A Brief Introduction to Language Testing Basics

First published in 2000
Revised in 2001, 2002

Randy Thrasher

JLTA

日本言語テスト学会事務局
〒389-0813　長野県埴科郡戸倉町芝原　758
TEL 026-275-1964 FAX 026-275-1970
e-mail: youichi@avis.ne.jp
URL: http://www.avis.ne.jp/~youichi/JLTA.html

The Japan Language Testing Association: JLTA Office
758 Shibahara Togura-machi Hanishina-gun
Nagano-ken, JAPAN (Zip code: 389-0813)
TEL +81-26-275-1964 FAX +81-26-275-1970
e-mail: youichi@avis.ne.jp
URL: http://www.avis.ne.jp/~youichi/JLTA.html

A Brief Introduction to Language Testing Basics

Introduction

The purpose of this workshop is to give language teachers an understanding of language testing. However, it is not a cookbook approach to the subject. It is not a collection of recipes for making the various sorts of tests that teachers may need. We will try our hand at designing and writing tests of various language skills — tests that can be used for the many purposes that teachers have. But we will move on to this hands-on part only after we have introduced test theory and the basic statistical underpinning of good measurement practice. Without such an understanding, teachers will be taking what John. B. Carroll has called “the sorcerer’s apprentice approach”[1] to language testing. You remember, I’m sure, the story of the sorcerer’s apprentice. He learned a few words that allowed his master to do wonderful things and he tried to use the same words himself but without his master’s deep understanding of what he was doing. And, not surprisingly, disaster struck.

It seems to me that most language teachers in Japan are in the same boat as the sorcerer’s apprentice when it comes to testing their students. They know a little bit but don’t have a deep understanding of what they are doing. And like the sorcerer’s apprentice they often end up causing problems for themselves and, more importantly and tragically, pain for their students. This workshop will hopefully allow you to begin to move beyond the apprentice stage and grasp a sufficient understanding of language testing to keep from making a fool of yourself and causing pain to your students. We begin by giving an overview and introducing the basic principles of language testing.

Tests and the role of measurement in foreign language education

Decisions must be made in education. They cannot be escaped. In a world where resources are limited and demand isn’t, somebody must decide who gets the scarce resources. Or the decision may concern the most efficient way to allot time or present materials. A moment’s reflection will convince you that decisions are necessary. Even deciding not to decide is a (probably unwise) decision. Given this situation, the question that a teacher must face is how to make the best possible decisions. And that is what language testing is all about.

We begin with a discussion of the basic components of decision making and define a

[1] Personal communication, summer 1980. Carroll made the comment over lunch to describe the methods used by a well-known language test theorist of the time whose lack of understanding of factor analysis had led him to make unsupportable claims about the nature of language.

test as a device for gathering information to help us answer some question. John B Carroll put it very elegantly when he said, “the purpose of testing is always to render information to aid in making intelligent decisions about possible courses of action.” (1991:31 quoted in Spolsky 1995:223) Thus, we view testing as the handmaiden of good decision making — the tool for providing data to allow us to make the most intelligent choice possible.

The nature of language measurement

Everyone is familiar with measuring things. We have no difficulty using a tape measure to find out how long or wide a table is. We’re quite capable of measuring out the ingredients needed to make a cake and the gauges on the dash board of our cars hold no mystery for us. We may not know just how the level of oil is determined but the general principle is clear enough. Yet, it is our very familiarity with measuring concrete things (tables, flour, oil, etc.) that gets us into trouble when we begin to measure abstract entities such as language ability. We find it difficult to realize that measuring something that we cannot see or touch is completely different from measuring someone’s height or weight. Part of our problem is that we cannot measure such abstract things directly. We cannot get inside people's heads to find out the degree to which they understood something said to them in a foreign language. We can only have them do some observable task — like answering questions about what they heard — and estimate their comprehension from their performance on that observable task. Yet there have been comparable problems in the world of physical measurement. For many years, the circumference of the earth could not be measured directly and scholars found many ingenious ways to estimate it using what could be measured directly. But the main difference in measuring concrete and abstract entities does not lie in the direct and indirect measurement split. It lies in the fact that the entities we must deal with when we measure language are totally different from objects in the physical world. The width of a table is a concept that is easily grasped and everyone understands the concept in the same way. Listening comprehension, on the other hand, is not an easily grasped concept and different people have different ideas about what is the exact nature of the skill (or skills) that constitutes it. In language testing we recognize this complex, poorly understood nature of listening comprehension by calling it (and each of the other entities we try to measure) a **construct**. By this we mean that listening comprehension is not a ‘thing’ out there in the world like a table that has width, length and weight, but that it is something formed in people’s minds and we must be careful to spell out as carefully as we can just what we mean when we use the term.

And the fact that we must measure constructs and not physical objects means that the life of a language tester is much more difficult than that of a carpenter, cabinetmaker, or cook. The world of the language teacher is full of abstract entities that must be measured and so we must form constructs in order to try to answer questions about such entities.

Let me try to illustrate what I'm talking about by looking at a construct we are all familiar with — vocabulary. We can count words. Words are enough like cans of soup or other prototypical things that we can use numbers to measure their quantity and not confuse people. If I tell you that this sentence contains 19 words, you will have no problem understanding what I mean. But if I tell you that Johnny knows 80% of the words that were introduced in lesson 12, and you know that that lesson contains 20 new words, you might be tempted to calculate that 80% of 20 is 16 and assume that the number 16 has the same status as the earlier number 19. In one sense it does, but there is another sense in which the two are totally different. In the first case we were counting objects — clusters of marks on the page. In the second case we introduced the word 'learn'. And this means we have set ourselves the task of measuring what we cannot directly count. We can only decide if 'learning' has taken place by examining some other thing that we can count. In this case we may decide that a child has learned the word if he or she can correctly answer the multiple choice questions at the end of the lesson that are designed to test the new words in that lesson. Notice that we have defined 'knowing the new words in the lesson' as success on a particular kind of test. Once we realize what we have done, we are then able to ask if our definition of 'knowing new words' is a good or reasonable one. Is being able to answer test questions really what we mean when we say that someone 'knows' a word? If someone asks you if you know a word, how would you justify a 'yes' answer? All of us carry around in our heads the meaning of expressions such as 'learn a new word' or 'know a particular word'. And it is these 'ideas in our heads' that are the stuff of language testing. It is the job of language testers try to refine these everyday definitions that we use without thinking about them. The first step in this refinement is to make these definitions explicit. Once this is done; once we have a statement of what we think 'learning a word' means, then we are ready to both test our definition and test whether Johnny has 'learned new words' or not. Obviously, our answer to the second question — How many of the words in Lesson 12 did Johnny learn? — will depend completely on the quality of our definition — our construct.

Almost all of the constructs used in language testing carry labels that we are familiar with; ‘grammar’, ‘vocabulary, ‘pronunciation’, ‘reading’, ‘writing’, ‘speaking’, etc. However, this very familiarity is a source of much of the misunderstanding we have in language testing. If a test writer told us that his test measures xyz or 5-7-D we would have to ask or figure out what xyz or 5-7-D means. If we find a test labeled ‘grammar’ we think we know what it is testing, but it is possible, even likely, that what we think ‘grammar’ is and what the test defines as ‘grammar’ are two different things. In other words, the construct grammar that we carry around in our head may not be the same as the one the test writer is using. This problem can perhaps be best illustrated by looking at a well-known construct outside of language testing; the construct of intelligence. We all can understand the statement, ‘Jo Ann is the most intelligent student in the class.’ We are quite used to equating that statement with another one; ‘Jo Ann has the highest IQ in the class.’ We know that IQ stands for ‘intelligence quotient’ so we assume that the two sentences mean the same thing. But, IQ is a construct that comes from a group of tests that trace their origin to Alfred Benet. Benet, and particularly his successors, defined intelligence as the ability to perform certain spatial and numerical tasks. A number derived from the child’s performance on these tasks, divided by the child’s age is said to be his or her IQ. The fact that this construct of intelligence does not match very well with what most people mean when they use the term became quite clear when a Harvard psychologist, Arthur Jensen, argued from IQ data that American Blacks are mentally inferior to Whites. Critics of Jensen were able to point out a fatal flaw in his argument. They were able to show that the IQ tests used contained tasks that were familiar to white children, particularly middle-class white children, but not familiar to the average Black child. In other words, the test writers’ construct of ‘intelligence’ was racially biased. It was not based on all human beings but only on a small subset of them. The lower scores of Black children were not a matter of intelligence as we usually think of the term. In fact, in recent years, there have been a number of challenges to the traditional IQ construct. One scholar (Gardner) cogently argues that there are five different types of intelligence.[2]

I recommend reading about the controversy over IQ and the supposed mental inferiority of Blacks. Stephen J. Gould’s book, The Mismeasure of Man, is an excellent, and sobering, discussion of this topic and its historical background[3]. We need to realize that test

[2] Another serious problem with IQ was pointed out by Steven P. Gould. He pointed out that, whatever intelligence is, it is clearly too complex an entity to be expressed as a single number.

