An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests

Paula V. Engelhardt

Tennessee Technological University, 110 University Drive, Cookeville, TN 38505

Abstract:

A number of conceptual multiple-choice tests have been developed addressing many different areas of physics. These tests are typically used to determine what difficulties students have with specific content and to evaluate teaching practices and curriculum. The purpose of this paper is to provide the reader with a general overview of the key aspects of the development process from the perspective of classical test theory and critical issues that distinguish high-quality conceptual multiple-choice tests from those that are not.

Introduction

Whether you are planning on writing your own test or are conducting a research project in which a conceptual multiple-choice test will be used, it is important to understand what qualities make such a test "good." This paper in the ComPADRE volume *Getting Started in PER* will focus on the process for developing a high-quality conceptual multiple-choice test by identifying the key characteristics. The process that will be described in this paper is based on classical test theory. This is not the first paper written regarding these important issues.[1-5] There are also numerous textbooks written on the subject.[6-14] The purpose of this paper is to provide the reader with a general overview of the key aspects of the development process and critical issues that distinguish high-quality conceptual multiple-choice tests from those that are not.

The most widely known conceptual multiple-choice test is the Force Concept Inventory[15] (FCI) which was introduced in 1992. The FCI focused on students' reasoning with the Newtonian concept of force. For most instructors, the items on the FCI seemed simple, *too* simple, in fact. Many, including Eric Mazur, believed that their students would have no difficulty with this new test. Mazur writes

> When I started teaching, I prepared lecture notes and then taught from them. Because my lectures deviated from the textbook, I provided students with copies of these lecture notes. The infuriating result was that on my end-of-semester evaluations—which were quite good otherwise—a number of students complained that I was "lecturing straight from (his) lecture notes." What was I supposed to do? Develop a set of lecture notes different from the ones I handed out? I decided to ignore the students' complaints.
>
> A few years later, I discovered that the students were right. My lecturing was ineffective, despite the high evaluations. Early on in the physics curriculum—in week 2 of a typical introductory physics course—the Laws of Newton are presented. Every student in such a course can recite Newton's third law of motion, which states that the force of object A on object B in an interaction between two objects is equal in magnitude to the force of Bon A—it sometimes is known as "action is reaction." One day, when the course had progressed to more complicated material, I decided to test my students' understanding of this concept not by doing traditional problems, but by asking them a set of basic conceptualquestions.[16,17] One of the questions, for

> example, requires students to compare the forces that a heavy truck and a light car exert on one another when they collide. I expected that the students would have no trouble tackling such questions, but much to my surprise, hardly a minute after the test began, one student asked, "How should I answer these questions? According to what you taught me or according to the way I usually think about these things?" To my dismay, students had great difficulty with the conceptual questions. That was when it began to dawn on me that something was amiss.[18]

This finding surprised many physics educators and spurred both additional physics education research and the development of additional conceptual multiple-choice tests.

Why was the FCI so influential? To answer this question, we first need to understand why the FCI was created in the first place. The FCI examines students' understanding of the concept of force. Hestenes et al. found that students' ideas about force and motion were inconsistent with the views of the scientific community and that traditional instruction did very little to change these beliefs. Hestenes et al. proposed that "effective instruction requires more than dedication and student knowledge" but also "knowledge about how students think and learn."[19] This need for knowledge about how students think and learn had been the focus of physics education research through the use of interviews since the late '70s. Interviews provide a wealth of data from a relatively small portion of the total population. A way of providing some of the same information in a more time-efficient manner was needed. Multiple-choice tests were the answer.

Multiple-choice tests have some advantages over interviews. Multiple-choice tests are objectively graded, and statistical methods can be applied to the resulting data. Multiple-choice tests can be given to large numbers of individuals at one time, providing larger sample sizes and increasing the generalizability of the results. Multiple-choice tests are less time intensive than interviews. When properly developed, distractors can be based on known student misconceptions, allowing the test to serve diagnostic purposes.[20]

Multiple-choice tests also have some disadvantages over interviews. These include the depth with which the test can probe. Interviews can probe more deeply into what students are thinking about particular phenomena. Mehrens and Lehmann suggest that some students may be more skilled at recognizing ambiguities or the correct answer without actually understanding the material presented.[21] Tamir conducted a study requiring students to justify their answer choices to a multiple-choice test which

showed that students who chose the correct answer were not necessarily able to provide adequate justification. From these results, Tamir suggests that multiple-choice tests may overestimate students' knowledge.[22] Nonetheless, for most instructors' purposes, the advantages of multiple-choice testing outweigh its disadvantages.

