotion has no impact on student performance, these would be the only marks used for the final report card grades, simply because these are the marks that indicate assimilation of information after instruction.

Again, the real world is not ideal.

The problem here is test anxiety. Every individual has an optimum stress level. If a person feels too little stress, that person's performance is poor. As stress builds, the individual's focus improves, and performance improves. If the stress level is too high, performance crashes hard. When summative assessments are worth larger and larger proportions of a student's mark, the stress level in that student increases, particularly when that student has future goals that depend on the marks in that course. The student stress goes beyond the

[2]The real solution to this situation is to go back and fill in the foundation from previous years, but I have yet to see a public school system that does this effectively.

ideal, and the performance crashes. Students who suffer from this test anxiety will then have summative assessment marks that are less indicative of their understanding than their formative assessment marks.

It should be noted that true test anxiety will impact multiple courses, and often impacts all of them. If a student hasn't reached the highest levels of Bloom's Taxonomy in a course, then that student will have poor retention in that particular course. The student may also have higher homework marks than exams in this case, particularly if additional support is provided for homework, because the student will be "going through the motions" on the homework without understanding why those motions are correct. Information will not be retained for the exam, and student performance drops once more. If exam marks are lower than homework for one student in one class, this cause is more likely than test anxiety.

2.3 Measurement Type Divisions

Final grades on an assessment or in a course can be reported in two primary ways, depending upon the intended goals for the reporting. The two main categories are norm referenced and criterion referenced reporting.

2.3.1 Norm Referenced

Norm referenced reporting is designed to report on a student's performance *relative to* his or her classmates. The report does not indicate the proportion of the course that the student succeeded with.

The advantages to this method of reporting apply almost exclusively to prospective employers and educational institutions. Those entities are primarily concerned with identifying the "best of the best," and compare the performance of one student to a group of peers. Those peers may have been chosen by age or grade, and they may or may not have met. These assessments are also frequently machine scored, which makes marking them remarkably efficient.

The disadvantages cannot be ignored. First of all, a student can see the results of this assessment, and know that he or she did "average." What does that mean? What was the average? Who was in the norming group? Which specific skills must a student improve in to increase this performance?

The basic methodologies for norm referenced assessments are pretty universal, and must be understood in order to correctly interpret the results of

such an assessment. In most cases, no single school can provide a population base large enough to generate reliable statistics for these assessments, so they draw students from multiple schools in multiple districts. The assessment is administered to these students before it is released to the public.

The results of all of these students are then compiled, and collated by total score. Individual assessment items are evaluated for quality using criteria that will be described in detail in later lessons, bad items are rejected, and student scores on only those assessment items kept are then compiled and compared. Performance of students on this assessment tool in later years is determined by determining where they fit relative to the members of this original "norming" group.

This is not a particularly accurate system in many cases. Often, in order to use enough students to build the reference results, the test is administered to a random sampling of schools in multiple regions using different curricula. In that case, a student's performance relative to the entire group may differ from his or her performance within the local regional group. As a result, a student's reported performance may not align with the local standards.

Things get worse when looking at improvement. Effective assessment items are difficult to write. As a result, students either cannot see or cannot keep test papers after they are written, which makes it difficult to analyze errors and improve future performance.[3]

The final problem is one that, sadly, seems rather pervasive in the field of education. When building a norm referenced assessment tool, one works by comparing the data obtained to one or more models. Unfortunately, it appears to be a common practice to decide on a model in advance, and reject data that doesn't fit the model rather than adjust the model to fit the data. The models used are good enough to apply to the majority of the population, but as student performance drifts further and further from the average performance of a student, the model fits less and less well. Ultimately, the results of the population's highest and lowest performers tend to be rejected and omitted from assessment results used for comparison.

For example, think about IQ tests. In North America, IQ tests try to measure a person's overall intelligence by modeling test performance on a bell curve (also known as a normal or Gaussian curve) with an average of 100 and a stan-

[3]Note that, in many areas, such school policies may be illegal. Technically speaking, the student is the author of any document he or she produces. In most regions, this means the schools cannot legally deny a student request to receive a copy of any work he or she has handed in. Note further that this only applies to the actual student output itself, and not to the assessment. If the student is told not to write in the test booklet, and to answer on a machine scored multiple choice form, then the school is only required to provide a copy of the multiple choice form, and *not* the assessment questions that provide a meaningful context for that form.

dard deviation of 15.[4] In short, approximately two thirds of the population will score between 85 and 115 on a North American IQ test. Even if we correctly compensate for language differences that occur (i.e. very smart people who only speak Spanish will get terrible scores on an English-based IQ test) and other regional differences, we have problems with the model that cannot be overcome. A bell curve is completely symmetrical about the average. This means that we would have just as many people scoring 80 points below average as there are scoring 80 points above average. This isn't the case. The number of people with tested IQ scores above 200 is far greater than those with negative scores. This skew distorts extreme results. Most standardized IQ scores will "report" scores as high as 160. However, an individual with an IQ of 145 could easily get the 160 score on most tests, as the average difficulty of the questions is around the 100 IQ mark. Once a person performs far enough above average, the odds of a perfect score increase.

Every norm referenced assessment also has two ambiguous cases. A student who gets every question correct or every question incorrect no longer fits on the scale. A student with a perfect score on an IQ test which accurately measured from 70 to 130 could have an IQ anywhere from 135 to ∞. Similarly, a student who gets every response incorrect cannot be accurately measured (and probably doesn't understand the instructions or know the language the test is written in.)

Norm referenced reporting on larger scales, such as University classes, leads to a new set of problems. If grades are normed and averaged on a class by class basis, then it becomes difficult or impossible to evaluate the performance of individual teachers or a complete course curriculum. The average mark and spread of grades in the class section with the most effective instructor looks identical to the average mark and spread of grades in the class section with the least effective instructor, simply because they are forced to conform to the same scale.

Research has shown that, when the goal of assessment is to communicate with students and inform them of their strengths and weaknesses, norm referenced reporting is the single most *ineffective* reporting method developed in recorded history.

2.3.2 Criterion Referenced

Criterion referenced assessments do not depend on mathematical models or on the performance of large groups of people. Instead, they measure a student's performance in direct comparison to the curriculum of skills being assessed.

[4] European IQ tests typically use the 100 average, but often use a standard deviation of 10 instead.

Detailed criterion referenced assessment is the type of assessment that benefits the students the most.

Criterion referenced assessment is well suited to reporting broad pictures of a student's performance in a complete course, or to reporting how each student is performing with individual skills. Less detailed versions include letter grades and percentages; students will know how much of the unit or course they need to improve, but not which specific aspects need to be reviewed. More detailed versions which give a skill by skill breakdown, or which include personalized teacher comments, tend to give the students the best possible chances to improve future performance by directing them to the exact skills which need to be improved. The drawback is that it requires a considerably higher amount of time to mark. Given the current trends in North American class sizes, teachers simply cannot provide this level of information in the time available without obtaining or developing a computerized system that helps them increase efficiency.

2.4 Upcoming Lessons

In lesson three, we begin to discuss validity, reliability and bias of assessment items, which are the three key concepts used to evaluate the quality of the assessment tools themselves. Lessons four and five will discuss methods of norm and criterion referenced reporting in detail, and lessons six through nine will cover increasingly mathematical methods to build assessment tools which can achieve the best of both worlds.

Chapter 3

Validity, Reliability and Item Bias

3.1 Item Quality

An assessment tool is only effective if it is evaluating the skill or curricular outcome which it is designed to evaluate. For example, assume a curricular outcome is "student is familiar with milestone events in superhero comic books published in the 1960s" and one needs to test this outcome on an exam.[1] A question such as

1. Fill in the blank: Peter Parker is also known as _____.
   (a) Batman
   (b) Hulk
   (c) Spider-Man
   (d) Superman

is a terrible way to measure. One can correctly answer this question having never read a superhero comic from the 1960s. In fact, there are millions of people alive today who would answer that question correctly, but who have never read a comic book in their lives. Effective assessment depends on identifying effective items and rejecting the ineffective items. The three primary measures which must be satisfied by an assessment item are *validity*, *reliability* and *bias*.

[1] I don't know which course would have this outcome, but I'd like to take it.

3.2 Validity

The first critical element to an effective assessment item is validity. An assessment item is considered valid if it measures the skill in question. These can be extremely difficult to write when there is more than one way to arrive at the answer to a question. The above comic book example is one that obviously has multiple approaches to the answer: one could watch an adaptation of the character in other media, or read a comic book from another era which contains the same information.

For a more subtle example, assume you are trying to assess a mathematics skill outcome described as "student will be able to multiply a three digit number by a two digit number, with or without regrouping." The question $273 \times 87 =?$ is more valid than $458 \times 11 =?$. Let us examine why that is.

Both questions involve the multiplication of a three digit number by a two digit number, and regrouping is optional. They both appear to satisfy the criteria of the skill outcome. However, in the second problem, the two digit number is 11, which is a relatively low two digit number. There are students who will solve this by adding 11 instances of 458. Thus, it is never actually determined whether or not the student can do the multiplication; the student who uses explicit repeated addition does so because he or she is unsure about how to correctly multiply by two digit numbers. The importance of using the "place holder" 0 in the second line is not understood. However, the first question involves multiplying by 87. This will ultimately be more valid, as there are far fewer students who will add 87 instances of a three digit number correctly. In this case, the validity differences between the two items is not as significant as the comic book question, but there is a definite difference.

Careful item creation can improve the validity of items by simply choosing items in which alternative methods are either difficult to implement or entirely impossible. However, the validity of the question can still be in jeopardy. If the item is a multiple choice or matching item, then it can be invalidated by a lack of plausible distracters. In other words, if the possible wrong answers are clearly wrong, the student can deduce the correct answer with no understanding of the skill. Let's return to the comic book outcome to illustrate this. Look at the following item:

1. Fill in the blank: Doctor Octopus first appeared in _____.
   (a) Amazing Spider-Man #3
   (b) Archie #4
   (c) Batman #7
   (d) drag

The prompt is one that would seem to require a knowledge of superhero comic milestones of the 1960s, which is what we want. However, three out of four possible responses are clearly wrong. Even without the success of the recent movie series, it isn't hard to imagine that anyone taking a course with an outcome such as this would know that Doctor Octopus is most strongly associated with Spider-Man, so the correct answer stands out from the others like a beacon. A better version would be

1. Fill in the blank: Doctor Octopus first appeared in _____.

   (a) Amazing Spider-Man #2

   (b) Amazing Spider-Man #3

   (c) Amazing Spider-Man #4

   (d) Amazing Spider-Man #5

In this case, students would either need to know the correct answer, or know the content of issues 2, 4 and 5 well enough to know that those could not possibly be correct.