[3] Gould’s book also discusses a number of other points that are useful to anyone with an interest in scientific measurement.

constructs and the constructs the general public carries around in its head are very often not the same, and that this gap can have serious consequences.

Once we realize the necessity of understanding just what the test writer's construct is, we have to ask how to determine it. Few tests come with the constructs spelled out. Even the few with manuals don't usually state what constructs were used. So we are left to figure out what the test writer meant by what he or she claims to be testing by looking at the content of the test itself. We have to ask what construct of, say 'grammar', lies behind the test items labeled as such. But it is also a very useful exercise to sit down and try to figure out just what we mean when we talk about vocabulary, grammar, and the like. In this workshop you will be asked to do just that with listening comprehension, reading, writing and speaking. Before we can talk meaningfully about how to test such things we have to ask what our construct of each of them is.

Validity and reliability

Validity

The central concept of language testing is validity. Validity is the degree to which a test provides results that are useful in answering the question we started with. We will discuss a newer understanding of the concept later in this section but let me introduce some traditional ideas about validity first. We have traditionally considered four sorts of validity but in the last 30 years or so two more validities have been added.

One traditional type is **face validity**. This sort of validity is determined by asking if the test appears to be a good measure of what we want to test. Some language testing theorists believe that this sort of validity is useless and even counterproductive. However, most of us realize that face validity has its place but recognize that over-reliance on it has slowed the improvement in test design. Face validity has a role to play because, if test takers do not believe that the test is really measuring what it is supposed to measure, they may not take it seriously and not work up to their full ability when they take it. But, there are two major problems with face validity. The first is that mere appearance is a poor guide in deciding if a test is or is not valid. The second is that, by demanding that tests of a particular skill look like tests we have seen in the past that carry that same label, we have encouraged the continued use of tests (or test formats) that may not really be useful in measuring the skill that we want to measure.

Another type of validity is **content validity**. It is most easily understood if we consider achievement testing. In this setting, this sort of validity asks if the content of the test matches the content of the course of study that it is supposed to be measuring mastery of. It is usually estimated by asking content specialists to examine the test to see if what is in the test matches what is in the content area being measured.

Concurrent validity asks if a new test is measuring the same thing as an older well-established one. If we assume that the older test is a valid measure of some skill, we can check to see if the results of the new test match the results of the older one in order to determine the validity of the new test. Concurrent validity is estimated by giving the two tests to the same group of test takers and comparing the results.

Since a test is used to make decisions, we can determine if those decisions were correct or not by asking if things turned out the way the decisions based on the test results predicted they would. The TOEFL is used to decide if non-English speaking students have sufficient ability in that language to succeed in an English medium college or university. We could compute **predictive validity** by comparing the test results with the academic performance of those who were allowed to enter such universities.[4]

We have already discussed the concept of a construct and stressed its importance in language testing. Therefore, since **construct validity**, the degree to which the test results match our definition or construct of the skill being measured, will be at the center of our discussion of the newer understanding of validity below, I will not say more about it here.

The most recent addition to the list of validities is what I first called **educational validity** but is now more commonly called **washback validity**. This sort of validity checks to see if the effects of the test on the teaching and learning are the positive ones hoped for by the test developers or unintended negative (detrimental) ones.

Reliability

The consistency of test results is called reliability. The concept is probably easiest to understand if we look at the rating of essays or speech samples. We want to know if the different raters are evaluating what the students produce in the same way and awarding the same marks. So we compute **rater reliability** by comparing the evaluations given by different raters to the same essay or speech sample. But there are a number of different ways of estimating the consistency of test results.

[4] Notice that, since we cannot include the people whose TOEFL scores were too low to allow them to be accepted in English medium universities in our validity study, we will not be able to detect what statisticians call Type One errors. That is, our study will not be able to determine if there is anyone that the test incorrectly assigned to the group lacking in proficiency. We will be able to detect Type Two errors — students who the test claims have enough English to succeed but, in fact, do not.

One way is to give the test twice to the same set of test takers. The two administrations must close enough in time so that no learning of the skill being tested will have occurred between the two and far enough apart that the test takers' performance on the second administration is not effected by their memory of the first. This **test-retest reliability** is computed by running a correlation between the two sets of results. If one half of the items in a test (usually the odd number items) are compared to the other half (the even numbered ones) we can compute the sort of reliability that is called **internal consistency**.

Another look at validity

We earlier talked about different kinds of validity but one of the most influential psychometricticans of the second half of the twentieth century, Samuel Messick, has argued quite convincingly that validity is a unitary concept. There are not five (or more) different sorts of validity, but there may be many sorts of evidence that can be presented to establish validity. In other words, the different sorts of validity we discussed earlier are only some of the possible ways of gathering data that could be used to establish validity.

Messick's other major contribution was to stress that validity is not a quality of a test. It is not the test that is or is not valid. What are or are not valid are the inferences that we draw using the test results. Our placement test itself is not valid. But we can determine if the division of students that we make on the basis of the results of that test is valid or not. Using the results of our placement test we decided that one group of students should go into a certain level of instruction and another group should be placed in a different level. If we got it right, if we placed the students properly, then the inferences we drew were valid. Perhaps this will become clearer if we imagine a situation in which our placement test ranked the test takers according to their ability in the skills needed in our course of instruction with perfect accuracy. The test is functioning well, but we cannot say it is valid because ranking the students is only one part of our task. Placement involves grouping as well as ranking. Our task is to turn this rank list into the groups we need to create the classes we need.

Many years ago, the English Language Institute of the University of Michigan used a placement test with two sub-tests. One was a test of grammar and vocabulary and the other was a test of listening comprehension. Each sub-test seemed to be doing a good job of ranking the students according to their ability in these two areas. In the pre-Messick view of validity, it could be argued that these tests were valid.

However, initially, placement was determined by adding the results of these two sub-tests together and dividing the students into levels according to their position in this overall ranking. But notice that, although this method of dividing up the students placed those who performed well (or poorly) in both of the sub-tests together, it also placed some very different students together. Those who were strong in grammar and vocabulary but weak in listening comprehension would end up in the same class with those who were strong in listening but weak in grammar and vocabulary. This is not a situation that will make a class that will be easy to teach. We gave the students a placement test to try to create classes containing student of the same level of ability or having the same strengths and weaknesses. The two sub-tests were doing their job but the way we used the results (the inferences we drew from the test results) did not produce the sort of classes we wanted. That is to say, the placement process we were using was not valid.

We solved the problem by using the test results in a different way. Instead of combining the results of the two sub-tests, we plotted them on a graph. We plotted the grammar and vocabulary scores on one axis and the listening scores on the other. This clustered those who were good in both skills in the upper right-hand corner of the graph. Those weak in both clustered in the lower left-hand corner. The other two corners were occupied by those students who were weak in one skill but strong in the other. Since this four way clustering was exactly what we wanted, we can say that our placement system is producing valid results. And we can see that Messick was right in claiming that tests by themselves are not valid. We can only decide if the inferences drawn from the results of such tests are valid or not.

Messick also claims that all validity is construct validity. Let's see what he means by that. In a 1996 article in Language Testing, Messick lists six aspects of validity or "standards for all educational and psychological measurement" (248). I would like to summarize these 'standards' and make the notion of operationalization, which is only implicit in Messick's list, explicit. As Messick does, I will frame these issues as questions.

1. Does our construct, and our implementation of it, include all and only the necessary elements?

a. Is there anything we need to add?

b. Is there anything we have left out?

2. Do we have these elements correctly weighted?

a. Are there elements that play a larger role in determining test performance than they do in determining the ability to perform that skill in the real world?

b. Are there elements that play a smaller role in determining test performance than they do in determining the ability to perform that skill in the real world?