The research purposes for using conceptual multiple-choice tests are generally two-fold: 1) to ascertain students' initial and final knowledge states and 2) to evaluate teaching, teaching methods and curriculum. Since neither of these purposes relate to student grades, multiple-choice tests have become a more palatable option. In fact this is a major requirement for using many of the conceptual multiple-choice tests developed in physics. These tests are *not* to be used as part of the determination of overall student performance in a course, but to aid instructors and students in the identification of misconceptions and areas of understanding that have yet to be fully developed. In terms of evaluating teaching, these tests can be used to determine how well a new teaching method or curriculum helps to remedy these known misconceptions and improve the quality of teaching.

As you have seen, the purposes for these conceptual multiple-choice tests are different from the more traditional classroom multiple-choice tests that we use to grade our students. A major difference between the two is in the use of results from qualitative research studies in the development of the distractors for the multiple-choice test and the detail and time dedicated to the development of individual items. The process of developing a conceptual multiple-choice test which utilizes this research data will be one focus of this paper. The other will focus on the characteristics of a high-quality conceptual multiple-choice test that one might want to use in a research study.

1. Characteristics of a High-quality Test

The development of a conceptual multiple-choice test is a multi-step process and when done properly will build in the characteristics of a high-quality test. A high-quality test should have the following five characteristics:[23]

1) Reliability
2) Validity
3) Discrimination
4) Good comparative data
5) Tailored to population

A brief definition will be given here and more in-depth discussion of each of these characteristics will follow later.

Both **reliability** and **validity** relate to the inferences that can be made from and confidence associated with the scores that are produced by the test after testing. Reliability is associated with the amount of error in the score (Does the test measure what it measures consistently?), while validity is related to the types of inferences that can be made about the scores that are obtained (Does the test measure what it says it measures?). As such, they are not a direct property of the test itself.

The relationship between reliability and validity is often illustrated using ideas from archery. The target is the bull's eye which represents what the test says it measures (validity).[24] The shots represent the scores on the test. In the cluster of hits on target A in Figure 1, you can see that they are grouped tightly together. This group of shots is very precise (reliable). In terms of the test, the test consistently measures something but this something is not what we wanted to measure. Target B shows a case where the shots are accurate in that they are mostly centered on the target but they are not very precise. In this case, we are inconsistently measuring what we intend to measure. In target C, the shots are both precise (reliable) and accurate (valid) indicating that we are consistently measuring what the test says it measures. The targets illustrate an important point which is that you can have a reliable test, but if that test does not actually measure what we intend to measure then it has very little value. Thus the ideal is to have a test that is *both* reliable and valid.

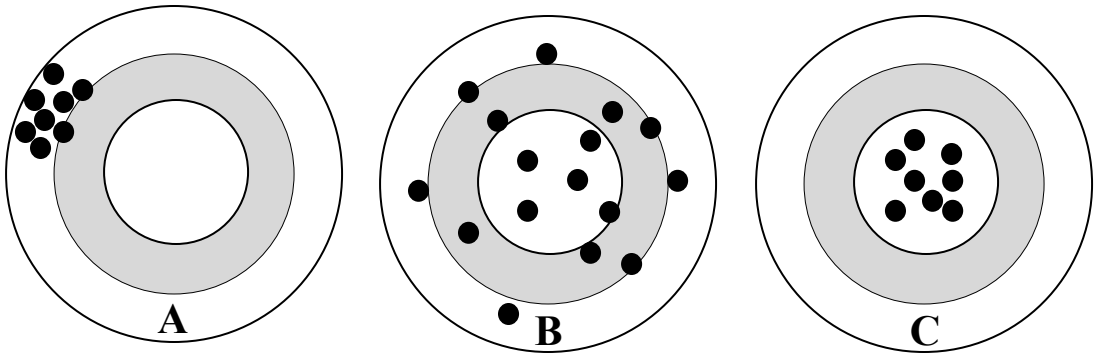

Figure 1: Comparison of Reliability and Validity

Reliability and validity both depend in part on the test's ability to differentiate between individuals taking the test. The discrimination power of the test tells us how well the test or individual test items differentiate between students who score well on the test or item and students who do not. Discrimination can be thought of as similar

to the Rayleigh criterion[25] as shown in Figure 2. The left portion of the figure illustrates a poorly discriminating test or item with a great deal of overlap, while the right portion illustrates a test or item that is discriminating well with little overlap.