The easiest way for an experienced teacher to improve the validity of the plausible distracters in multiple choice items is to determine the most common mistakes, and use those responses to complete the incorrect options. This is particularly important in math and mathematical science courses, in which students may realize their calculated response is incorrect when it is not found in the list of responses. For this reason, math and mathematical science teachers should seriously consider including "none of the above" as an option on every calculational question, and using it often enough to make it a plausible option. Note that overuse of this option can also reduce the validity of a question: students who have so little understanding of a topic that they fail to produce any of the numbers listed as common errors will correctly respond "none of the above" when they do not truly understand the skill.

Of course, it is entirely possible that the teacher is unable to recognize that an invalid item is invalid. As a human being, it is a natural psychology: if one is trying to write an item that can be solved by method A, it may not cross one's mind that another approach is possible, particularly when method A is the most efficient approach. One would need to apply the item analysis techniques of lesson six to determine whether the items truly perform as anticipated on the assessments.

3.3 Reliability

The reliability of an assessment item is similarly important. Reliability speaks to reproducibility. A reliable assessment item is one which will produce the correct response from students of sufficient ability over multiple applications. An item can be valid but not reliable, or vice versa.

Imagine a metaphorical dart board. If a student answers a question correctly, the dart hits the bulls eye. If a student has a misconception, the dart will land outside the bulls eye. If the item is valid, understanding the skill in question will be the only way to hit the bulls eye. If the item is reliable, the darts a given student throws will always land near each other.

Now, imagine an item which is valid, but not reliable. The darts will be scattered evenly across the board. The average position will still be the bulls eye, but the darts will be spread out across the board (and possibly the wall it is hanging from.) An example of a question of this type would be measuring the curricular outcome "student can multiply numbers with several digits with regrouping" by asking "what is $72,695,402,394 \times 42,394,634,630$?" It is virtually impossible to answer that question without understanding the skill, but the number of individual arithmetic operations involved make small mistakes very likely, so students who do have a fair understanding of the skill could quite possibly get it wrong. It is not a particularly reliable assessment item.

Conversely, imagine an item which is reliable, but not valid. Results are reproducible, but not accurate. The darts will still be bunched together tightly, but not near the bulls eye. One common example of this would be the "does your child have ADHD?" questionnaires that are often distributed. The same parent will likely give the same response each time he or she is asked to respond to the questionnaire, but the questionnaires are not accurate diagnostic tools in many cases.[2]

A reliable and valid question will see students with similar averages and

[2] As is probably quite obvious, this is a pet peeve of the author. Varying reports put the number of students misdiagnosed with ADD and/or ADHD at anywhere from 80% - 90% of the students who have been given the label. True ADD and/or ADHD results from an inadequate blood flow to a certain lobe of the brain, and the brain craves more sensory input than it receives. As a result, student attention drifts to look for new stimuli, or in the case of ADHD, the student moves to generate the input cravings that the environment alone cannot satisfy. The true conditions are unmistakeable before the student reaches school age, impact all subjects which do not involve unusual amounts of physical activity, and cannot be reliably diagnosed without a medical professional and possibly a CAT scan. A student who exhibits ADD symptoms in a single subject only is more likely to be struggling in that subject, to the point where maintaining his or her attention on the subject for a full class is frustratingly difficult, and the student's attention wanders to take a mental break or to look for comprehensible stimulus. It is akin to watching a foreign film without subtitles; if you cannot understand all of the content relatively easily, your attention *will* drift.

abilities responding correctly and incorrectly with similar frequencies. If the plausible distracters are sufficiently well crafted from common mistakes, one may even find that students with similar ability levels below the difficulty of the question tend to choose the same incorrect response.

3.4 Bias

Bias occurs in almost all assessments. It occurs when student performance on assessment items varies between two groups of students with equal ability levels. As such, bias skews both the validity and reliability of assessment items, so we need to make deliberate efforts to reduce bias as much as possible.

The most obvious source of bias is language. A student who has consistently gotten marks around 75% in English language science courses from grades K-8 will likely continue to do so. Now, imagine a student who consistently scored around 95% in Spanish language science courses in the same K-8 range. Now this Spanish student (who has been learning English for, say, 6 months) moves to the same school as the first student, and starts taking the same science course in English. The student who used to get 95% will likely see a sudden and severe drop in grades due to this language difference, which is a form of bias.

Now, if a course is in English, it is in English. There is nothing the teacher can do about that. However, the teacher *can* control the level and complexity of English used to teach the course. If the grade nine concepts can be effectively explained using grade six grammar and vocabulary, then using the grade six grammar and vocabulary will reach a larger number of students, despite their difficulties with the language. If assessment items are also written at this level, bias is reduced on the exams.[3]

There are far more subtle sources of bias. The most common sources are cultural, which become particularly pronounced in an ethnically diverse classroom. There are gender biases as well, primarily due to cultural differences. If you walk into a typical North American seventh grade classroom and ask students what C4 is, you will typically find that the male students are far more likely to know the answer than the female students.[4] If you were to ask the same question in a war torn nation, you will likely find the bias disappears, because exposure to the information is far more widespread. This can be difficult to do while still assessing students at all levels of Bloom's Taxonomy. To achieve the highest levels, one must ask students to apply knowledge, skills and concepts to personal experiences, which means the teacher must be careful to select personal

[3] Obviously, this doesn't work in all areas or in all subjects. It's impossible to effectively measure grade nine level grammar using grade six level grammar, for example.

[4] C4 is a type of plastic explosive.

experiences that would be shared by all students. This is extremely difficult to do in a diverse classroom.

3.5 Upcoming Lessons

The concepts of validity, reliability and bias will be explored in more detail in lesson six, at which point we will be equipped with the mathematical tools necessary to perform detailed analyses of specific test items.

Chapter 4

Norm Referenced Assessments

4.1 Norm Referenced Results

One of the two main ways to report student achievement is through norm referenced results. With this style of reporting, students are compared to age or grade peers, and performance is measured relative to this group. Because of this, interpreting norm referenced results correctly can be difficult. These interpretations are easier to make if one understands how norm referenced results are produced in the first place, and work onward from there.

4.2 Norming Groups

In order to produce a norm referenced assessment, one must first find a group of typical students who may be assessed with the tool. The group must be large; to do an accurate statistical analysis, you need a lot of data.[1] Furthermore, you need a spread of data.

Imagine you are creating a pencil and paper assessment tool used to measure performance of grade five students. The industry standard way to label school

[1] As a rule of thumb, one wants 30 students to form a bell or normal curve. This is enough to rate students in a single post secondary course, but when the goal is to build a standardized tool for any peer group, groups need to number in the hundreds or thousands to work effectively.

grade levels is with numbers that have one decimal place. The number before the decimal is the current student grade, while the number after the decimal is the number of months the student has completed in that grade. So, if the new school year begins on September 1, then a grade five student would be in grade 5.0 on September 15, moving up to grade 5.1 on October 1, and so forth, reaching grade 5.9 on June 1.[2] For accurate testing, you would want to know how typical students perform on the assessment at different points in their academic careers, so you would want to administer this grade 5 test to typical students every month of the year to build data.

There is a drawback to this, of course. Logistically, one typically needs to pay a school to use that school's students to build data. Furthermore, administering the same test to the same students every single month skews results, as the students start choosing answers they remember being confident in on the last assessment, or the answers their friends told them were right while discussing it after school the first time, and so forth. Therefore, it is more common to use "seasonal norms" by administering the assessment once per season, and then using mathematical interpolation[3] techniques to "predict" the values in between administrations.

Now, in any group, one will run across students who perform significantly above or below average. There are two ways to deal with this when building a reference. It is common to administer the assessment to students outside the intended grade to get the actual data to model; a test intended for grade five is typically given to students in grades four, five and six to generate the normative data. However, drifting too far from the intended grade level skews results. Administering a grade five assessment to a grade one class is virtually worthless, as students will be emotionally frustrated in trying to complete it, and the responses given would be effectively random. Administering the grade five test to a grade twelve class would be worthless on the other end; students would feel their time was being wasted by an assessment that easy. Typically, norm referenced results that are more than a year apart from the intended grade level are compared to data mathematically extrapolated[4] from the original set. As such, accurate results are highly dependent on the quality of the model used to fit the data. The different models will be discussed in more detail in lesson seven.

[2] The grade level assigned in July and August for the traditional ten month school schedule is ambiguous. Most bodies increment the grade on July 1, so this student would be in grade 5.9 on June 30, and then grade 6.0 from July 1 to September 30 inclusive. Some bodies leave the student grade as 5.9 through July and August, and increment on September 1.

[3] Mathematically speaking, interpolation is like playing "connect the dots" when you have some idea about what the picture should look like, but when some dots are missing. Given the expected picture, which would be the theoretical model of student performance in this case, one tries to deduce where the missing dots are and reports the values as such.

[4] Mathematically speaking, extrapolation is similar to interpolation, but is used to continue to "connect the dots" beyond the limits of the original picture. Because the reference dots are only on one side, it is harder to get accurate data in this fashion.

A final obstacle in this type of assessment is the fact that one needs a variety of possible scores to build an accurate model, which means longer assessments in many cases. With more assessment items, one is more likely to assess multiple skills at once. As such, students who excel in one category but struggle with another may not fit the model very effectively. For example, if a student who excels at computation but struggles with reading is given a typical math assessment, the student may perform well above average on the computational items, but struggle with word problems due to the difficulties with reading itself.[5] Assessments that measure multiple skills are called *multidimensional assessments*, while assessments which measure only a single skill are called *unidimensional assessments*.

To overcome these issues, many modern norm-referenced assessments are being administered with technology, and will adapt their level to the student's functioning level. Thus, the assessment isn't tied to a particular grade, and can provide accurately normed results for students at all levels of ability, and with fewer assessment items. They tend to model each question individually, using advanced techniques discussed in lesson seven.