3. Do these elements interact in the same way in the test task and in real-world performance?

4. Does our scoring scheme support what we are trying to do?

   a. Does our scoring scheme evaluate the test performance in the same way real world performance of that skill is evaluated?

   b. Does it allow us to draw correct inferences from the performance?

5. Is there anything about our test that will cause the test takers, or a portion of them, to perform in a less than optimal fashion?

6. Are the results generalizable?

   a. Are the results comparable across time

   b. Are the results comparable across settings

The Whole Skill and Nothing But the Skill

Our construct must be complete but, at the same time, it should not include anything that shouldn't be there. The same is true of the tasks we put into out test. The tasks we pose should force the test taker to demonstrate their proficiency in all of the sub-skills that make up the ability we are trying to measure and no others. Yet, a look at what are called test method effects will show the impossibility of this with any one type of test.

We learned earlier that we usually cannot measure language constructs directly. We have to get the test takers to do some task that we can measure and from their performance on this task estimate their proficiency in the construct we are interested in. These test tasks or methods are necessary but they also get in the way of measuring what we really want to measure. As Genesee and Upshur put it, "Test methods can have an effect on test taker's scores because they call for certain kinds of skill or knowledge that is independent of the content itself."(143) These test method effects can arise because the test takers do not perform up to their full ability because they lack familiarity with that particular kind of test format. Many Japanese students find cloze tests difficult because they have not yet learned how to perform at their best on such tests. Many fail to realize that they must use more than just the immediately surrounding context to fill in the blanks. And the opposite problem can occur. In this country, many students have been trained in taking multiple-choice format tests and can often outperform their ability in the matter being tested. In both of these cases, and in additional ones as well, the format or method of the test has an influence on the scores the test taker receives. In fact, it is impossible to escape such test method effects. Regardless of the test method we use, it will influence the scores of the people who take that test.

Since we cannot escape test method effect we have to try to control for it. And we can only do this by having the test takers demonstrate their ability by using a variety of test formats. In most classroom tests it is not practical to include a number of different test formats in a single test. But it is possible to use a variety of test methods over the course of instruction. Testing grammar, for example, by using multiple-choice items, completion items, and by examining the accuracy of what the student writes or speaks, provides both better coverage of the construct grammar and gives the students the opportunity to demonstrate their knowledge of grammar in a number of different test formats. By using different test formats we hope that the test method effects will begin to cancel each other out.

Concern about test method effects is one of the reasons for the insistence on authentic test tasks. It is difficult to see how the skills that allow a student to outperform his or her ability on a multiple-choice test will be useful when that student is faced with the task of communicating in the real world. However, we would expect a test task that closely resembles what test takers will have to do when they use language in the real world will have method effects that will also be useful in doing that task outside the test situation.

Getting the Weights Right

Clearly, in any task (or construct) some sub-skills are more important than others. And we would like to have our overall evaluation accurately reflect the various levels of importance of the sub-skills. However, some sub-skills are easier to test than others and so there is a natural tendency to test what is easy to test and avoid testing what is difficult to test. But this often means that we end up with a weighting of sub-skills that does not really reflect the way those skills contribute to the successful performance of the overall skill that we are interested in. And, as Messick points out, a failure to get the weighting right poses a threat to validity.

Keeping Bias At Bay

If there is something about our test that causes a portion of our test takers perform in a less than optimal fashion, we have test bias. Our test, for some reason, is biased against these test takers. Bias in language testing terms is not something that is intentional. It merely means that our test disadvantages certain test takers. The test designers and writers had no intention of creating an unfair situation for these test takers but something in the test caused this effect. And it is this issue of test bias that the second half of Messick's fifth question is addressing. Test bias is a threat to validity.

Generalizability

Messick's last question uses a term that you may not be familiar with —

generalizability. Any test is a sample of the tasks that could possibly be used. It is always the case that, on the basis of the test takers' performance on the tasks that are in the test, you must decide their performance on the thousands of tasks that they will have to perform in the real world. Test tasks can be different than real world tasks or very good approximations of them, but they can never be exactly the same in either nature or scope. Even if the test task seems to be exactly the same as what you would do in the real world, there is always the difference engendered by the fact that the task is used as a test. Driving a car with the department of motor vehicles tester sitting next to you is a very different experience from driving after you have received your license. And the 'driving' tasks you do for your test are only a few of the things that you will have to do when you actually get behind the wheel as a licensed driver. But those tasks have been carefully selected to allow the inspector to decide how you will perform in the real-world driving situation. The same situation holds for language tests.

A test provides information that hopefully will allow us to decide how the test taker will perform in non-test (real world) situations. In technical terms, we must generalize from the test behavior to behavior in the real world. The more generalizable the test results are, the better the test. If a vocabulary test only told us whether the students know the words that actually appear on it, it would be of limited value. Even if we were only using the test to check if the students have mastered the words taught in a particular course of study, we would still have to generalize the results of the test. It is usually impossible (and probably a waste of time) to test every word that was taught. But even if it were possible to test every word taught, we are not interested in learning only if the student can use the word on a test but rather if he or she can use it to communicate in the real world. We can never escape from the necessity of generalizing. Therefore, it is important to select test tasks that yield results that are generalizable.

Kinds of tests

Recall that we began by defining a test as a device to gather information to help us answer some question. That is, all testing begins with a question that we need to find an answer to. One way to classify tests is by the kinds of questions they are designed to help us answer. If we are faced with deciding who should be allowed to enter a particular course of study or academic organization, we need a **screening test**. These tests are sometimes called **selection tests** or **entrance exams**. If we must divide the students we have accepted into classes, a **placement test** would be useful. If our task is to discover just what the students' strengths and weaknesses are, a **diagnostic test** called for. If we want to know the degree to which the students have learned what we taught them in a particular course of study, we would build an **achievement** or **mastery test**. The content of an achievement test is determined by the course of study. In fact, the rule for creating a good achievement test is to test what you taught in the way you taught it. But we have a different situation if our task is to find out what level of ability a person has in the language generally or in some aspect of it. In building such a **proficiency test** we need to avoid basing our test content or method on any particular course of study.

These labels give us a useful way to categorize tests but just knowing that we want to build a test that has a particular label does not help us much. Knowing the category of test does not tell us what should be tested. If, for example, we have the task of designing a screening test, we must decide what skills we need to test in order to acquire data that will allow us to divide the students in the way we want them to be assigned to classes. In this particular case we would have to ask what skills are needed at each level of instruction and select or design a test that measures these skills. In the English Language Program (ELP) at International Christian University (ICU) it was found that tests of listening comprehension, reading comprehension, grammar, and vocabulary are the most useful in determining which of the three levels of the program is most suitable for each student. Because the results are needed in a hurry and since it is believed that the negative washback of a multiple-choice test is quickly offset by the more naturalistic measurement that is done in the classrooms and in the program-wide tests that start as soon as classes begin, a completely multiple-choice format test is used for screening. However, a second step in the screening process is to look at additional information on those students near the two cut-off points. These students are interviewed before a final decision is made on their placement.

In addition to classifying tests according to their broad function, there are several other classifications that are even more important because they have greater consequences for test design and analysis. There is an important distinction to be made between **passive** and **productive** tests. In passive tests, the test takers do not need to actually produce the language being tested but in production tests they do. The interview that is part of the second step in ELP's screening process is an example of a production test. The students have to use English to respond to the interviewer's questions. The multiple-choice test of the first stage is the classic example of a passive test.

However, the classification of tests that probably is of the greatest consequence is the distinction between **norm-referenced** and **criterion-referenced** tests. A norm-referenced test compares the students taking the test to the other students taking the test or to a group (called the norming population) that have taken the test in the past. A criterion-referenced test compares the students to some standard or criterion. The TOEFL is probably the best-known example of a norm-referenced test in our field. The driving license test is a good example of a criterion-referenced test. A norm-referenced test is designed to spread the test-takers out across the range of the skill being tested. Since a criterion-referenced test is designed to see who has mastered a particular skill and who hasn't, the test takers tend to cluster together into two groups — those who have mastered the skill and those who haven't. Whether test results spread out the test takers or cluster them, determines what sorts of statistical procedures are appropriate for determining validity and reliability. Correlation is usually not an appropriate procedure if the results are too tightly clustered.