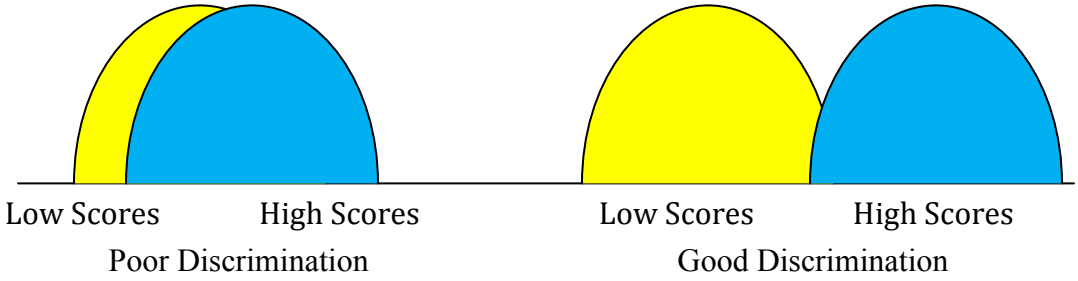

Figure 2: Illustration of Discrimination

Comparative data is obtained by the administration of the test to various groups of students for whom the test is designed. The test needs to be designed with the target population in mind; otherwise, the level of questions may not be appropriate, either too easy or too difficult.

While many conceptual multiple-choice tests exist for physics and astronomy concepts,[26] there are still areas that have not been fully investigated and do not have a test. If you find yourself in a situation where there is no test available or the available tests do not adequately meet your needs, you may have to design a conceptual multiple-choice test to fill the niche. All of these characteristics must be part of the design process and ***not*** a consequence of the process. The process that will be described in this paper is based on classical test theory. How to develop a conceptual multiple-choice test which incorporates these key characteristics will be the focus of the next section.

2. Design of a Conceptual Multiple-choice Test

Beichner suggests the flowchart[27] shown in Figure 3 as a suitable approach for constructing a conceptual multiple-choice test. Each box in the flowchart can consist of many individual steps. An overview of the process will be given first.

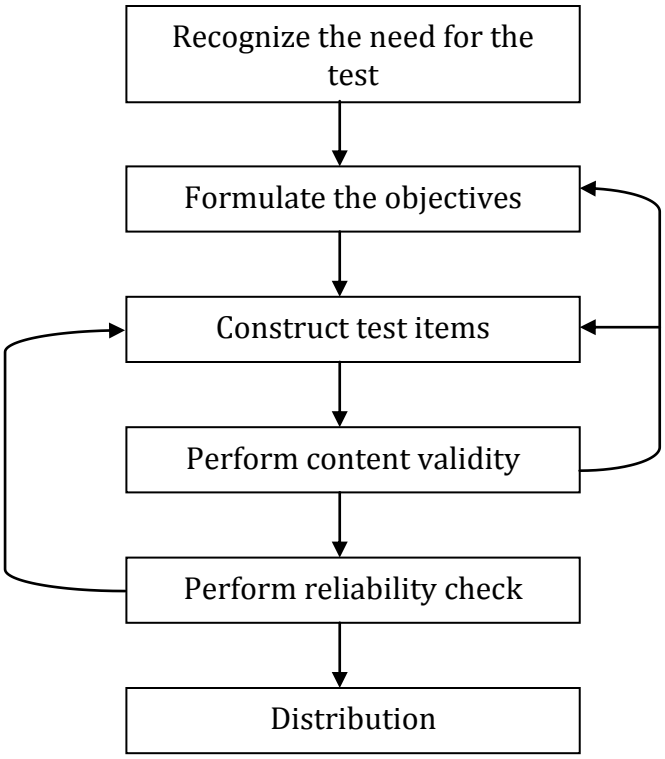

Figure 3: Flowchart for Test Development

The process begins with recognizing a need for a new test and continues with the development of a set of objectives. These objectives are used to guide test item construction. Once the test items have been constructed, a content validity check is performed. The test is presented to an independent panel of experts to evaluate how well the individual items match the objectives and to evaluate the items in general for accuracy, formatting and grammar. At this point the process could continue to the reliability check or return to re-evaluate either the objectives or test items. The reliability check is performed during a large-scale administration of the final form of the test. Problems with the reliability check result in a re-evaluation of individual test items. With a reliable version of the test, criterion and/or construct validity checks are performed. Only once this process is complete and the test has been shown to be ***both*** reliable and valid, is it ready for distribution.