4.3 Typical Score Reports

The scores on norm-referenced assessments are typically published in one (or more) of four ways:

1. **Percentiles**: These are distinct from percentages. This compares the student to a percentage of the population that did as well as the student, or worse than the student. A student who is average among his or her peers will score at the 50th percentile on a properly normed assessment. A percentile score of 10 doesn't mean the student only got 10% of the questions correct; it means that 90% of the population scored higher than that student. In a class of 100 students, if the lowest mark on a test is 25%, then a raw score of 25% means scoring in the 1st percentile. This gets limiting when students get the same score: if 10 out of 30 students tie for the highest mark on a test, it can be mathematically ambiguous when it comes to determining which percentile to assign. Technically, they are all in the 100th percentile by definition, but the next lowest attainable percentile would be the 67th. That's a rather large range of ambiguity, which is why the average difficulty of most norm-referenced paper based tests is higher than the average difficulty of the typical classroom test.

[5] For example, the student may misinterpret a word problem requiring division, and then correctly multiply the numbers provided instead.

2. **Grade Equivalents**: These are the scores used most often. A grade equivalent of 4.2 GE means that the testing student performed as well *on that particular testing tool* as you would expect from a typical student who had completed 2 full months of grade 4. Note that this is *NOT* a grade level. It's easier to see that from a top performer: an honours grade 8 math student could score, say, 11.4 GE on a grade 8 math test, because most students would be in the eleventh grade before they do that well on a grade 8 test. That does *not* mean that same grade 8 student could skip ahead to grade 11 math and expect to succeed. There are too many intermediate skills that he or she has never seen. It is entirely possible for a strong grade 3 or 4 student to score a grade 5 equivalent on a grade 2 test, and score a grade 2 equivalent on a grade 5 test. It also has the disadvantage that the actual ability represented by divisions in the numbers is variable. The average grade 12 student learns more in a month than the average grade 2 student, so the difference between grade equivalents 12.2 and 12.3 represents more skills and more learning than the difference between grade equivalents 2.2 and 2.3. It is not an "equal ability" scale. Grade equivalent results are particularly sensitive to the limitations of extrapolation. If a score is more than a year away from the intended grade level for the assessment tool, then the score really only means "too easy" or "too hard."

3. **Normal Curve Equivalent (NCE)**: This avoids some of the difficulty with percentile score. Instead of grouping the student population into 100 categories, it groups the actual scores into 100 categories. So, if the lowest score in a class on a test out of 50 is 17, and the highest score is 42, then there is a 25 point range in test scores. Each mark is worth $100 \div 25 = 4$ points on the NCE scale, so a score of 17 out of 50 has NCE 4, 18 out of 50 has NCE 8, and so forth. Assigning a student a score with the NCE score means comparing them to average performance relative to the test instead of performance relative to their peers. Unlike grade equivalents, this is an equal ability scale.

4. **Standard or Scale Score (SS)**: This is often reported, but the meaning is hardest to interpret. Part of the mathematical process is norming a test is setting up a completely arbitrary reference scale to measure scores against. This is the standard score, also referred to as a scale score. This is typically an equal ability scale, but it's generally meaningless outside the context of that particular assessment. Instead, it is most often used internally to represent a student's ability, and then converted into one of the three types of scores above.

Note that all of these reporting methods share a common limitation: the student reading the results has no idea which areas need improvement and how much improvement is needed. For example, if a student scores in the 50th percentile, he or she is at class average. That student doesn't know if the

class average was 50% or 80%, and so on. Some computerized norm referenced assessments are starting to provide norms in different areas, but even then, norm referenced results do not provide the information needed to determine what needs to be improved.

4.4 Upcoming Lessons

In lesson five, we will discuss criterion referenced assessments and the manner in which the results are reported. In lessons six through nine, we will discuss different means of analyzing assessment items for a variety of purposes.

Chapter 5

Criterion Referenced Assessments

5.1 Criterion Referenced Assessments Defined

Criterion referenced assessments are those which report student achievement relative to the skills detailed in the curriculum, regardless of peer performance. The most common forms are:

- Letter grades
- Percentage grades
- "Fractional" grades (e.g. "3 out of 4")
- Skill checklists

The final item on this list, the skill checklist, is probably the least common system. It is, however, the one which makes it easiest for the student to determine which skills need to be improved for the future.

5.1.1 Letter and Percentage Grades

These are the most common types of grades presented on report cards in public schools. They are most effective in communicating the proportion of the course the student in question has internalized from the course, at least in principle.

So, if both letter grades and percentages are the same basic tool, why are both in use? It's a question of precision and accuracy. The difference between getting an A and getting a C is fairly obvious and intuitive. The difference between getting 64% and getting 66% is negligible in most cases, but the difference between getting 49% and 51% can make a tremendous difference when 50% is considered the minimum passing grade. Statistically speaking, the uncertainty on a final report card grade may actually exceed 1%, and yet those grades can determine a student's ultimate fate. When statistical error pushes a 51% student into a 49% grade, there's no going back. When one moves to a letter grade system, one can be much more confident that students are getting accurate report card results. There will always be borderline students, but an A/B/C/D/F system has four borders instead of the 99 borders in the percentage system, so students are less likely to be near one.

The down side to this type of grade is that the student doesn't know exactly which skills do and do not need to be refined and improved in the future. One might guess at the relative proportion of skills that need reviewed, but only the most self-aware students will know which particular skills those are.

5.1.2 Fractional Grades

Fractional grades are most commonly found on individual assignments rather than final report cards. These are the simplest to compute, in that they give the total marks earned relative to the total marks possible. (For example, getting nine questions right on a twelve question multiple choice test usually[1] gives the mark of $\frac{9}{12}$.)

As these marks are typically accompanied by the actual assignment or test which earned the mark, students will often have some idea of what it is they need to work on to improve. However, this information does not come from the mark itself, but from the actual assignment with its individually marked questions. The fractional marks alone are no more or less informative than letter or percentage grades.

[1] In some circumstances, tests are marked "right minus wrong" or with wrong answers being treated somehow differently than blank answers. For example, some tests give four marks for a correct answer, zero for no answer, and negative one mark for an incorrect answer. The idea is to discourage guessing. Instead, one gets a more accurate measure of student confidence levels and risk taking patterns than actual student ability.

5.1.3 Skill checklists

The final major criterion referenced assessment is the skill checklist. The reason other criterion referenced reporting methods do not inform students of specific skills needing improvement is that the entire curriculum's outcome has been compressed down to a single scale. One needs to maintain multiple scales in order to communicate this information. That's where the checklists come in.

A skill checklist reporting scheme is one which lists all outcomes within the curriculum in a checklist, and indicates on an outcome by outcome basis how the student performed. The only method found to be more effective when communicating with students is providing them with detailed, free-form comments on their achievements. Checklists like these are not common in schools, although they can be found in other areas, such as swimming lessons or martial arts instruction.

5.2 Determining a Final Grade

There are, of course, advantages and disadvantages to this type of grading. The first question is determining how to calculate the grade. We have already discussed the merits and drawbacks of including homework marks in the final grades at all. Beyond that, one must determine the relative weightings of different assessments. How should quizzes compare to exams? What about labs and large reports and projects?

If the primary purpose of giving this grade is to communicate the proportion of the course which a student has internalized, then the seemingly best answers to these questions are counterintuitive and rarely implemented.

The final report card grade should not be a weighted average of different assessments. Rather, it should communicate the curricular outcomes that the student has proven capable of managing by the end of the course. Furthermore, it should be based solely on those curricular outcomes, and not on attendance, behaviour, or other social outcomes which are not explicitly listed in the curriculum.

To compute the grade, one needs to have specific knowledge of the curricular outcomes assessed by each assessment item administered throughout the course. This becomes the source data of a skills checklist. The second counterintuitive point is related to giving credit for each skill: if the student proves competency with a skill at the end of a course, that student should get credit for the skill *regardless* of prior performance. If that student missed the relevant

questions on a unit exam but has since learned the material, then *that student has still learned the material, and should receive full credit for doing so.* Once the skills checklist has been completed, an overall grade can be computed from it. However, the student benefits most from the checklist itself. If a teacher wants to best serve all possible recipients of the final grade, then the students should receive the checklist, the overall criterion referenced grade (as a letter, percentage, or whatever system the school has chosen to implement) as well as the relevant norm-referenced information (such as average, standard deviation, and so forth.) This is a tremendous amount of work, and may not be easily implemented given that most schools mandate report card formats from the administrative level. However, it is still the ideal.

5.3 Planning: Making the Ideal Feasible

In order to reduce the work required to give the ideal report card down to a manageable level, one must do an immense amount of planning ahead. It may seem counter-intuitive, but one needs to plan the course in reverse.

Begin by writing a final examination which is an accurate representation of the curricular outcomes in the entire course. Keep track of each curricular outcome and the question(s) connected to it. Log them someplace, preferably in some sort of electronic format. Then backtrack the lessons to cover the skills in the appropriate prerequisite order, filling in quizzes and unit exams along the way. This is *not* to say one should "teach to the test." If the curricular outcome is that "students can use the Pythagorean Theorem to solve for the length of one side of a right angled triangle when given any two other sides," and the final exam asks students to solve for b in $a^2 + b^2 = c^2$, then the classroom notes and homework should still have an equal share of solving for each of a, b and c. Do *not* simply focus on the one variable that you know is on the test this year, as the students may not be prepared to properly apply the skill if the assignments and applications in subsequent years (including the eventual workplace) require an alternative application of the skill. The electronic assessment item log should include information about all assessment items on all assessments which will contribute to the final report card grade.

The author's preference for this electronic log when classroom teaching was to create a spreadsheet with a tab for each student, and where each column on a student's tab represented a single assessment tool. One tab was a master tab, listing all of the assessments from the course to be used for the final report card grades, with both the correlated curricular outcomes and the marks available for each assessment item. Each student's tab then had a section which was the curricular skills checklist. This was completed automatically through con-

ditional sums on the spreadsheet.[2] Then, with each assessment administered, a student's marks were entered on a question by question basis. It took an extra minute to enter the marks this way for each student in the class, but that extra half hour in the day after marking unit exams allowed me the opportunity to hand each and every student a personalized checklist of achievement before the standardized provincial final exam, letting each of them know *exactly* what he or she should focus on while preparing for the final.[3] The spreadsheet format also allows one to adjust the weighting of each assessment item or tool, and to provide students with an immediate update on their current standings, as well as norm referenced classroom data.