Classical Test Theory

Number-right score

The traditional way to score a test is to count the number of correct answers or to total the points given for each task on the test. Obtaining this number-right score is so obvious a way to deal with test papers that it may seem to some to be the only way to go about the job. But in recent years new ways of determining a test taker's score have been devised and so we now speak of the number-right score understanding of language testing as classical test theory. In the final section we will introduce these newer ideas but we first must understand the traditional way of thinking about test results.

True Score

No test result can be perfectly accurate. Even if our test is highly valid and reliable, the test takers' results will not be a perfect reflection of their ability in the skill being measured. Some students may be nervous and this may not allow them to demonstrate their true ability on the test. Others may be relaxed and confident and their feeling of well-being allows them to do somewhat better than their true ability would normally permit them to do. Students sitting next to the window may be distracted by what is happening outside while other students will be able to focus their attention completely on the test. All these factors and perhaps a hundred others may cause a test taker's score to be less than a perfect measure of his/her ability. In language testing terms we speak of the score the test taker actually received as his/her **observed score** and the score that would perfectly reflect his/her real ability in whatever is being tested as his/her **true score**. The difference between the person's true score and his/her observed score is simply called **error**.

The problem we face is how to determine a person's true score. We have the observed score but we need to figure out how much error is present. In classical test theory we try to do this by making two assumptions about the error component. We assume that it is **random** and that it is **normally distributed** around the observed score. Random means without any pattern. If there were a pattern such as the true scores consistently being lower than the observed scores, the error would not be random. A normal distribution is merely the pattern of results that often occurs if we measure a large number of just about anything. If you were to measure the height of all the students in your school, you would find that most were close to the average but a few would be much shorter than average and a few others would be much taller than average. If we plotted our measurements on a bar graph we would get the famous bell curve or normal distribution. The reasoning behind the assumption of Classical Test Theory that error is normally distributed is that, if we measured the test taker's ability again and again and again, these measurements would fall into this classical 'normal distribution' pattern.

The normal distribution has several well-known statistical characteristics and, since we assume that error is normally distributed, these characteristics can be used to estimate what is called the standard error of measurement (SEM). The standard error of measurement is calculated by multiplying the standard deviation of the test scores times the square root of one minus the reliability coefficient.

$$\text{SEM} = \text{SD of test scores}\sqrt{1\ \text{Ğreliability}}$$

We can see from this formula that we will get a small SEM if the SD of the test scores is small and the reliability is high (close to one). This is reasonable. The more reliable the test is the less wobble (error) there will be in the scores on that test. It may not be so obvious why the SEM should increase with the SD of the scores. But as Henning mentions (74), this is because the observed score variance (what the SD indicates) contains error variance.

The SEM is expressed as a number with plus and minus in front of it. For example, the reported SEM of the traditional pencil and paper TOEFL was ± 15. This means that a person's true TOEFL score falls somewhere between 15 points below the score he/she received (their observed score) and 15 points above it. His/her true score is most likely nearer the observed score than the ends of the 30-point area in which it might lie. And there is even a small chance that the true score is actually more than 15 points above or below their observed score. By reporting the SEM, we put those who must interpret the test results on notice that the score which is reported is not an exact measure of the candidate's ability. Many times university admission officers forget this and rigidly apply a cutoff score to decide who gets in and who doesn't. A person with 549 is denied admission and one with 551 is accepted. But, with an SEM of ± 15, these TOEFL scores do not really support such decisions. When we take the SEM into account, we have to admit that the TOEFL is saying that there is not much difference between these two candidates and the difference we see is probably a matter of chance, not a difference in ability.

Correlation and Classical Test Theory

The primary statistical tool for estimating reliability and validity in classical test theory is correlation. The **internal consistency** of a test is determined by seeing how half the test (usually the odd numbered items) correlates with the other half (the even numbered items). Or the same test is given to the same group of test takers twice to determine **test-retest reliability** by computing the correlation coefficient of the two sets of results. **Rater reliability** is the degree to which different raters agree on the scores they give for the same performance. **Concurrent validity** is computed by correlating the results of a new test with the results of the same group of test takers on an established test. **Predictive validity** is determined by checking to see how well the test results correlate with performance in some real-world activity that the test is supposed to be a measure of. The list could go on. But there are drawbacks to the classical test theory. We will see some of the practical problems with this approach to language testing in the next section but there are theoretical considerations as well. Statisticians believe there are better ways of doing the job that needs to be done. The manual of the statistical package SYSTAT puts it this way. "[T]he latent trait model is generally regarded by test experts to be superior to the classical model… Indeed, despite the popularity of the classical model (and its associated statistics such as Cronbach's alpha, item-test correlations, and factor loadings) among nonprofessionals and applied researchers, the latent trait model is the one used by well-known psychological and educational testing organizations" (II-446). But before we begin to discuss the latent trait models mentioned in this quote, we need to first consider how test items are evaluated in the classical model.

Item Analysis in Classical Test Theory

Item analysis is undertaken to determine two major things; the difficulty of the item and how well the item discriminates between those who have more of the ability being tested and those who have less. **Item difficulty** is simply the percentage of test takers who got that item correct.[5] **Item discrimination** is a bit more complicated. The easiest way to grasp the

[5] Some people call the percentage correct the item **ease** and reserve the term **difficulty** for the percentage

meaning is to look at the way item discrimination was traditionally calculated. In the traditional way of computing discrimination, the test takers are first ranked according to their total score. Then the results of the top 27% of the test takers are compared to the bottom 27%. We can see how this is done if we look at Table One below. The table contains the results of 11 students on 8 items. A 1 means that the test taker got that item correct and a zero means an incorrect answer was given.

Testees	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Score
A	1	0	1	0	0	1	1	1	5
B	1	0	1	0	0	1	1	1	5
C	1	0	1	0	0	1	0	1	5
D	1	0	1	1	0	1	0	0	4
E	1	0	0	1	1	1	0	0	4
F	1	0	1	1	0	0	0	1	4
G	1	0	0	1	0	1	1	0	4
H	1	0	0	0	0	1	1	1	4
I	1	0	0	1	0	1	0	0	3
J	1	0	0	1	0	1	0	0	3
K	1	0	0	1	0	1	0	0	3
Diff.	1	0	.45	.64	.09	.91	.36	.45	
Disc.	0	0	1	-1	0	0	.67	1	

If you look at the right-most column you will see that the 11 test takers have been ordered according to the number of items they got correct. The item difficulty was calculated by adding up the number of ones and dividing by the number of test takers. Item 3, for example, was answered correctly by 5 out of 11 students. Dividing 5 by 11 yields 0.45.[6] Notice that two of the items are extremely easy. Everybody got Item 1 correct and 10 out of 11 got Item 6 correct. Two others were quite difficult. No one got Item 2 correct and only one person got Item 5 right.

In this particular case, the item discrimination can determined by comparing the top three test takers (A, B, and C) with the lowest three (I, J, and K). Since there are 11 test takers all together, three of them constitutes 27% of the total. For each item, we compute the discrimination index by subtracting the number of correct responses in the low group

incorrect. It doesn't make any real difference whether you use the percentage correct or percentage incorrect as long as you are aware which is being used in a particular discussion. If the percentage correct is used, easy items have high numbers (close to one) and difficult items have low numbers (close to zero). The reverse is true for percentage incorrect.

[6] In filling out the table, any symbol could be used for a correct answer as long as it is different from the symbol for a wrong answer. But 1 and 0 are very useful, particularly if you use a spreadsheet program such as EXCEL to record the test takers' results. In EXCEL you could compute the difficulty of all the items by inserting the formula which takes the sum of the ones and divides it by the number of test takers. The formula only needs to be typed in once (at the bottom of the left-most item in your matrix). The Fill Right command will do the same calculation for the remainder of the items.

(students I, J and K) from the number of correct responses in the high group (students A, B, and C) and dividing this result by the number of people in one of the two groups (in our case 3). The last row in Table One shows the result of this calculation. As you can see, these eight items do not show very good item discrimination. Four of them have a discrimination value of zero. Any time the number of correct responses in the high group equals the number of correct responses in the low group, the item discrimination index will be zero. Item 4 shows that discrimination indexes can be negative. Any time the number of correct answers in the low group exceeds the number of correct answers in the high group, the item discrimination will be negative. A negative item discrimination value indicates a serious problem with that item. It is clearly working at cross-purposes to the rest of the items in the test. Such an item claims that those test-takers who the other items indicate have more of the ability being tested have less and those the other items claim have less of the ability have more. Such items need to be revised or discarded.