2.1 Recognizing the Need

The first step in the design of a conceptual multiple-choice test is to recognize a need for a new test. As an illustration, we will consider why the author chose to develop DIRECT.[28]It was time to choose a doctoral dissertation project. At the time, the author was planning a project that would examine the effectiveness of different types of educational software programs in helping students overcome conceptual difficulties with dc resistive electric circuits. An objective, easy-to-grade test that would allow a comparison of student learning using the three software packages was needed. Other tests were already available; however, many of these were designed either as a research tool or to test a new curriculum. Those that were designed as a research tool had restricted content focusing, for example, only on current. The Electric Circuits Conceptual Evaluation (ECCE) was developed specifically to assess the effectiveness of two new curricula (*Tools for Scientific Thinking*[29]and *RealTime Physics*[30]) but was rejected due to its alignment with the curriculum and the additional coverage of inductors and capacitors. As a result of the examination of other tests, the author decided to create a new test (DIRECT), which took on a life of its own and became the entire dissertation project. As this illustration shows, the test development process is not something that can be completed in an afternoon but requires months or more to complete.

For contextual purposes to guide this section of the paper, suppose that we are planning a study to ascertain the effects of a new curriculum on students' understanding of thin lenses by using a conceptual multiple-choice test. We have examined the literature and have found no tests available, which requires us to create a new test. We need a test that covers thin lenses that is appropriate for high school and algebra- and calculus-based university courses, including two-year colleges. We will be incorporating known student misconceptions reported in the literature as distractors.

2.2 Writing the Objectives

Having identified a need in a specific area of physics, we next determine what concepts are to be included on the test. For our thin lenses test, we first need to decide what students should understand about thin lenses. A review of the topics related to thin lenses covered in typical textbooks can aid in the development of this list which might include the ray model of light, the law of refraction, image formation, the lens equation, and ray tracing. This list of topics does not indicate what the students should be able to do with those ideas. A list of instructional objectives needs to be created which articulates what students can do with those ideas. Few university instructors have been trained to write instructional objectives. There are textbooks[31,32] which discuss in detail how to write good instructional objectives. Some textbook

instructor's guides[33] are also including student learning objectives which can be helpful. Table 1 shows what a list of instructional objectives for thin lenses might look like.

Table 1: Instructional Objectives for Thin Lenses

The student can
use the ray model of light to describe how light leaves the source and interacts with the lens.
use the law of refraction to explain the bending of light within the lens.
use the three principal rays to locate where the image is formed.
calculate using the lens equation the location of the image given the focal length of the lens and the object distance.
do all of the above for both converging and diverging lenses.

Writing the instructional objectives is a crucial step in the development process. As Nitko[34] points out, a specific list of objectives aids in the planning of the test procedures or evaluation of an existing test. It also aids in the determination of the content validity of the test, making matching items to objectives easier. Once the objectives have been written, one can either proceed to writing items or begin part of the content validity check, which is discussed later in section 3.4.

Another useful tool is to create a table of specifications.[35] The table of specifications is usually illustrated in table form. The rows represent the content areas while the columns represent the cognitive levels, which are often based on Bloom's taxonomy.[36]A simple example is shown in Table 2 for a 30 item multiple-choice test. The table of specifications helps ensure appropriate coverage of the material and can assist in the evaluation of the content validity of the resulting test.

Table 2: Table of Specifications for Thin Lenses

Content	Level		
	Recall (10%)	**Interpret (60%)**	**Apply (30%)**
Ray model of light (25%)	1	4	2
Law of refraction (25%)	1	5	2
Lens equation (25%)	1	4	3
Ray tracing (25%)	1	4	2

Prior to developing individual test items, it is important to ensure that the area covered on the test is appropriate and that no major omissions have been made. The panel of experts that is formed as part of the content validity process will conduct this review. Taking this step early in the process will ensure that the items written for the test will adequately cover the domain.

2.3 Writing Items

Using the instructional objectives and/or the specifications table (if available) as a guide, individual test items are written for each objective. As items are written, a table should be created that matches the item with its corresponding objective. This table will be used during the content validity check in determining percentage agreement.

Writing a test item is not simply a matter of selecting a group of items from a commercially available test bank. The items need to written with the goal of eliciting students' ideas related to the particular objective. That is not to say that perusing the test bank and items from various curricular packages and interview questions would not be useful. These items can be adapted or used with appropriate permission and credit. This step, however, should *not* be taken lightly. This step can make or break the usefulness of the final test product. Good test items will be neither too difficult nor too easy and will be able to differentiate between students who do well overall on the test or item from those who do not. We will come back to some specific issues related to writing test items shortly.

So it is important to carefully write the individual test items, but how many do you write per objective? Is one item per objective sufficient? Popham points out that this is not an easy question to answer, but it is an important one. He recommends that the

number of items depends on whether the decisions that will be made from such a test will be high stakes (difficult to reverse) or low stakes (easy to reverse).[37] For high-stakes decisions, such as a competency test for graduation, Popham suggests 20 items per objective. For low stakes, such as alerting an instructor to the misconceptions held by the class, he recommends 5 items per objective.[38] In general, writing multiple items for each objective allows for triangulation of responses. Very few physics or astronomy conceptual multiple-choice tests have 5 items per objective. Therefore, a minimum of three items per objective is suggested so that responses can be triangulated.