5.4 Upcoming Lessons

Lesson six will conceptually describe mathematical techniques used to analyze individual assessment items in terms of their effectiveness and performance for students in the long term. It will then cover the same material in full mathematical detail for the mathematically inclined. This lesson is typically the highest level material related to assessment that one sees when working on an undergraduate degree in education. Lessons seven, eight and nine continue from there, detailing some advanced models and techniques, again covering concepts first in detail, and then following up with the full mathematical glory in sections which can be omitted without making later lessons difficult to follow.

[2] In both Microsoft Excel and OpenOffice Calc, the SUMIF function serves this purpose quite nicely.

[3] I did this on my teaching practicums, as I moved into private education before holding a regular classroom position for a full year. My cooperating/supervising teacher, who was formally in charge of the class, later informed me that the classes I worked with performed 10% higher than the provincial averages on those standardized assessments. Quality checklists shared with students before high stakes exams can turn into very effective study guides.

Chapter 6

Single Classroom Analysis

6.1 Preface

The mathematics involved in this lesson are optional, but still minimal. An adept fifth grader could likely handle the math in this lesson. (The same is not true of the math in later lessons.) All lessons involving math will be handled the same way. An initial conceptual overview covers a completely non-mathematical description of the relevant topics, which will provide all background needed to follow the conceptual discussions in later lessons. After this conceptual overview is complete, an optional section with the full mathematical glory of the topic follows.

6.2 Evaluating Assessment Item Quality - Conceptual Overview

Teachers of a small number of classrooms do not have immediate access to the advanced analysis techniques available to large scale assessments due to the low student populations. However, there are a number of highly useful tools available to teachers on these scales which can be implemented to good effect.

It is also important to note that one should always analyze multiple assessment items for a single curricular outcome. If one only has a single assessment item and the class performs poorly, one is unable to determine if the problem is with the assessment item itself or with the lesson plan for the class in which

that curricular outcome was taught.

The three key analyses which must be performed evaluate the difficulty, discrimination and bias of various assessment items.

6.2.1 Difficulty

The difficulty of an assessment item is the most intuitive concept. How difficult is the item for the students in question? This is also the analysis which benefits most from having multiple assessment items for a single curricular outcome. This also speaks to the validity of an assessment item.[1] If a series of items relating to the same curricular outcome are valid, they will have comparable difficulty values. If an analysis of the assessment items reveals inconsistent difficulty values, then one or more of the assessment items has a problem.

The difficulty of a concept is inherent to the concept itself. The overall class understanding of that concept will depend primarily on the difficulty of the concept and the effectiveness of the instruction that introduced the concept. Thus, if a class is given three or more different assessment items for this curricular outcome and produce two (or more) distinct levels of performance on that outcome as measured by those items, then the validity of the items is in question. Imagine at first that only one of the items is out of line with the difficulty of the others. If it is more difficult than the rest, it is likely that the item is poorly phrased or presented, so that the students are unable to recognize the item as an application of that particular curricular outcome. If it is less difficult than the rest, one must look carefully to see if students are solving the problem by an alternative means, reducing the validity of the item. With only two items, it can be difficult to determine which is anomalous. If there is only one item, comparison becomes impossible. Note that items being compared need not be on the same assessment. A quiz, a unit exam, and a final exam can all be compared, although interpretation can get harder. Should the students perform poorly on a quiz, it is likely the material will be reviewed and retaught, making later assessment items appear less difficult because instruction has improved. (In this case, strongly consider rejecting the original quiz from the final report card grades.) Should the performance on later items decrease, then it is likely that students did not have sufficient time to fully assimilate the information and move through all levels of Bloom's Taxonomy, in which case the information about the skill was not retained by the students. Future courses should adjust the time spent on relative topics.

The difficulty of valid, reliable items should be maximized through improved instruction. There is no reason teachers shouldn't aim for the highest possible

[1] Validity was first discussed in section two lesson three.

average from their students, provided that high average is an *accurate* representation of the understanding students have demonstrated for the curricular outcomes assessed.

6.2.2 Discrimination

The discrimination of an assessment item is indicative of the item's reliability.[2] A reliable item will discriminate between the highest and lowest performers in a class. Top performers will answer the item correctly more often than low performers.

The concept behind calculating discrimination is simple: one the entire assessment has been graded (and not just the single assessment item in question), sort the entire class by their grades. Choose two equally sized divisions of students, where one division has the highest performers, and the other has the lowest performers. Compare the two groups. If the group of high performers performed better on the individual item than the low performers, the item is functioning properly. If the average performance in the two groups is comparable, then the item is not a reliable item, and should likely be rejected from any final calculation of grades. If the low performing group outperforms the top performers on the item, there's something terribly wrong. This typically happens on multiple choice items with a mistake on the answer key. If the answer key is correct, the item must be rejected: it is *not* a reliable item. Generally speaking, the most reliable items are the most discriminating items.

6.2.3 Bias

An assessment item that exhibits bias will be both invalid and unreliable. A bias in an assessment item means students can be lumped into groups which have different levels of performance despite comparable skill levels.

For example, imagine a class in which the average grades from female and male students are comparable.[3] To identify gender bias on a particular assessment item, perform the same steps done when analyzing discrimination, with the exception of choosing the groups. Instead of grouping students by grade, group them by the criteria which defines the bias. In this case, use one group with all of the female students, and another group of all male students.[4] If there is a significant difference in the performance of these two groups, then one needs

[2] For a refresher on reliability, see section three of lesson three.
[3] This should be every class with statistically significant populations of both genders.
[4] Androgenous students can be safely ignored in this analysis.

to examine the structure of the item; it somehow relates to information from the local culture which has a gender bias within.

The next most obvious groups which may have bias are those who are learning in a second language.[5] This bias can be hard to eliminate; all one can do is create assessment items using the simplest language possible. If the purpose of the class is to teach the language, such as English class, this may not be possible at all.

Finally, one needs to pay particular attention to opinion related assessment items. One should always check for bias comparing the papers written by students who support the same opinion as the teacher to those written by those on the other side of the argument. If there is a skew towards those who agree with the teacher, then the teacher may not be grading the assessment items based solely on the information presented by the student presentations.

6.3 Mathematical Overview

The mathematics in this lesson, as mentioned earlier, could likely be handled by an elementary school student. All readers are encouraged to read on. If the math is overwhelming, but the conceptual overview is understood, the reader may abandon the rest of this lesson without fear of being lost in the conceptual discussions still to come.

6.3.1 Difficulty

The mathematical definition of difficulty is simple, but counterintuitive. The higher the difficulty value an item has, the *easier* the item; the most difficult items have the lowest difficulty score.

The difficulty of an assessment item is the average score students achieved on that item. Mathematically speaking, this can be computed most easily as

$$\text{Difficulty} = \frac{\text{Total marks earned by class}}{(\text{Student population}) \times (\text{Maximum value of item})}$$

If the item is worth a single point, as with most machine scored items, this reduces to

$$\text{Difficulty} = \frac{\text{Number of students who answered correctly}}{\text{Number of students who wrote the assessment}}$$

[5] This assumes that most of the class is learning in their first language. Roles are reversed in immersion classrooms.

So, in a class of 35 students, if 30 students answer a single-point multiple choice question correctly, the difficulty is $\frac{30}{35} \approx 0.86$. A difficulty of 0 indicates all students got the item wrong, and a difficulty of 1 indicates that all students got the maximum possible mark on the item.

6.3.2 Discrimination

The discrimination calculation begins by separating the student population by grades and performance. When choosing the groups of highest and lowest performers, try to take the top and bottom quarters of the class population. Also try to have at least ten students in each group; you may need to take the top and bottom thirds of the class to achieve this. Make sure you have the same number of students in each group. Try not to simply divide the class in half, as there will be too many borderline students in each group who will skew the results towards neutral discrimination.

With the groups selected, the discrimination becomes

$$\text{Discrimination} = (\text{Average mark of high performers}) - (\text{Average mark of low performers})$$

In the case of assessment items worth a single mark (such as most machine scored questions) and equal sized groups, this can be written

$$\text{Discrimination} = \frac{(\text{Correct high performers}) - (\text{Correct low performers})}{\text{Number of students per group}}$$

Perfectly discriminating questions have discrimination of 1, although values of 0.4 to 0.6 are typically the highest one expects in practice. (It is difficult to exceed these values without having extremely difficult questions, though one should aim for the highest discriminating non-biased questions possible.) A question with 0 discrimination doesn't discriminate: the two populations had equal performance on the assessment item. A question with negative discrimination saw better performance from the low performers, which is a serious problem with the question, and typically indicates an incorrect answer key.

6.3.3 Bias

The mathematics used to check for bias on a rudimentary level are virtually identical to those used to calculate discrimination. The only difference is the selection of the groups. One wants to group the entire class by the trait which may be biased. As this is unlikely to have equally sized groups, one should

use the first formula, with the average marks within groups, to equalize the comparison. The discrimination when checking for bias should be at or near 0.

6.4 Upcoming Lessons

The next three lessons focus on item analysis of large scale assessments. The specific content within is not typically covered at the undergraduate level, but the shift towards large scale assessments built on this infrastructure is so clear that today's teachers, parents and students need to have at least a rudimentary understanding of the underlying concepts.

Chapter 7

Parametric Item Response Theory

7.1 The Need For Better Analysis

The tools presented in the previous lesson are adequate for a single classroom, but may not be adequate for larger populations. The results are closely tied to the effectiveness of the instruction within that particular classroom, and that will vary from class to class. Furthermore, the difficulty of a question is tied specifically to the students in the analysis group. The difficulty of an assessment item given to a grade three classroom will probably not match the difficulty measured when giving the same item to a grade nine classroom, whatever the item is. We need a way to evaluate individual items on a larger scale, and with more accuracy and flexibility.

7.2 The Three Rasch Paramaters - Conceptual Overview

Georg Rasch was the first to develop and popularize a model for working with detailed analysis of assessment items which was consistently decoupled from any single individual classroom or grade level. His work actually formed a series of models, based on one, two, or three parameters.