We can now see how to calculate item difficulty and discrimination, but why bother? We take the trouble to do these calculations because we want our test to be a good one. We don't want our test to be either too difficult or too easy for the test-takers. It is not only that a test that is too difficult will discourage students and one that is too easy will give them a false sense of confidence. The most important reason is that we want the test results to accurately reflect the ability of the test-takers. A test that is too difficult only tells us that all the students are of a level of ability below what the test is measuring. A test that is too easy tells us that all of the students are of a level of ability that is higher than what the test is measuring. In other words, in either case, the test does not fit the ability level of the test takers[7]. And we want our test to be able to measure the ability of all who take it. This means that our test should be able to divide the test-takers who have more of the ability from those who have less. If our task is to divide the test-takers who have more ability from those who have less, extremely difficult or extremely easy items are just about worthless. For finding out which test-takers have more and which have less of the ability being tested, only items 3, 7 and 8 look reasonable.[8]

[7] In theory, there is one case in which a perfectly easy test would be acceptable. If we were testing to see if students had mastered what they had been taught, a perfect learning situation (and a test that accurately measured that learning) would result in all the students getting all the items correct. In practice, an achievement test which all the students could ace is more likely an indication that the teacher set the teaching goals too low or that the test did not measure true mastery. I have never taught a class in which all students mastered everything that was taught. This may be a reflection of my less than perfect teaching skills but a more reasonable explanation is that it merely reflects the differences in both native ability and motivation that will be found in any collection of students.

[8] Some teachers find this focus on trying to discover which test-takers have more and which less of what is being tested unfair. They complain that we should not be judging students in this way and forcing them to compete against each other. But by using the word 'judge' and 'compete' they reveal that the problem does not lie with the test results but in the way those results are used or interpreted. If not having mastered a point that was taught or being at a level ability which is lower than some other test-taker is seen as a judgment against that person, implying that he or she is not as good (in some moral or individual worth sense) we have the problem that these teachers are complaining about. But test results are not required to be used in this way. We would laugh if someone claimed that a person who is 168 centimeters tall is a better person than one who is 166. We should also laugh anyone who makes similar claims about differences in language test scores. This is not to say that the

This desire to accurately separate the test-takers who have more of what we are testing from those who have less is also our primary reason for calculating item discrimination. If some items are working against the majority (that is, have negative discrimination) the test as a whole is not doing the best job it could to divide the test-takers according to their ability. And, of course, items that have low discrimination also are only weakly helping. This is why it is usual to set a lower bound on discrimination. We accept all items above a certain level (usually 0.2 or 0.3).

The description of the calculation of item discrimination given above was used only to illustrate the principle involved. If you wish to analyze the items in a small classroom test you might use the method I've described. But a more powerful statistical procedure is available. Using the point bi-serial correlation allows the data from all test-takers (not just the high and low 27%) to be used.

But lets focus our attention for the moment on item difficulty only. One of the weaknesses of classical test theory is that difficulty is determined completely by the people who take the test. We can see this if we return to Table One. The difficulty estimates we calculated were obtained by looking at all 8 test takers. But what would happen if we only looked at the first five? I've put their results into Table Two.

Testees	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Score
A	1	0	1	0	0	1	1	1	5
B	1	0	1	0	0	1	1	1	5
C	1	0	1	0	0	1	0	1	5
D	1	0	1	1	0	1	0	0	4
E	1	0	0	1	1	1	0	0	4
Diff.	1	0	.8	.4	.2	1	.4	.6	

Now compare these difficulty estimates with the estimates we got when we looked at the results of all 8 test takers.

Diff.	1	0	.45	.64	.09	.91	.36	.45	

All except the first two changed. Most increased (Items 3, 5, 6, 7, and 8), but Item 4 became more difficult.

If we were to determine the difficulty of these same 8 items using only the results of test takers G through K we would see yet another change.

issue these teachers raise is not important—it is. But the problem does not lie in the test scores themselves and it points up the importance of reporting test results in ways that discourage such misunderstanding.

Table Three

Testees	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Score
G	1	0	0	1	0	1	1	0	4
H	1	0	0	0	0	1	1	1	4
I	1	0	0	1	0	1	0	0	3
J	1	0	0	1	0	1	0	0	3
K	1	0	0	1	0	1	0	0	3
Diff.	1	0	0	.8	0	1	.4	.2	

As before, only items 1 and 2 had the same difficulty estimates.

This is an artificially small sample but it illustrates clearly that, in classical test theory, our estimates of item difficulty depend on the people who take the test. If we give our test to people with high ability in the area being tested, the items will be easy for them. If we give the same items to people of less ability, the items will appear more difficult. This means that when we try out or, as language testers say, **pretest** our test items, we need to try them out on people who are of the same level of ability as the people who we expect to take the final version of the test. This is easy to say, but it is difficult to find a suitable pretest group.

Once we have tried out the items that we would like to put into our test, we have to decide which ones are good enough to go into the final form. We have already mentioned that we should eliminate items with negative or very low discrimination power and that, except in very special cases, items of very high or very low difficulty should be avoided. But what level of difficulty should the items that we include have? The answer is, it depends. The purpose of the test and the proficiency of the group that you expect to take the test should determine the appropriate level of difficulty. Let’s look at one example of how the purpose of the test would dictate the appropriate level of difficulty of the items needed. Let’s say that we wanted to use our test to decide which test takers have at least some minimum level of ability. Perhaps that ability level is the minimum needed to succeed in a course of instruction and we want to use our test as a screening device. Those who have the needed level ability should pass and those who don’t should fail. Notice that we don’t need to know if a test taker has just barely the required level or much more than the required level. Likewise, we don’t need to know if a student who doesn’t succeed missed reaching the necessary level by a little or a lot. In this pass/fail case we would want to load the test with items of a level of difficulty that matched the necessary ability level. If we loaded the test with items of, say, 0.4 difficulty, we would divide the test takers into two groups; the 40% who got the items correct and the 60% who didn’t.

However, if our task were to find out the ability level of each member of a group consisting of people with a range of proficiency, we would choose items of varying difficulty. We would include some easy items, some of intermediate difficulty and others that were hard.

For ease of presentation we have only talked about the analysis of items that can be scored on a correct/incorrect basis, called in testing terms, **dichotomously scored**. Most

multiple-choice items are dichotomously scored but they do not have to be. It is possible to give **partial credit** for answers to such items. It is rarely done because usually the disadvantages of such non-dichotomous scoring are greater than the advantages. And test items that require the test taker to produce the language being tested can be scored dichotomously. Note that what makes scoring dichotomous is the number of possible scores an item can be given. It doesn't make any difference how much a correct answer is worth. If a test taker gets 3 points for a correct answer and none for an incorrect answer, the item is still being scored dichotomously. However, if it were possible for some answers to be given 2 points rather than just either 3 or zero, the item would no longer be dichotomously scored.

The item difficulty and discrimination of such non-dichotomous items can be calculated. The difficulty becomes the percentage of points the group of test takers obtained out of the total possible points. If 11 people took the item mentioned above, the total possible points would be 33. If we total the number of 3's and 2's awarded, we could divide this number by 33 to get the difficulty of the item. Item discrimination would be computed by subtracting the number of points obtained by the lowest 27% from the points obtained by the top 27% and dividing by 9 (the total possible points that the 3 members in either the high or low group could be awarded). Recall I pointed out that, in determining the item discrimination of dichotomously scored items, a more powerful statistical procedure that utilized the data from all the test takers could be used. However, with non-dichotomously scored items, we cannot use this point bi-serial method to calculate item discrimination using the data from all test takers. The point bi-serial calculation requires dichotomous data.