Multiple-choice items involve writing both a **stem** (the question) and answer choices. The answer choices consist of the correct answer known as the **key** and the incorrect answers known as **distractors**.[39] There are many resources[40] which provide suggestions and examples for writing good items and stems. Here is a list of suggestions provided by Linn and Gronlund:[41]

- The stem of the item should be meaningful by itself and should present a definite problem.
- The item stem should include as much of the item as possible and should be free of irrelevant material.
- Use a negatively stated stem only when significant learning outcomes require it.
- All of the alternatives should be grammatically consistent with the stem of the item.
- An item should contain one correct or clearly best answer.
- Items used to measure understanding should contain some novelty, but beware of too much.
- All distracters should be plausible. The purpose of the distracter is to distract the uninformed from the correct answer.
- Verbal association between the stem and the correct answer should be avoided.
- The relative length of the alternatives should not provide a clue to the answer.
- The correct answer should appear in each of the alternative positions an approximately equal number of times but in random order.

- Use sparingly special alternatives such as "none of the above" or "all of the above."
- Avoid the use of the word "and" in alternatives.[42]
- Do not use multiple-choice items when other question types (short answer, fill in the blank, matching) are more appropriate.

In writing the distractors for each stem, a useful step is to first write the stems in an open-ended format. The set of stems are then administered to a group of students within the target population. The student answers are then categorized into groups that eventually form the distractors and key. The number of distractors then will in part be determined by the number of different responses given and by the number of possible misconceptions elicited by the item. In reducing the possible number of answer choices, you want to keep in mind that the distractors need to be plausible to the students; keeping the top four or five incorrect responses is reasonable. One does not want to have an unlimited number of options; however, as cognitive research suggests that most people can only keep about seven plus or minus two digits or unrelated words in short-term memory.[43] Using the open-ended format to help determine appropriate distractors can help identify, early on, any difficulties with specific questions and provides answer choices using common student phrasing. Additionally, this step can identify the reliability of the research that has been guiding the item creation. In other words, students' written answers may uncover new or different ideas than indicated by earlier research.[44]

To confirm the categorization of student answers, at least one other individual should be given a random subset of the answers provided by students and asked to categorize them into groups of the correct answer and distractors which will have multiple sub-categories; this is known as **inter-rater reliability**. The category labels should be agreed upon, and any disagreements should be discussed and reconciled. The percentage agreement between the two categorizations should be computed. This is simply the number of times the two individuals categorized the answer in the same way for each question. The total percentage agreement or correlation should be reported in any future publications.

It is recommended that more items are written for the test and field tested than will actually be used, so that only the most discriminating and well functioning items are chosen for inclusion on the final form of the test. Once this set of items has been written, it is time for a content validity check.

2.4 Content Validity

Validity is a measure of whether or not the test measures what it says it measures, as well as what interpretations can be inferred from the scores on the test. Linn and Gronlund state that the

> interpretations and uses of assessment results are likely to have greater validity when we have an understanding of (1) the assessment content and the specifications from which it was derived, (2) the nature of the characteristic(s) being mea-sured [*sic*], (3) the relation of the assessment results to other significant measures, and (4) the consequences of the uses and interpretations of the results. However, for many uses of a test or an assessment, it is not practical or necessary to have evidence dealing with all four considerations.[45]

For convenience, the methods for establishing the validity of a test have been divided into three categories based on the intended uses of the test scores. These categories are: content-related, criterion-related, and construct-related evidence. As the *APA Standards* point out,

> the use of category labels does not imply that there are distinct types of validity or that a specific validation strategy is best for each specific inference or test use. Rigorous distinctions between the categories are not possible. Evidence identified usually with the criterion-related or content-related categories, for example, is relevant also to the construct-related category.[46]

Evidence of content-related validity is established early in the development process while criterion-related and construct-related validity are established after the test has been administered to a group of students. These latter two types of validity will be discussed later in sections 2.5.3.1 and 2.5.3.2 and we will focus for now on content-related validity.

Evidence of content-related validity examines how well the test items cover the content domain it purports to test. For example, DIRECT[47] purports to test direct current (dc) resistive electric circuit concepts. Content-related validity then would answer the question, "Do the items on DIRECT adequately test dc resistive electric circuit concepts such as voltage, resistance, current, power and energy?" Evidence of content-related validity is important for achievement tests where the subject matter is clear.[48,49] Thus, content-related validity provides an understanding of the test content and the specifications from which it was derived.