Mathematically speaking, a parameter is a variable which is set for a partic-

ular question. It is, in essence, a number which varies from assessment item to assessment item, but not from student to student. The the first two of Rasch's three parameters will be familiar to those who have read lesson six.

7.2.1 Difficulty

Rasch's first step was to refine the definition of difficulty. He proposed a continuous ability scale, and that students at all stages of academia could be plotted somewhere on this scale. In other words, rather than having a measuring scale for each individual grade, all questions of all difficulties in all grade levels can be mapped continuously, as though schools were designed around year-round schooling. He also proposed that an individual student who had a 50% chance of answering a question of a given difficulty (say, θ) would have an ability equal to that question's difficulty. This is a reasonable assumption, and nicely sets up an infrastructure for students in a particular classroom who are not performing on the same level as their peers.

With this scale established, Rasch proposed that the difficulty of an assessment item be measured by the ability at which a student has a 50% chance of answering correctly. In other words, items and students are set along this scale independently, and item difficulty is rated by where the item falls on this scale. This now reports the intrinsic difficulty of the item, rather than the performance of a specific group of students on that particular item.

In some simple models, this is the only parameter used. All other parameters are set to some arbitrary value for every item on the assessment under scrutiny. When Rasch's third parameter is introduced, the difficulty parameter must be reinterpreted. Rather than representing the ability of a student who has an exact probability of 50% of correctly answering the question, it becomes the ability of a student who has moved half way from the introduction of the skill to complete mastery of the skill. The difference arises because, when a one or two parameter fit is used, it is assumed that students cannot respond to an assessment item correctly by guessing.

7.2.2 Discrimination

For any given curricular outcome, there is a point at which no students are expected to answer the question correctly, and a point at which virtually all students are expected to answer the question correctly. Rasch's discrimination parameter deals with the area between these points.

For example, think of a grade six level math question. Grade one students

are not expected to answer it correctly, but grade twelve students are. Grade six students are expected to answer with a mix of results, some students answering correctly, and others not. Similarly, some of the most capable grade five students may deduce the steps needed to answer correctly, while other students will still struggle with the question when working at the grade seven or eight level. Rasch's discrimination parameter is related to the width of this period in which some students answer correctly while others answer incorrectly. A higher discrimination value leads to a shorter difference in ability between students who answer correctly and students who answer incorrectly. More complex skills tend to have lower discrimination values.

7.2.3 Pseudo-chance Level

Rasch's third parameter is the one most often omitted from implemented models. The *pseudo-chance level* of an item is the probability that a student with minimal ability will answer the item correctly with no knowledge of the skill involved. In other words, it is the probability of guessing correctly. In a four question multiple choice question, for example, this parameter should be at or near 25%. In a true-false question, this parameter should be at or near 50%. In most cases of omission, it is set to 0%, as though the only students who provide the correct response to an item are those who understand the skill well enough to arrive at the correct answer through actual implementation of the skill.

7.3 The Logic of Omission

All three of Rasch's parameters are reasonable, and would be expected to apply to any assessment item. Yet, the difficulty parameter is the only one consistently applied. Why is this done, and how realistic are the assumptions behind this?

Most decisions to limit things to a single parameter are based on the same, single criteria: lack of resources. Assessing these parameters with any useful level of accuracy requires administering the items to hundreds or thousands of students. If you double the number of parameters, you need to more than double the relevant student sample population. Furthermore, you need the computer processing resources to perform the analyses, and computing time scales more quickly than the student population, too. Moving from one to three parameters could mean increasing the required budget by more than a factor of 10. Others choose to reduce the parameters because of a mathematical anomaly which may occur only with the three parameter fit, as a result of students with inconsistent results. These cases will be described in more detail in lesson eight.

So, how does one ensure that the results are accurate with fewer parameters in use? This is done by performing a little trick common to the social sciences but abhorred in the harder sciences: if you cannot change the model to fit the data, you change the data to fit the model. Assessment items which do not conform to a one or two parameter fit are discarded entirely. This practice is effective for norm-referenced assessments, but is open to dispute for criterion-referenced assessments, particularly with the more complex skills. The more complex the skill, the lower the discriminating power, the more likely students will just give up and guess, and the less likely it is to conform to a one or two parameter fit. Thus, norm-referenced assessments based on only one or two parameters tend to have extremely high statistical uncertainties at the highest levels of difficulty where the average complexity of a skill becomes large. Still, at lower levels and with less complex norm referenced skills, the one and two parameter fits work well enough.

Most pencil and paper based norm-referenced assessments are modeled using only the two parameters listed here. Very few incorporate Rasch's third parameter. Moreover, older pencil and paper assessments are often modeled as though all items on the assessment are at the same difficulty level with the same discrimination, whatever that may be. This is why the extrapolation methods mentioned in lesson four tend to become inaccurate when a student departs too greatly from the anticipated ability of students writing that individual assessment tool: when the third Rasch parameter is assumed to be zero, and when all items are treated as equivalent, the small errors that are present in the ability regions of interest become large errors when one departs that region. What the author will never understand is why Rasch's third parameter is set to 0 when not explicitly studied. It takes virtually identical amounts of computing power to set it to 0 as to 0.2, 0.25, 0.5, etc. to represent the number of options presented to the student on a machine scored assessment. If all assessment items with n choices are to have a single, arbitrary value, that value should be $\frac{1}{n}$.

7.4 The Three Rasch Paramaters - Mathematical Overview

The remainder of this lesson will incorporate some mathematics from late high school or early post-secondary, depending on the region. It can be omitted without reducing the accessibility of the conceptual aspects of future lessons.

With all combinations of parameters, items are modeled in terms of the probability P that a student of ability θ will correctly respond to the assessment item.

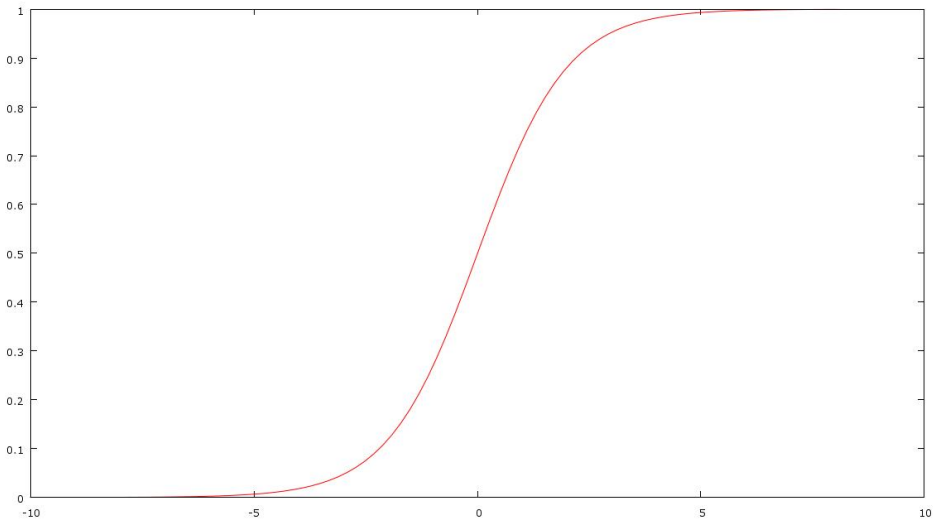

Figure 7.1: A single parameter Rasch model

7.4.1 Difficulty

If we label difficulty with variable b, then the formula describing the probability of a student of ability θ correctly responding to the assessment item is given by:

$$P(\theta) = \frac{e^{\theta - b}}{1 + e^{\theta - b}} \tag{7.1}$$

On mathematical scales, $0\% = 0$ and $100\% = 1$, such that a plot of such a graph for an assessment item of difficulty 0 would look like the graph seen in figure 7.1.

This graph is the item characteristic curve for the assessment item. In a model using only this single parameter, the only distinguishing features among the graphs would be their positions along the horizontal axis, such that three different item characteristic curves graphed side by side would look like figure 7.2.

This begs the question, how is the difficulty scale set? It seems odd to have 0 as the middle of the scale. There are methods to do this based on intrinsic data from the set and based on student ability levels, and these "natural" methods almost always have 0 in the middle of the scale. The ability scale here is the same as the scale score or standard score introduced in lesson four. It's arbitrary, and chosen by the entity creating the assessment. As the Rasch scale is an equal ability scale (meaning the difference in skill levels between scores 1.2 and 1.3

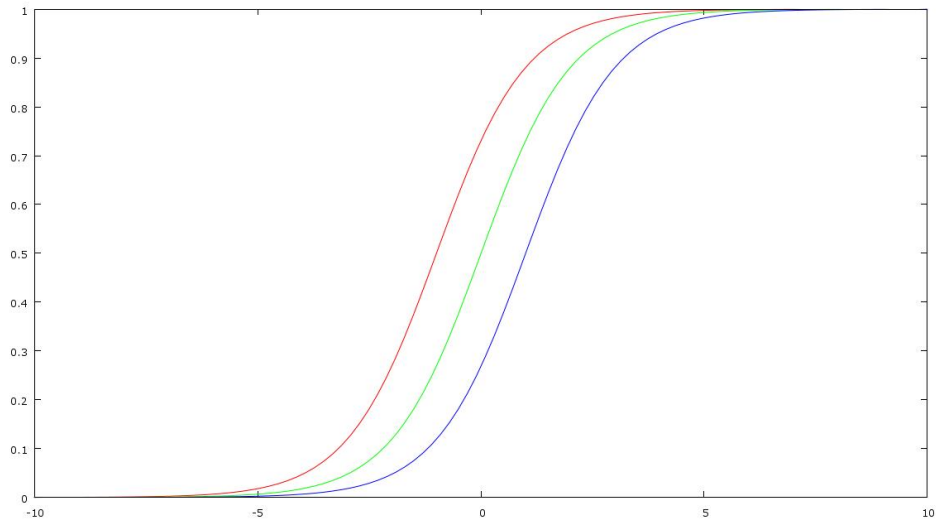

Figure 7.2: Three item characteristic curves modeled with a single parameter Rasch model. The only differences between the curves are their difficulties, which are the points at which the probability of answering correctly is exactly 0.5.

reflects the same difference in skill levels between scores 11.2 and 11.3), grade levels or grade equivalents are not appropriate scales to use, as the model fits will not conform to the actual student results. Typical practice is to fix two points on the scale at the extremes, with one point indicating an assessment item every student answered incorrectly (to define the upper limit of difficulty) and a second point indicating an assessment item every student answered correctly (to define the lower limit of difficulty.) If these points are set at 0 and 1000 (as is common practice) then many practical questions will fall in the 100-400 point range. Surprisingly, this range is not centered about 500: regardless of student ability, it is much more likely that assessments will include items every student will get right than items every student will get wrong. (i.e. the average student is more likely to approach the low end of the scale than the high end.)