The final thing I want to mention about item analysis is the possibility and advisability of checking to see how well the distractors (wrong answer choices) of multiple-choice items are functioning. One of the hardest test writing tasks is to write good distractors, so we need to find out if our distractors are working the way we expect them to. We expect a distractor to look correct to a person who doesn't know the point being tested but wrong to the person who does. If no one selected a distractor, it is not doing its job. And the same is true if more of the high group than the low group selected the distractor. There is no set statistical procedure for checking what is sometimes called distractor efficiency. However, a rough rule of thumb followed by many test developers is to eliminate or rewrite distractors which do not attract at least 10% of the people who marked the item wrong. Distractors that are not selected by significantly more of the low group than by the high group should also be rejected. The likelihood of all distractors being acceptable under these rules is so low that many test developers pretest items with an extra distractor, expecting at least one to be weak.[9] Most classroom tests are not pretested and do not have to meet such rigid standards if they are, but it

[9] It is also common practice to pretest more items than you will actually need in your test. The pass rate (the percentage of acceptable items in the items pretested) varies with item types (from my experience a larger percentage of grammar items are found to be acceptable than vocabulary or reading comprehension items) but a rough rule of thumb is to pretest 50% more items than you think you will need in the test itself.

is wise for the developer of these tests to check to see how the distractors are functioning. It will show which ones ought to be rewritten if the item is to be used again. It should also help in learning how to write better distractors.

Standard Scores

The basic reason for converting **raw scores** to **standard scores** is to make the results of different tests (or different versions of the same test) comparable. Teachers typically give a number of tests or make a number of evaluations of their students during a term of instruction and must have a way of combining the results of these various tests and evaluations into a single assessment or grade. Even a single test with several subtests raises the same question. How shall we combine the part scores into a meaningful total or overall result?[10] One solution is to merely add up the raw scores, but this approach has a number of serious drawbacks. The most obvious one is that this approach means that we get 'self-weighting'. Each set of raw test results determines the weight that that test will have in the total. To see why this is so, look at the results displayed in Table Four.

Student	Raw1	Rank1	Raw2	Rank2	Raw3	Rank3	Raw Sum	Raw Sum Rank
Kenji	8	1	80	1	60	3	148	3
Taeko	6	2	78	2	70	2	154	2
Yoko	4	3	76	3	80	1	160	1

Although Yoko was the lowest in two out of the three tests, she ends up ranked first overall. Kenji was the top student in two out of three tests and yet he ends up ranked at the bottom overall. If we look at the **mean** and **standard deviation** of the three tests, the reason for the strange overall ranking becomes clear.

Student	Raw1		Raw2		Raw3
Kenji	8		80		60
Taeko	6		78		70
Yoko	4		76		80
Mean	6		78		70
SD	2		2		10

It is obvious that the standard deviation is what determines the weight of each test in the final

[10] One very important issue that we do not have time to deal with here is whether it is really meaningful to combine the various scores. If the various tests or parts of a test are measuring quite different abilities, any attempt to combine the scores will produce a meaningless result. A person's weight is a useful figure and so is that person's height, but a combination of the two figures will not be of much help. In fact, the combined figure will be misleading. A very tall but relatively light person could have the same 'score' as a very short but heavy person. The same sort of misleading total can come out of the combination of language test results. Therefore, the first question we must ask is if it is meaningful to combine the results of the different measures we have. In many cases, more than one overall score or a profile of abilities is needed.

ranking. If the mean were the determining factor, the second test would have the greatest weight, but that clearly is not the case. Kenji was the best in test two. But he was at the bottom of test three and this is the test that determined the overall ranking because its standard deviation was five times that of either of the other tests.

Obviously, the situation shown in the first table above is not the sort of outcome we would like to have. But we can now see how to avoid such a counter-intuitive result. All we need to do is to make sure that the standard deviation of each of the tests we combine is the same. And the most common type of score modification does just that.

There are three common standard scores. The principle underlying all three is the same. The only difference is the value each uses for the 'standard' mean and 'standard' standard deviation. The standard score most used by statisticians is what is called z-scores. Since the computation of z-scores makes the principle underlying standard scores clear, we will explain the math involved and then convert the three test scores for Kenji, Taeko, and Yoko to see what effect combining z-scores rather than raw scores will have in the example we looked at above.

We compute z-scores by first subtracting the mean of the test from the result that each student got. Then we divide this by the standard deviation. Let's try this with the test results given in the first table above. On Test One Kenji got 8. The mean was 6, and the standard deviation was 2. This means that Kenji's z-score for Test One is 8 minus 6 divided by 2 [(8 – 6)/2 = 1]. Taeko got 6, so the calculation for her is [(6 - 6) / 2 = 0]. For Yoko, the z-score calculation for Test One is [(4 – 6) / 2 = -1]. Table Six gives the raw scores and z-scores for each of the three tests, the sum of the three raw scores and the three z-scores, and the rank of the students based on total z-scores.

Students	Raw1	Z 1	Raw2	Z 2	Raw 3	Z 3	Raw Sum	Z Sum	Total Z Rank
Kenji	8	1	80	1	60	-1	148	1	1
Taeko	6	0	78	0	70	0	154	0	2
Yoko	4	-1	76	-1	80	1	160	-1	3

We can see that, if each set of test scores has the same mean and standard deviation (which is what converting to z-scores accomplishes), the students get ranked in a way that seems consistent with their results on all three tests and not just one.

By subtracting the group mean from each score we have a set of scores with a mean of zero. By dividing the result of subtracting the group mean from each score by the standard deviation of the group we get a z-score with a standard deviation of one.[11]

[11] The z-scores in our example above are all whole numbers but in actual practice z-scores range from three to

As we said, the other kinds of standard scores differ only in what the mean and standard deviation are set at. If we set the mean at 50 and the standard deviation at 10 we get the standard score that in Japan is called hensachi. The CEEB scale (the one used most commonly to report test results in the US) has a mean of 500 and a standard deviation of 100. Some commercial tests, such a TOEIC, report scores using their own standard. This means that they have set their own mean and standard deviation.

Recall that this whole discussion started in order to explain why it is not a good idea to combine raw scores from different tests. We saw that, if we do that, we may get counter-intuitive results because of self-weighting. The weight of each test in the combined total is determined by the size of its standard deviation. We don’t want tests to weight themselves but there are times when we would like to give more weight to one test and less to another in putting together an overall score. And we can do this if we first convert all the raw scores to standard scores and then multiply these scores by the weighting factor we think is appropriate. If we think that the listening component should have twice the weight of the grammar and vocabulary components, we can multiply the listening scores (expressed as standard scores, of course) by 2 before combining them with the scores on the other tests (also expressed as standard scores).[12]

The need to combine scores into a meaningful overall measure is not the only reason for converting raw scores to standard scores. Such a conversion is also necessary if we want the results of one administration of a test to be comparable with those of a parallel form of the same test. For example, we would like each administration of our entrance exams to yield the same results. The people who pass this year should be of the same ability as those who passed last year. Or the people who passed one version of the exam should be of the same ability as those who passed a parallel version. Yet, we know from some highly publicized Nyushi Center results from supposedly parallel forms of the same test (one for present high school students and the other for those who graduated from high school in previous years) that there is reason to believe that this doesn’t always happen, even if we try to make the tests as similar as possible. And there are equally strong reasons for doubting that entrance exams for the same university are of the same level of difficulty year in and year out. Each year the test is different and so are the test takers.

But there are ways to assure that this year’s test is of the same difficulty as last year’s

minus three. Statisticians prefer to use z-scores because they are easy to calculate and use in other calculations but scores are rarely reported to test takers in this form. The reasons should be obvious. Unless you understand how z-scores are calculated and know how to properly interpret them, you will be shocked to receive a negative test score or have to wonder what a score such as 0.45 could mean. That is why the other types of standard scores have been developed.

[12] Although it is usually not possible when we are combining classroom tests into a final grade, it is best to allow the purpose of the test (rather than the test developer’s or teacher’s opinion) to determine the weighting of the various part scores. If, for example, you are able to try out a screening test on a group which includes both those who have succeeded in the course for which you are screening and those who have not, you can check to see what weightings best separate the two groups.

or that one form of this year's test is equally difficult as another form. Ideally, we would like to give both forms of the test to the same group of people or the same test to both groups of test takers, but this is rarely possible. But it is possible to do **test equating** by using what are called **anchor items** to determine if this year's group of test takers is of the same ability as last year's[13]. Anchor items are items from the former test embedded in the present one. By comparing the performance of this year's test takers on these items with the performance of those who took the test last year, we can see if one year's test takers are better than the other or if both groups are the same. If the later is the case, we can use this year's results as is. But, if the two groups are not the same, that is, one group performed better than the other on the anchor items, then we have to adjust the scores of this year's group so that they are parallel with last year's group. If this year's test takers did better on the anchor items than last year's group did, that means that this year's group will have a higher mean than last year's group. Lets say that the mean score of this year's group on the anchor items is 10% higher than the mean of last year's group on the same anchor items. If we convert the results of this year's test takers to standard scores using, not the mean of their raw scores, but the mean of last year's group on the anchor items, this would have the effect of raising the scores of this year's group. And this is the result we want because their performance on the anchor items showed that they are better than last year's group.[14]

We will introduce another way of equating scores on different versions of the same test when we discuss item response theory below.