It should be noted that evidence of content-related validity is more than just a superficial look at the test items and saying that they appear to measure the content. When the items of a test appear at quick glance to measure a particular concept, this is known as *face validity*. Although it is important for a test to have face validity so that students are motivated to perform well on the test, it is *not* the same as establishing the content validity of the test through more rigorous methods.[50]

To begin the content validity check, an independent panel of experts is formed. This panel of experts consists of individuals who are knowledgeable about the area being tested. These individuals may have developed curricular materials within that area, have done research into students' ideas about this area, or are well-known for their teaching in this area.[†] The literature does not give a set number of individuals who should serve on the panel of experts as this will depend on the amount of work already done in that particular area of physics as well as the target population. If the test will be used in high school classes, high school teachers familiar with the curriculum and standards should be included on the panel. Five individuals would provide some consensus, although it would be prudent to invite more than five individuals as some will decline to participate due to other commitments.

The literature is unclear if compensation should be provided to the panel of experts for this time and expertise. The author has not provided compensation other than giving appropriate credit in publications for the contributions provided by the panel of experts. Providing compensation may be detrimental to the process in that the panel may feel obligated to provide positive feedback. Additionally, funds need to be available and not all test development projects are funded.

To establish the content-related validity of a test, the test, instructional objectives, and/or table of specifications are given to the independent panel of experts. The panel of experts reviews the instructional objectives and/or table of specifications to determine if the content coverage is adequate. The review of instructional objectives can be done before the individual test items have been written to help guide the item writing phase or after the items have been written. The test and instructional objectives are matched by the panel of experts. The test is also examined to ensure that the answer key and items are free from errors. Typically, the evidence for content-related validity is reported as a percentage agreement of items-objectives.

[†] For our example of thin lenses, the following PER researchers would be appropriate to ask: Fred Goldberg would be appropriate for this panel as he has done research into student understanding of geometrical optics.[51] Dewey Dykstra was part of a group who created a set of curricular materials[52] related to light. Eric Mazur's book, *Peer Instruction*[53] contains a set of ConcepTests for light.

Alternatively, one could use a correlation to indicate relative agreement. In either case, the results should show high levels of agreement, at or above 90%.

The content validity assesses whether the items adequately sample the domain. For our thin lenses test, do the items adequately cover the concepts related to thin lenses? The multiple-choice version of the test along with the objectives and/or specifications table is given to the panel of experts. The panel of experts will take the test and match each test item with its corresponding objective. The panel also notes any grammatical errors or problems with the wording of the items that might either confuse the student or provide a clue to the correct answer. The percentage agreement on matching items with objectives is reported.

A way to assess whether a single item has content validity has been proposed by Hambleton and Rovinelli.[54] They assume that in an ideal case each item will only match one objective. When collecting data from the panel of experts, the panel is instructed to indicate the degree to which an item matches that objective. A value of 1 indicates that the item does measure that objective, 0 if unsure, and -1 if the item clearly does not measure that objective. An item-objective congruence of 1.00 indicates the item is matched to one and only one objective by all judges. If a single item can be matched to more than one objective, the index will be less than 1.00. The formula for calculating the item-objective congruence for item i to objective k is given by

$$I_{ik} = \frac{N}{2N-2}\left(\mu_k - \mu\right)$$

Where N is the number of objectives, μ_k is the judges' mean rating of item i on the k^{th} objective and μ is the judges' mean rating of item i on all objectives.

After the content validity check has been completed, it may become necessary to edit and/or revise the individual test items. Major revisions would require the panel of experts to re-examine the content validity of the test. Minor revisions would not. This process can take several iterations before a final version of the test is ready for field testing.

2.5 Field Testing

Once the final version of the test is completed and the content validity check has been done, the test is now ready for field testing. Field testing is part of the reliability check shown in Figure 3 and consists of many steps. The test developer now solicits

test sites to administer the test so that reliability, validity, and item analysis can be evaluated.

Two common approaches used in the past to solicit test sites in physics were 1) to post a request to a listserv such as PHYS-LRNR[55] or 2) to present a poster about the test at national meetings of the American Association of Physics Teachers.[56] Newer venues such as ComPADRE[57] and PER-Central[58]are now available and may take on a greater role in the solicitation process. The number and type of institutions to be solicited will depend on the target population for which the test is intended. For our thin lenses test, we wish to solicit university test sites that offer algebra-based and calculus-based introductory physics courses, as well as high school test sites. The university sites should include two-year colleges as well as four-year institutions.