7.4.2 Discrimination

The discrimination parameter a is added to equation 7.1 as follows:

$$P(\theta) = \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} \tag{7.2}$$

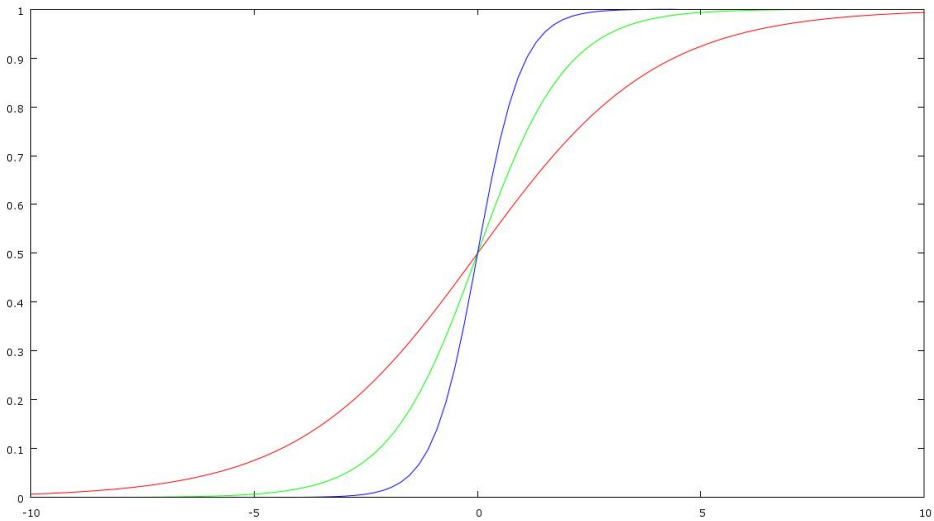

Figure 7.3: Three item characteristic curves with difficulty 0 but with three different discrimination values. (red: $a = 0.5$, green: $a = 1$, blue: $a = 2$)

For three item characteristic curves with the same difficulty (0) and three different discrimination values, see figure 7.3. Note that higher discrimination values lead to steeper item characteristic curves. These are less complex skills, which move from introduction to mastery in shorter periods of time.

7.4.3 Pseudo-chance Level

The pseudo-chance level parameter c is added to equation 7.2 as follows:

$$P(\theta) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}} \tag{7.3}$$

As mentioned in section 7.2.1, the introduction of this parameter alters the meaning of the difficulty parameter. In a one parameter fit, the probability of a student of ability b correctly answering the item is given by

$$P(b) = \frac{e^{b-b}}{1 + e^{b-b}} = \frac{e^0}{1 + e^0} = \frac{1}{1+1} = \frac{1}{2}$$

or 50%. Once the third parameter is introduced, this becomes

$$P(b) = c + (1 - c)\frac{e^{a(b-b)}}{1 + e^{a(b-b)}} = c + (1 - c)\frac{1}{2} = \frac{1 + c}{2}$$

which is the probability exactly half way between the probability c of guessing correctly with no knowledge of the skill and 100%. This is exactly half way between the lower and upper limits, which places the student half way between introduction and mastery of the skill.

Note also that many texts interpret the discrimination parameter directly in terms of the slope of the line. The definition used here, related to the time elapsed between introduction to and mastery of a skill, has two advantages:

1. It is easier to understand to a non-mathematical audience.
2. It does not require reinterpretation with the introduction of the third parameter. The higher the pseudo-chance parameter, the lower the slope of the line, but the difference in ability levels between when the line "looks" straight at the lower limit to the naked eye and when it "looks" straight at the upper limit remains unchanged.

Three different item characteristic curves with $a = 1$ and $b = 0$ are seen in figure 7.4. The red curve is a standard Rasch fit with $a = 1, b = c = 0$. The green is the same curve with $c = 0.25$, as it would be in most four question multiple choice assessment items. The blue curve shows $c = 0.5$, as it would be on a typical true-false type of question.

7.5 Upcoming Lessons

In lesson eight, we will discuss further analyses made possible by the Rasch models, and discuss how they pertain to computerized adaptive testing, while will become much, *much* more common in the future. In our final lesson, we discuss item response theory which does not involve any parameters.

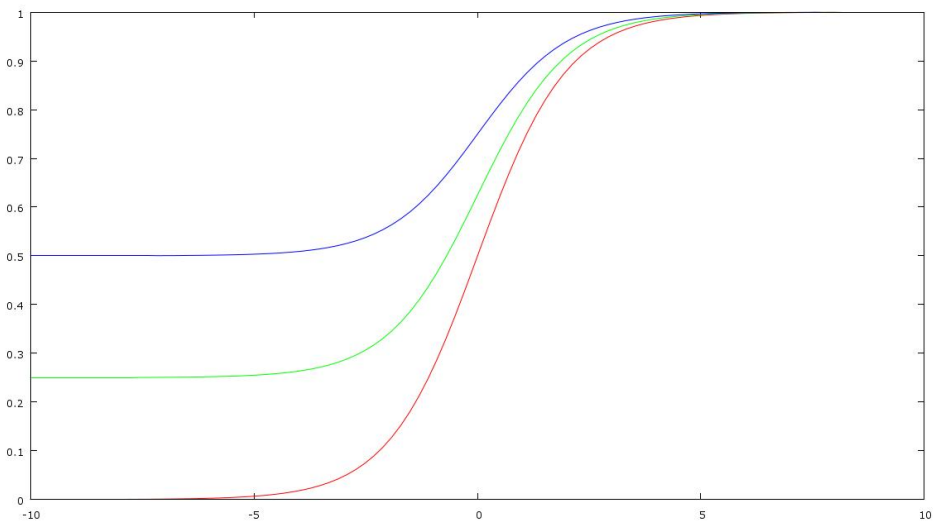

Figure 7.4: Three different item characteristic curves. The only difference in the parameters is found in the c parameters. (red: $c = 0$, green: $c = 0.25$, blue: $c = 0.5$)

Chapter 8

Computerized Adaptive Testing

8.1 Limitations of Traditional Theory

Assessments which do not use item response theory, introduced in the previous lesson, have a number of limitations, many of which have been alluded to, but which have not yet been collected into a single list. Here is that list:

- All items on a single assessment must be treated as having the same difficulty, the same discrimination, and the same bias in order to produce norm-referenced models of the assessment. This is remarkably unlikely.
- Evaluating the difficulty, discrimination and bias of test items on a pencil and paper test by traditional means is more indicative of the class population than of the properties of the test itself.
- It is impossible to evaluate the performance of a single student outside the context of his or her classmates.
- Reliability is difficult to measure precisely and in isolation for individual test items. Only its secondary effects on discrimination can be seen, and these can be obscured by skill complexity.
- Statistical uncertainties in student performance cannot be measured, or even estimated precisely.

It is very difficult and time consuming to apply item response theory to a teacher-created pencil and paper test. Doing so can improve the quality of items over time, but a considerable amount of time is required to collect enough statistics to make the results useful.

Item response theory was first applied to standardized pencil and paper tests written for specific grade levels. Rather than looking at the probability of getting a particular item correct, the student's percentage correct on the assessment is evaluated instead. A particular mark was mapped to a mathematical scale, and that was reported as the student's ability. It suffers from the same problems as before, particularly when dealing with students who are working significantly above or below the test level. Computerized adaptive testing can overcome these problems quite nicely.

8.2 Computerized Adaptive Testing - the Concept

With a computerized adaptive test (CAT), each student in a group can receive an entirely *different* but still perfectly valid (and reliable) test. The first step is to develop a very, very large number of assessment items for a database. Then, test creators give the items on the CAT to a large number of students to define all relevant Rasch parameters for the items in question. At this point, the CAT is ready to be administered. When a new student starts the test, he or she is given an assessment item from the database of items which is appropriate to the student's age grade level. Before the second question is served to the student, the first question is marked. If the student responded correctly, the computer draws a more difficult item from the database. If the student responded incorrectly, the computer draws a less difficult item from the database.

The process continues until the student has at least one response correct and one response incorrect.[1] At this point, the computer uses statistical methods to take its best guess at the actual ability level of the student. It then serves up a question from the database which provides the most information about a student's estimated ability and administers that question. With this new information, it refines the estimate of the student's ability. As more questions are served, the student's ability is known with greater and greater precision. Ultimately, the CAT ends, either because a predetermined level of precision is reached, or because a set number of questions have been administered.

[1] If a student gets every item correct or every item incorrect, then an ambiguous case occurs. The way this is handled depends on the individual test. In any event, the student either passed unconditionally or failed quite spectacularly. Either way, his or her fate is quite sealed.

This is a fantastic tool for norm referenced testing. If a grade eleven student is reading at the grade four level, the results obtained from a pencil and paper test designed for grade eleven would be almost worthless, while a CAT would move down through the database to find the items at the student's current functioning level. As all students are measured relative to the same absolute score, the norm referenced results would be accurate regardless of the student's ability.

A CAT is not as easily applied to criterion referenced testing. Questions are rarely cross-correlated to curricular outcomes, so the criterion referencing is more difficult to manage. Though certainly possible, it is more likely that the CAT would be used as a "locator" test to determine a student's ability, and any criterion referenced assessment would be used as a secondary follow-up assessment at that level. This combination of assessments could then be used to produce a very effective curriculum customized to the needs of that particular student, which is a tremendous asset to facilities which can provide individualized instruction.

8.3 Computerized Adaptive Testing - The Math

There are three mathematical components to the above conceptual description.

1. Determining the amount of information provided by a particular assessment item. This step requires calculus.
2. Estimating the student's ability. This can be done using statistics, but it most commonly accomplished through calculus.
3. Determining the precision to which a student's ability is known. Although this is derived from statistics, the application requires algebra alone.