Reporting test results

Recall that I said that a test was valid to the degree to which it provided information to usefully answer the question that was posed. And a test must be valid, not only for the one who designs and builds it, but also for those who take it. In other words, it is not enough for the test writer and/or teacher to be able to correctly interpret the results, each test taker must be able to make sense of his or her results. But just what does it mean to 'make sense of results'? Humans have a built-in desire to make sense of things, so test takers will undoubtedly take the test results to mean something. But unless that 'something' is the correct understanding of the results, the test is, for that student, not valid.

There are two common ways in which test reports are misunderstood. The person who receives the report often believes the score is more exact than it actually is. As we saw when we discussed standard error of measurement, any score contains some error. If we simply tell one student that he got a 71 and another that she got a 69, the first one will

[13] The use of anchor items can only be justified if the items used are secure. We could only use anchor items to check to see if this year's entrance exam is of the same difficulty as last year's if we are willing to give up the practice of releasing each test after it has been administered.

[14] In practice we would compute their standard score by taking into account both the difference in the means of the two groups on the anchor items and any differences in the standard deviations as well. But since our task here is to explain the principle involved in test equating, we won't discuss how this is done but direct the interested reader to the books and articles on test equating listed at the end of this chapter.

probably think that he is better than the second one. The second student may well think that she has failed to reach the passing level of 70. But, as we saw, a SEM of as little as 5 would make both of those conclusions less than certain. One way to help test takers avoid thinking that their test score is more accurate than it actually is, is to report both the score and the SEM. One testing organization in the US instructs students to plot their score on a printed scale. Next it tells them to count out from this point in both directions the number of points of the SEM. It then explains that their true ability falls somewhere between these two points, most probably close to the score they received.

Another common error that is made in interpreting test scores is to assume that the scale on which the score is reported is from zero to 100 and that each unit on the scale is the same size. The two students above who got 69 and 71 would have made this assumption. But the scale the test reporter is using may not be a zero to 100 one. If these scores were T-scores (hensachi in Japanese) these results would be very good indeed. This is so because T-scores have a mean of 50 and a standard deviation of 10. These results would be roughly two standard deviations above the mean. These students would be in the top 2 or 3% of all who took the test.

The Latent Trait Model

Item Response Theory (IRT)

In our discussion of item difficulty in classical test theory we said that difficulty was defined as the percentage of people who got the item correct. As we pointed out then, this means that the difficulty estimate we get is completely dependent on the people who took the test. This population dependent difficulty estimate leaves us with a number of problems. To escape these problems we need a population free estimate of item difficulty. To illustrate how we might determine item difficulty that isn’t dependent on the first people who take our test let me set up an artificial situation in which we have 9 test takers, each with a different level of ability in the skill that we are testing. Let’s assume also that we have 10 test items, each of a different level of difficulty. We’ll leave aside for now the question of just what it means to say that each item has a different level of difficulty. Hopefully that will become clear as we go through this explanation. Let’s make one more set of assumptions. Item 1, the easiest item can be answered by all 9 test takers. Item 2 can be answered by all but the weakest test taker. Item 3 can be answered by all but the two weakest test takers and so on. That is, we have set up the situation shown in Table Seven. A one means that test taker answered that item correctly and a zero means that it was answered incorrectly.

Student/Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A (weakest)	1	0	0	0	0	0	0	0	0	0
B	1	1	0	0	0	0	0	0	0	0
C	1	1	1	0	0	0	0	0	0	0
D	1	1	1	1	0	0	0	0	0	0
E	1	1	1	1	1	0	0	0	0	0
F	1	1	1	1	1	1	0	0	0	0
G	1	1	1	1	1	1	1	0	0	0
H	1	1	1	1	1	1	1	1	0	0
I (strongest)	1	1	1	1	1	1	1	1	1	0

As I said, this is an artificial situation. The data from actual test administrations are never this straight forward. But this ideal situation is exactly what we would like to create. We would like to have the situation in which we can know that any test taker with the same ability as student A will get the same results. Item 1 will be answered correctly and the other items will be answered incorrectly. A test taker of the same ability as student E will get items 1 through 5 correct but will miss items 6 through 10. In other words, we would like to have fixed item difficulties that did not change when we gave the test to different people. We would like our measure of item difficulty to be like the scales we know in the physical world. A kilo of potatoes contains 1000 grams just as a kilo of flour does. A meter of cloth is the same length as a meter long board. In the physical world we have scales that are the same regardless of what we are measuring and we got such scales by fixing the interval between points on that scale. The French revolutionary government decided that a meter would be 1/1,000,000th of the distance from the equator to the North Pole. The length of the interval is arbitrary. To this day, the US persists in measuring distance in inches, feet, yards, and miles. But the distance from one mark on an American yardstick to the next mark is set. The distance between the mark for one inch and the mark for two inches is exactly the same as the distance between the mark for 24 inches and 25 inches. But how can language testers create such a scale for item difficulty?

Item Response Theory or IRT is an attempt to create such a scale. But, before I begin to explain just how IRT works, I should point out that, as its name indicates, it is a theory. We speak of the metric system. In the case of physical measurements we can set the intervals of the scales we use. However, in our attempt to measure abstract entities such as item difficulty, the best we can do is to define difficulty in a way that allows us to measure it using a scale that we can justify on theoretical grounds.

IRT assumes that if we test enough people on a group of items of a sufficiently wide range of difficulty we should be able to arrive at a table like the one above. That is, we should be able to fit all the items measuring the same underlying trait or ability on a scale of difficulty that is represented by the columns of the table and we should be able to fit all the test takers on a scale of ability that is represented by the rows of the table. The mathematics involved is somewhat complicated and the calculation procedure time consuming, but essentially, by a process of iteration, the best fit between the difficulty of the items and the ability of the test takers is determined.

In classical test theory, we first allowed the test takers to determine the difficulty of the test items. Recall that the difficulty of an item was the percentage of the test takers who got it correct. But we ranked subsequent test takers by using the item difficulty that had been determined using the first group of test takers. IRT does not let the test takers determine item difficulty. Instead, it looks at both item difficulty and test taker ability together. The IRT procedure begins by making its best guess at the difficulty of the items and then checking to see how closely the resulting order of test taker abilities resembles the ideal of the table above. It keeps adjusting its guesses until the table of items and test takers is as close to the ideal form as possible.

In this process we may discover that some items don’t fit very well. That is, they don’t help us in the process of arriving at the ideal form of the table. These are items we would not want to use in the future. It is even possible for some test takers not to fit. This might seem like a defect of IRT but it is actually an advantage, because it indicates that that test taker is clearly different from the others. Perhaps that test taker cheated or has a set of abilities that is unusual. Most learners of Japanese as a foreign language who come from countries where kanji are not used are generally more proficient in spoken Japanese than they are in the written language. Students from countries that use kanji show the opposite pattern. And, not surprisingly, an IRT analysis of an Australian test of Japanese that was first tried out on Japanese as a foreign language learners from non-kanji countries showed that the results of persons from kanji using countries didn’t fit.

Computer Adaptive Testing

We learned earlier that it is important to give students a test that is at the appropriate level of difficulty. That is, the difficulty of the test should match the ability level of the test takers. If the test is too easy, not only will it bore the students, but we will not be able to learn about any differences in ability among those who take it because everybody will be able to answer the questions correctly. A test that is too difficult for those who take it is also of little use. It will discourage the students because they will not be able to answer any of the questions correctly and it will not give us any information about differences in ability that may exist among the test takers.

However, in the traditional pencil and paper test it is not easy to match the difficulty of the test to the level of ability of the test takers. The first problem is that, before they take the test, we cannot be sure of the level of ability of the test takers. Even if we have a good idea of the range of abilities in the group that will be taking our test, the best we can do is include items that cover the whole range of ability. But this means that, for any one test taker, some of the items will be very easy and some impossibly difficult. And, since the test must include items of the whole range of difficulty, it is bound to be longer than needed for a particular student. Test takers with a high level of ability will find all but the most difficult items easy and boring. Test takers at the other extreme will find most items beyond their ability. Even a test taker in the middle range of ability will find some items too easy and others too difficult.