A representative sample may include thousands of students for reliability studies, but only a few students in each course (high school, university algebra-based and calculus-based) may be required for the validity study. When conducting the item analysis of the test, Crocker and Alginanote that "another longstanding rule-of-thumb is to have 5 to 10 times as many students as items."[59] For our 30-item thin lenses test, we would need between 150 and 300 students for the item analysis. Each of these steps will be discussed next.

2.5.1 General Test Statistics

With any test that is given, it is common to report the following information to describe the variance in the scores obtained by the test. In general, the following information is reported:

- Mean
- Standard deviation
- Standard error of mean
- Range
- Distribution of scores

2.5.1.1 Mean

The mean is the average of all of the scores on the test. It can be calculated by

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

where x_i are the individual scores and N is the total number of students.

2.5.1.2 Variance and Standard Deviation

The deviation score is how each individual score differs from the mean ($x_i - \bar{x}$). The variance in scores is determined by summing the square of the deviation scores and dividing by the number of students. The variance (σ^2) is given mathematically by

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

The variance is always positive and describes "the dispersion of a set of scores around the mean of the distribution."[60]

The standard deviation (σ) is just the square root of the variance and also gives a measure of spread of the scores in a distribution.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$$

The larger the variance or standard deviation, the more the scores on the test are spread out. The smaller the variance or standard deviation, the more the scores on the test are close together. For many of the measures we will consider, we will want the variance to be as large as possible so that the scores on the test are spread out and not clumped tightly together.

2.5.1.3 Standard Error of the Mean

The standard error of the mean gives an estimate of how close the obtained mean is to the true mean. It depends on the standard deviation (σ) and the number of students taking the test (N). The standard error of the mean is given by

$$\sigma_{mean} = \frac{\sqrt{\sigma}}{N}$$

The smaller the value of the standard error of the mean is, the smaller the uncertainty in the value of the mean. Ideally, the standard error of the mean should be as close to zero as possible.

2.5.1.4 Range and Distribution of Scores

The range is the difference between the highest and lowest scores plus one. Ideally, the range should be as large as possible so that the variance will also be large. To help visualize the distribution of scores, it is necessary to plot the portion of students obtaining each possible score. Figure 4 plots the possible scores on DIRECT version 1.0 along the horizontal axis and the total number of students obtaining that score on the vertical axis.

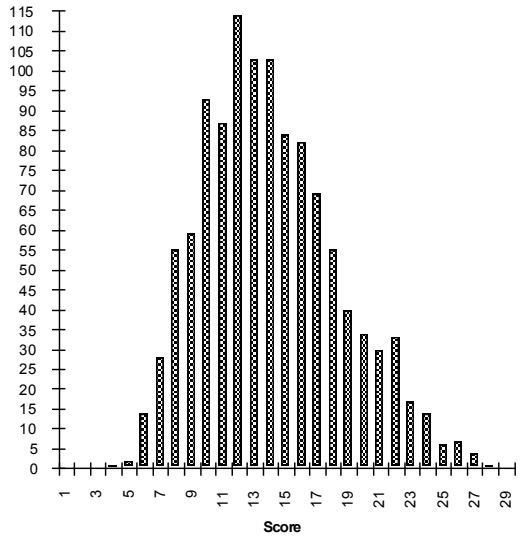

Figure 4: Distribution of Scores for DIRECT Version 1.0 for the Overall Sample[61]
The overall mean for version 1.0 of DIRECT was 14 out of 29 or approximately 48%. Figure 4 shows that the distribution is not symmetric but tails off to the higher end of the scores; thus, version 1.0 of DIRECT has a slight positive skew, indicating a difficult test. A negatively skewed distribution would tail off toward the lower end and indicate an easy test.

2.5.2 Reliability

An important attribute of a high-quality test is the reliability. Anastasi defines reliability as

> the consistency of scores obtained by the same persons when reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions. In its broadest sense, test reliability indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to chance errors.[62]

In testing we want to be able to generalize "what we see today, under one set of conditions, to other occasions and conditions."[63] There are three types of reliability: stability, equivalency, and the internal consistency of the individual items. Stability refers to how well the scores remain constant over time. A student taking the test on Monday should have a very similar score on Tuesday, barring external factors that might affect performance, such as illness. Equivalency refers to how well the scores on two different versions that cover the same content relate to one another. A student taking form A should expect to receive an equivalent score as another student who is taking form B for the same level of performance. Internal consistency is an indication of how homogeneous the test items are. The more homogeneous the test items, the more likely the test items measure the same concept, such as force in the case of the FCI. Internal consistency is especially important with tests used for assessment.[64]

The choice of which reliability one needs to calculate depends in part on the number of forms of the test one has and on the number of testing sessions. As you can see in Table 3, in order to determine the internal consistency associated with the test, only one test form and test session are required. This is why in practice the internal consistency of the test is the most often calculated. The methods for conducting a reliability study for each type of reliability will be discussed.