The equations generated are all relatively straightforward combinations of statistics and calculus. Diverting into their specific derivations would require spending an excessive amount of time and paper on material not directly related to assessment, so they will be presented without proof.

8.3.1 Information Functions

Assessment items only provide useful information if their rated difficulty is close to the student's ability. If an item is too easy or too difficult, learning that the student gets the item wrong or right tells us next to nothing. Therefore, it is

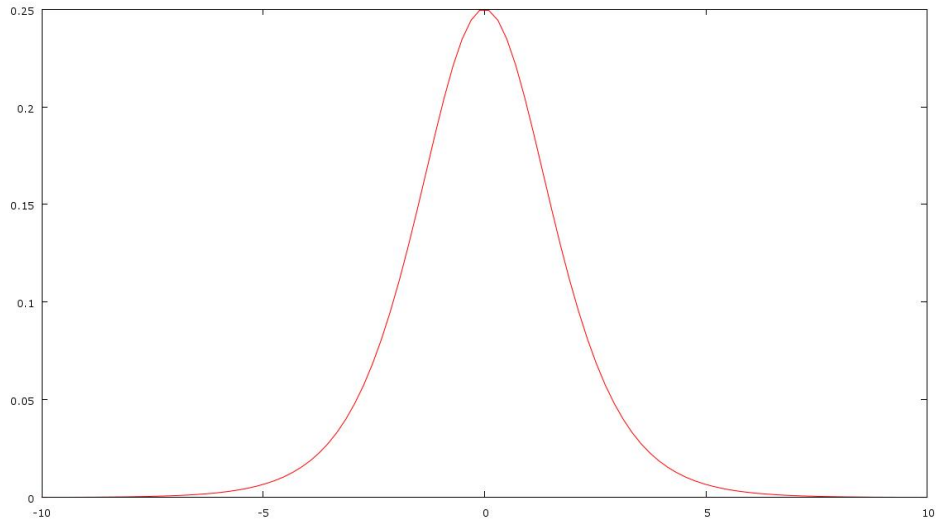

Figure 8.1: The information function corresponding to an item characteristic curve with $a = 1$ and $b = c = 0$.

not surprising that a mathematical formalization of the information function would depend on the same θ variable that the item characteristic curve $P(\theta)$ depends on. Regardless of the number of parameters involved in the Rasch fit, the information function $I(\theta)$ related to the item characteristic curve $P(\theta)$ is given by

$$I(\theta) = \frac{\left(\frac{dP(\theta)}{d\theta}\right)^2}{P(\theta)(1 - P(\theta))} \tag{8.1}$$

where $\frac{dP(\theta)}{d\theta}$ is the usual derivative from calculus. For a full three parameter fit,

$$\frac{dP(\theta)}{d\theta} = a(1 - c)\frac{e^{a(\theta - b)}}{\left(1 + e^{a(\theta - b)}\right)^2}$$

If the fit has less than three parameters, then simply substitute the assumed values of a and c into the expression. For the special case with $a = 1$, $b = c = 0$, the information function can be seen in figure 8.1.

Although the information function and item characteristic curve have entirely different vertical scales and interpretations, they are both determined by the same set of Rasch parameters and θ. As such, it is useful to plot them on the same axes, despite the incomparable vertical scales. This has been done in figure 8.2. Notice that the information function peaks at $\theta = b$ and drops rapidly

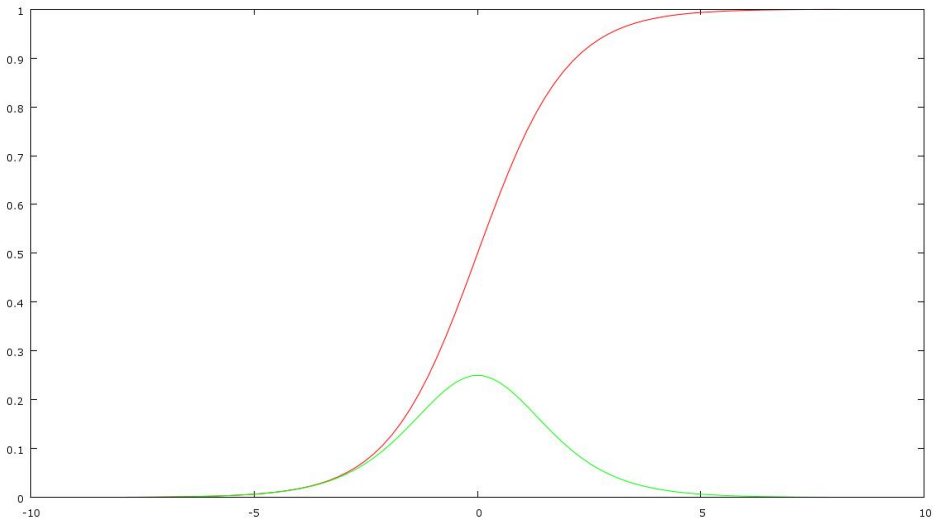

Figure 8.2: An item characteristic curve and its information function plotted on the same axes to compare their dependence on θ, despite the incompatibility of their vertical scales.

to each side. This is consistent with intuition: information is only garnered about a student's ability from an assessment item of similar difficulty.

Notice also that the information function depends on the square of the derivative of the item characteristic curve. Less discriminating items have shallower slopes, as do problems with high pseudo-chance levels. These problems provide less information about a student's ability, and should be avoided if the sole purpose of the assessment is to generate norm referenced results.

Finally, it is important to note that information functions add linearly when multiple assessment items are administered. This will be important when we discuss the calculation of uncertainties and standard errors two subsections from now.

8.3.2 Determining Student Ability

Determining a student's ability is the most mathematically involved portion of item response theory and computerized adaptive testing. Imagine the student has completed n assessment items thus far, and the response provided for assessment item j is represented by u_j, where $u_j = 1$ for a correct response and $u_j = 0$ for an incorrect response. The likelihood L that the student who has

ability θ and has provided responses u_1 through u_n for the n assessment items answered thus far is given by

$$L\left(u_1, \ldots, u_n, \theta\right) = \prod_{j=1}^{n} P_j^{u_j} \left(1 - P_j\right)^{1-u_j} \tag{8.2}$$

where the θ dependence of $P\left(\theta\right)$ has been omitted to make subscripts and superscripts more clear, and where $\prod$ is the standard product notation.[2] The most likely ability level for a student to have is the maximum value of this function. The maximum value occurs where $\frac{dL}{d\theta} = 0$.

This is not a fun integral to evaluate, particularly given the rapidly changing nature of L with each completed assessment item. It can be simplified somewhat by noting that the maximum of L will correspond to the same θ as the maximum of $\log L$, which can turn the function being differentiated into a sum rather than product, which is computationally preferable. Still, numeric methods are used to find the interesting point. Things get even more complicated when a student's responses are inconsistent in a three parameter fit. Although impossible in a one or two parameter fit, a three parameter fit combined with student responses which include correct responses to relatively difficult questions and incorrect responses to relatively easy questions can result in a likelihood function which has a maximum at no finite value of θ. Some assessments use one or two parameter fits specifically to make this problematic situation mathematically impossible. Unfortunately, inconsistent results are something of a hallmark of students with learning disabilities,[3] so omitting the third parameter to ensure these situations are never represented means ensuring a group of students who most need a test which will adapt to student performance will not be accurately assessed by them.

This is the step which ensures application of item response theory to assessment will likely always be performed through the use of computer technology instead of by human hand. Determining which question a student needs to do next just takes too much time during a testing situation.

[2]If the reader is unfamiliar with $\prod$, it is to repeated, sequential multiplication what $\sum$ is to repeated, sequential addition. For example, $\prod_{j=1}^{5} j = 1 \times 2 \times 3 \times 4 \times 5 = 5! = 120$. One starts by substituting the $j = 1$ starting value that appears below the $\prod$ into the expression (j) to the right of the $\prod$ and evaluating that. Then you add 1 to the value to obtain $j = 2$, and substitute that into the expression to the right of $\prod$. This continues until you reach the value $j = 5$ listed above the $\prod$. At this point, you take the values of everything you calculated after each substitution (1, 2, 3, 4 and 5) and multiply them together.

[3]The author personally hates the term "learning disability." In his experience, the term "learning anomaly" is far more accurate.

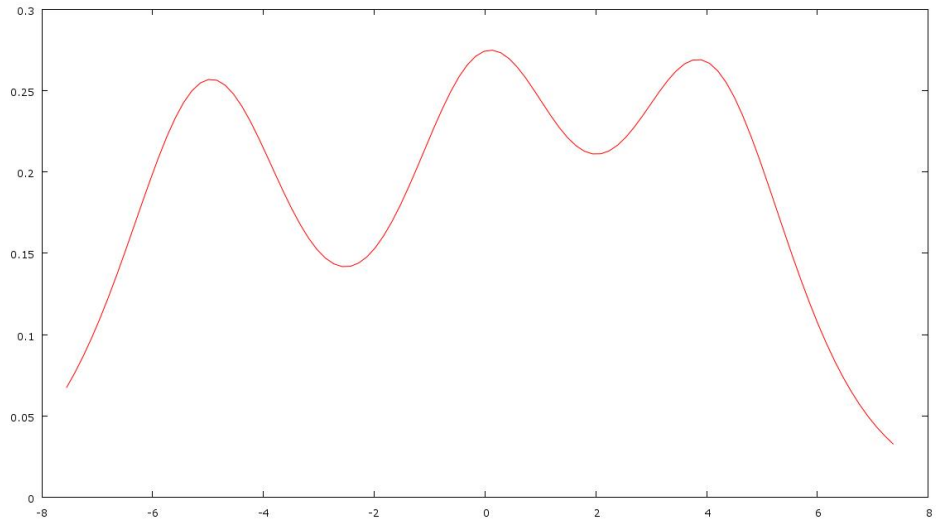

Figure 8.3: The combined information function after administering three assessment items of difficulties 0, 4 and -5.

8.3.3 Computing Uncertainties

Now that we have a method by which to estimate a student's ability, we need to determine the accuracy of that estimate. The standard error $SE(\theta)$, or statistical uncertainty, in the measure of a student's ability as θ after administering n assessment items is given by

$$SE(\theta) = \frac{1}{\sqrt{\sum_{j=1}^{n} I_j(\theta)}} \tag{8.3}$$

where we have used the previously mentioned fact that information functions add linearly.