What we would like to be able to do is deliver a test that is tailored to the level of ability of each student — and computer adapted testing allows us to do just that. The idea is simple. We have a pool of items and the computer draws items from this pool that fit the level of ability that that particular test taker shows. The strategy is to begin with items of middle level difficulty. Test takers who get these items correct are then presented with items of greater difficulty. If those test takers can answer these items correctly, they are given yet more difficult items until items that they cannot answer are encountered. The same procedure is followed with all test takers. The next set of items that are presented to the test takers are determined by their responses to the earlier ones. Test takers of lower ability are given ever easier items until they encounter a set that they can consistently answer correctly.

This approach is possible because IRT allows us to place items on a scale of difficulty that is not affected by the persons who take the test. That is, we have a scale of difficulty that doesn’t change (in statistical terms, is invariant) and, if we have a pool of items that represent the full range of this scale, we can determine a test taker’s ability by checking to see which items he/she can answer correctly and which he/she cannot. It is like trying to see how high each member of a group of people can jump by first placing the bar in the middle position. For those who clear this height, the bar is raised and we continue to raise the bar for each of them until it is at a level that that person can no longer clear. For each of those who fail to clear the bar in the middle position, we keep lowering it a notch until it is at a height that that person can clear. We can do this because the scale we use to measure height (feet or meters) does not depend on the people doing the jumping. The notches at which we place the bar correspond to the items we have in our pool. Just as we have to have notches that are low enough so that the weakest jumper can clear the bar at that height and high enough so that we can determine the height that the strongest jumper can clear, in our test pool we need items that span the full range of ability of the group we are testing.[15]

Notice that in computer adaptive tests we have moved completely away from classical test theory. Comparing the number of items each test taker got correct is meaningless because the items that each student answered are not the same. What we can report is the highest level of difficulty at which each student was able to correctly answer the items presented to him or her. There is no such thing as a raw score. We have only the student’s position on the IRT scale of difficulty. From a measurement point of view, this is a great step forward, but test

[15] It is recommended that the pool should contain not only items whose difficulty cover the full range of ability of the group to be tested but also items that reflect the number of people at the various ability levels within the group. This usually means building a bank or pool of items whose difficulties have a normal distribution.

takers may have difficulty accepting this change. It may seem unfair to them. Since they do not understand the measurement principles involved they cannot help feeling that it's unfair to give tests of different difficulties to different people. However, if they understand that the different difficulties of the items in the test are like the notches on the high jump bar (that is, all part of a single continuous invariant scale) the test takers should be able to overcome their feeling that computer adaptive tests are unfair.

It is important to help students realize that computer adaptive testing is fair because such testing has a number of advantages for both test takers and the users of test results. There are practical advantages and educational advantages.

One practical advantage is the saving of test time. A computer adaptive test usually requires less time to administer because the test takers do not have to deal with items that are clearly too easy or too difficult for them. The computer supplies the test taker with the items that will be most useful in deciding just what his/her level of ability is. This feature of computer adaptive tests also provides an educational advantage. Test takers will neither be bored doing many items that are too easy for them nor discouraged by having to do a lot of items that are too difficult for them. That is, the test will be challenging to the test takers and this should increase their level of motivation

Another practical advantage that may not be obvious to test takers is that they will be able to retake computer-adaptive tests without waiting for a new version of the test to be developed. At present, a test such as the pencil and paper TOEFL can only be offered a limited number of times each year because it requires time to develop a new form. Since computer adaptive testing relies on a large pool of test items and each test is tailored to the needs of that particular test taker, that person can be retested without waiting for a new form of the test to be created. The computer creates a new form of the test each time a test taker sits down to take the test.

Computer adaptive tests also suffer less than conventional pencil and paper tests do from the problem of guessing. This means that the test results provide a truer picture of the test taker's real ability. In a traditional pencil and paper test, the test taker of 'average' ability will encounter some items that are very easy and others that are impossibly difficult. Such test takers run the danger of missing some of the easy items because they are bored with such childish stuff and don't pay close enough attention to what they are doing. In other words, they don't answer some items correctly that they should be able to answer. This means that their score will be lower than what it should be. The same test takers may guess at the items that are too difficult for them. If they guess successfully, they will get a score that is higher than their true ability would warrant. In both cases, there is a gap between their score and their true ability. However, in computer adapted testing, items are presented in sets of roughly equal difficulty. This means that guessing and errors from inattention are easier to spot and, if there is any question about the person's real ability, more items of appropriate difficulty can be administered until the doubt is resolved. This reduction in the guessing factor in computer adaptive tests has the practical result of allowing such tests to produce an ability estimate that is more accurate.

In addition to these advantages, computer adaptive testing opens the way to using graphics and video to make the test items more closely reflect real world language use situations. The use of computers to deliver test items also holds the promise of enabling test developers to test language ability in new and more interesting ways.

References:

AERA, APA, and NCME. (1999). *Standards for Educational and Psychological Testing.* American Educational Research Association.

Banerjee, J., C, Clapham, P. Clapham, and D. Wall, eds. (1999). *ILTA Language Testing Bibliography 1990-1999, Language Testing Update*. Center for Research in Language Education, Department of Linguistics and Modern English Language, Lancaster University.

Clapham, Caroline, and David Corson, eds. (1997). *Encyclopedia of Language and Education, Vol 7, Language Testing and Assessment.* Kluwer Academic Press.

Davies, Alan, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley, and Tim McNamara, eds. (1999). *Dictionary of Language Testing.* CUP.

Hambleton, R.K., and H. Swaminathan. (1985). *Item Response Theory: Principles and Applications.* Kluwer-Nijihoff Publishing.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research.* Newbury House Publisher.

Gardner, Howard. (1983). *Frames of Mine: The Theory of Multiple Intelligences.* Basic Books.

Genesee, Fred & John A. Upshur. (1996). *Classroom-Based Evaluation in Second Language Education.* Cambridge University Press.

Gould, Stephen J. (1981). *The Mismeasure of Man.* Penguin.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates.

McNamara, T. (1996). *Measuring Second Language Performance*. Longman.

McNamara, T. (2000). *Language Testing*. OUP.

Messick, Samuel. (1996). "*Validity and washback in language testing*". Language Testing Vol. 13 No. 3. Edward Arnold.

Spolsky, Bernard. (1995). *Measured Words*. Oxford University Press.

Spolsky, Bernard, ed. (1999). *Concise Encyclopedia of Educational Linguistics.* Elsevier Science Ltd.

SYSTAT. (1999). *Software Documentation*. Statistics I. SPSS Inc.

Wright, Benjamin D., and Mark H. Stone. (1979). *Best Test Design*. MESA Press.

池田央（1994）『現代テスト理論』、朝倉書店

池田央・大友賢二監修：大友賢二・笠島準一・服部千秋・法月健（訳）（1997）『言語テスト法の基礎』（Bachman, Lyle. 1990. *Fundamental Considerations in Language Testing*. OUP)　みくに出版

清川英男・濱岡美郎・鈴木純子著（2003）『英語教師のための Excel 活用法』　大修館書店

小池生夫主幹編集（2003）『応用言語学事典』　研究社

大友賢二（1972）（訳注）『英語の測定と評価』（Harris, David P. 1969. *Testing English as a Second Language*. McGraw-Hill Book Company.）　英語教育協議会（ELEC）

大友賢二（1996）『項目応答理論入門：言語テスト・データの新しい分析法』　大修館書店

大友賢二・Randy Thrasher 監修：中村優治・根岸雅史・渡辺良典・智原哲郎、安間一雄・清水裕子・石川祥一・法月健・中村洋一訳（2000）『《実践》言語テスト作成法』（Bachman, Lyle, and Adrian Palmer. 1996. *Language Testing in Practice.* OUP)　大修館書店

大友賢二監修・中村洋一著（2002）『テストで言語能力は測れるか ～ 言語テストデータ分析入門 ～』　桐原書店

靜哲人・竹内理・吉澤清美 編著（2002）　『外国語教育リサーチとテスティングの基礎概念』　関西大学出版部

渡辺真澄・野口裕之編著（1999）『組織心理測定論：項目反応理論のフロンテイア』　白桃書房

和田稔（1999）（訳）『言語テストの基礎知識』（Brown, J. D. 1996. *Testing in Language Programs*. Prentice Hall.）　大修館書店