For example, assume we have administered three assessment items. For simplicity, assume $a = 1$ and $c = 0$ for all three. If the three items have difficulties of 0, 4 and -5 respectively, then their combined information function can be seen in figure 8.3. Notice how the items add nicely when the relative difficulty levels are close together. The standard error function for this information function can be seen in figure 8.4.

For a better comparison of the θ dependence of the two functions, they can be plotted against the same horizontal axis with differing vertical axes as seen in figure 8.5.

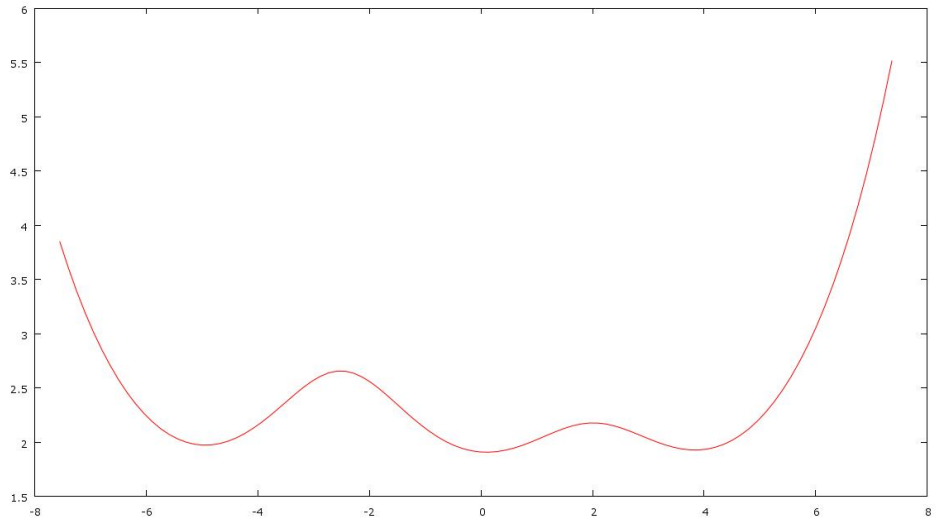

Figure 8.4: The standard error function corresponding to the information function in figure 8.3.

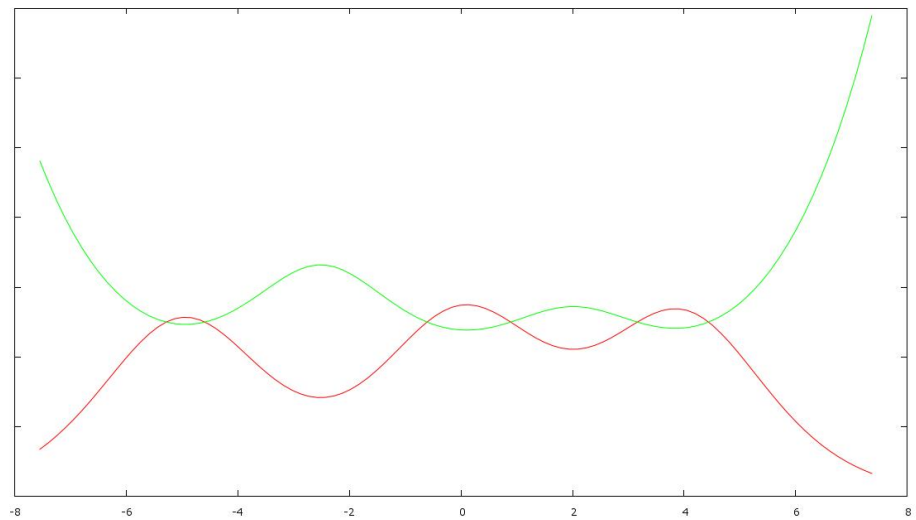

Figure 8.5: Information and standard error functions plotted against the same θ access with different vertical axes.

The fit is complete when the standard error at the anticipated value of θ has dropped below a specific threshold value. In many cases, particularly those using only single parameter Rasch fits, the assessment continues until a specific number of questions have been administered. Because all such information functions are identical apart from position on the θ axis, if the test administrators can accurately predict the rate at which the test will adapt to any given student's ability, then they can determine in advance how many questions are required to reach the required minimum standard error. This reduces the computational requirements during the test administration, as standard error and information functions would no longer need to be tracked. If one uses a more accurate two or three parameter fit, then the calculations must be done along the way.

8.4 The Upcoming Lesson

In our final lesson, we will describe methods of measuring student abilities with questions that do not fit into any parameterized model of any kind.

Chapter 9

Nonparametric Item Response Theory

9.1 Limitations of Parametric Item Response Theory

When assessment items are modeled with a Rasch model, certain assumptions are present which do not always apply. One such assumption is that the likelihood of responding correctly to a particular item increases in a smoothly defined fashion. This is not always the case. In some cases, as students are exposed to related skills, they rapidly assimilate previously presented information, causing a spike in difficulty. A second assumption is that, at some point, every student will respond to a question correctly. This is not necessarily the case. If the question is to multiply a 30 digit number by a 30 digit number without the aid of a calculator, odds are excellent that any given individual will make a computational error at some point in the process (which involves a minimum of 1741 distinct operations, not counting regrouping) and get the item incorrect. Finally, it assumes that the odds of correctly answering a question are strictly increasing with student ability. While this is usually true in the classroom situation, there are exceptions[1] to this rule. When item response theory is applied to political opinion polls and other non-classroom situations, these situations become far more common.

[1]The author's favorite exception: "given that $x + y = 2$ and $xy = 3$, what is $\frac{1}{x} + \frac{1}{y}$? High school students will attempt to solve this through substitution and typically get stuck when they reach an equation which can only be solved by taking the square root of a negative number. Junior high students haven't seen that method yet, and solve the problem by looking for a common denominator and finding that $\frac{1}{x} + \frac{1}{y} = \frac{x+y}{xy} = \frac{2}{3}$.

9.2 Benefits and Limitations of Nonparametric Item Response Theory

Nonparametric item response theory concerns itself only with relative relative performance of the individuals being assessed. No absolute ability or difficulty scales are used at all. Instead, examinees are sorted and ranked by ability and ability alone. When it is done, you may know that Alice has more ability than Bob, that Carol comes in third and that Ted is at the bottom of the pile, but you will not know the degree to which the differences arise. Because of these limitations, nonparametric item response theory has very limited applications to academia. It may be used to screen applicants for a private institution of some sort, but is nearly useless when measuring student performance in any other respect. It is, however, frequently applied outside academia.

Most opinion polls that are analyzed by professional statisticians are based on this theory. The analysis is performed almost exclusively through variance and covariance calculations from statistics, and the methodologies lend themselves nicely to questions that fall on a rating scale, such as "strongly agree, agree, neutral, disagree or strongly disagree" with each possible response being assigned a numeric value. In this type of analysis, conversion of responses may be necessary. For example, if left wing politics lead to a level of agreement with item 1 but disagreement with item 2, then the values of one of these items should be reversed so that the numeric values assigned are comparable. Similarly, items that are regarded with the same importance are to be treated equally, they must have the same extremes for numeric values. For example, if the five options above are valued such that "strongly agree" is assigned the value 2 and "strongly disagree" is assigned -2, then another item on the same survey with only "yes" or "no" options available must be assigned with values of 2 and -2 (in some appropriate order) to be equitably compared to the earlier item.

There is one other benefit to nonparametric item response theory which can be useful in the analysis of long assessments. The larger the assessment, the more likely that it is multidimensional, meaning it will assess multiple skills at once. Students who are more adept at one group of skills than another can have seemingly inconsistent results with parameterized item response theory. The prevalence of covariance measures in the nonparametric version makes analysis of correlations simple. The theory can identify which assessment items measure the same skill or group of related skills, and which belong to a different group. Armed with this information, traditional parameterized item response theory can be applied along two distinct lines, producing two distinct scores in the different skill groups. As such, it can be an effective companion to a parameterized analysis of an assessment tool, but still does not serve as an effective academic tool in isolation.

9.3 Final Recommendations

Given the variety of assessment tools and purposes available, recommendations will depend to a degree on the intended purpose of the assessment. The author's recommendations are as follows:

Goal	**Recommendation**
To give a general picture of student achievement within a course or curriculum.	A letter grade serves this purpose best.
To inform students of their strengths and weaknesses. (This should be a goal of every academic course.)	A skills checklist derived from summative assessments, by a mile. This gives students exactly what they need to improve in the long term. A letter grade, percentage, or norm-referenced mark of some form may be a useful supplement, but does not provide enough information on its own to reach the goal.
To screen candidates for acceptance to a job or academic program.	A norm referenced assessment. If space is limited, use a nonparametric item response theory analysis. If space is plentiful, use a parameterized analysis.
To evaluate the effectiveness of instruction.	Norm referenced analysis of criterion referenced assessments. Compare the average student performance to the actual curricular outcomes, with particular attention to any frequent results on the students' skill checklists.

9.4 Conclusion

The world of assessment is changing right now. The traditional methods of running with a teacher's gut instincts are being overturned by methodologies that are based on research and mathematical methods to best determine what abilities students have and how to improve them. The rate at which this change will occur will depend greatly on the funding provided to education in the future, and that will depend on the voiced opinions of voting taxpayers. Speak your mind to the local politicians about education (and every other issue that concerns you), and vote for the one who listens best.

Bibliography

Three primary reference materials were consulted while writing these lessons. They are:

- *Assessment of Student Achievement* by Norman E. Gronlund, Sixth Edition, ISBN 0-205-26858-7. This is a standard undergraduate textbook on the topic, and more recent editions are available.
- *Fundamentals of Item Response Theory* by Ronald K. Hambleton, H. Swaminathan, and H. Jane Rogers, ISBN 0-8039-3647-8. This was the main source for lessons 7 and 8.
- *Introduction to Nonparametric Item Response Theory* by Klaas Sijtsma and Ivo W. Molenaar, ISBN 0-7619-0813-7. This was the primary reference for lesson 9. Honestly, I find it lacks details of the more mathematical aspects of the advanced topics, instructing readers to buy software written by the authors which will do the job for them rather than teaching readers how to do the job themselves. Just because I referred to it, that doesn't mean I recommend it.

Index