Table 3: Reliability in Relation to Test Form and Test Session

Number of Testing Sessions	**Number of Test Forms**	
	One	**Two**
One	Internal consistency	Equivalency
Two	Stability	Stability and Equivalency

2.5.2.1 Test-Retest

One method for computing the stability of a test is to use the test-retest method. In this method, one form of the test is given on two separate occasions under similar testing conditions. The separation between the two testing sessions is typically two weeks but can be anywhere from one day to several months. The students' scores from the two administrations are then correlated using the Pearson product-moment correlation coefficient, which is given by

$$r_{xy} = \frac{\sum_{i=1}^{N} x_i y_i}{(N)(\sigma_x)(\sigma_y)}$$

x_i is a student's deviation score on the first administration of the test, y_i is the deviation score for the same individual on the second administration of the test, N is the number of students in the sample, σ_x is the standard deviation of the scores on the first administration and σ_y is the standard deviation of the scores on the second administration. The Pearson Product-Moment Correlation coefficient (hereafter called the correlation coefficient) varies between -1 and 1; the more positive the correlation the higher the reliability.

Recall that reliability provides an estimate of the amount of error in the scores. The test-retest reliability coefficient indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to errors due to the separation in time. To estimate the amount of error variance due to the difference in time, subtract the correlation coefficient from one. The resulting value gives you a measure of the error variance due to time sampling. For example, suppose the correlation coefficient between two administrations of the same test separated by a month was 0.70. The amount of error attributable to time sampling then is 30%. The remaining 70% is attributable to true differences in the test scores. Evaluating the amount of error due to the difference in time is important when the test will be used for pre-post analysis.

2.5.2.2 Alternate Forms

The experience of having taken the test once already plays a role in the students' performance on the second administration. An issue with longer gaps between administrations is that there is a possibility of a developmental spurt in some individuals or additional learning could have taken place. As a result, the test-retest reliability coefficient tends to give an overestimate of the reliability. One way to correct this problem is to have two forms of the test. The questions on each form are

not the same, but the content is. This eliminates the issue of students becoming sensitized to the test items.

When using two forms of the test and one testing session, one can assess the equivalency of the two forms. Alternatively, when using two forms of the test and two testing sessions, one can assess the stability as well as the equivalency of the two forms. In either case, the method is similar in design to the test-retest with the difference being the amount of time between the two administrations.

Table 4: Test Administration and Test Form for Alternative Forms Methodologies

	Administration 1	**Administration 2**
Group 1	Form A	Form B
Group 2	Form B	Form A

As Table 4 shows, both groups take both forms of the test. When using only one testing session, there is only a short break between the two administrations to prevent test fatigue. When using two testing sessions, the separation is usually two weeks up to six months. The correlation coefficient is used to calculate the reliability coefficient.

The equivalency reliability coefficient indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to errors due to the content. To estimate the amount of error variance due to the content, subtract the correlation coefficient obtained using 2 forms and 1 session from one. The resulting value gives you a measure of the error variance due to content sampling. An estimate of the error variance due to content sampling plus time sampling is calculated by subtracting the correlation coefficient obtained from 2 forms and 2 sessions from one.

2.5.2.3 Internal Consistency

The internal consistency is measured by administering a single form of the test once to a group of students. The methods for determining the internal consistency of the test compare the correlations between separately scored parts of the test. There are three methods for evaluating the internal consistency of the test: 1) Split-halves, 2) Coefficient Alpha, and 3) Kuder-Richardson 20 (KR-20). As Crocker and Algina point out, coefficient alpha and KR-20 should yield identical results.[65] One disadvantage of using internal consistency methods alone for assessing the reliability

of a test is that these methods do not account for time sampling and only account for content sampling.

2.5.2.3.1 Split-Halves

Once the test has been administered to the group of students, the test items are divided in halves. One way is to simply divide the test by odds and evens – all odd numbered items form subtest 1, while all even numbered items form subtest 2. Another way of dividing the test into halves is randomly. A third way is to attempt to match items based on content. As you can see there are a number of ways to create this division.[66] Each half is now scored for each student, and the correlation coefficient is calculated between the two subtests. Since this procedure has now reduced the number of items on our test in half, the Spearman-Brown prophecy formula is used to determine what the reliability coefficient would be for the whole test. The Spearman-Brown prophecy formula is given by

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}}$$

where r_{tt} is the reliability for the whole test and r_{hh} is the correlation coefficient between the two halves.

There are some drawbacks to using the split-halves method to determine the internal consistency of a test. Since there are a number of ways to divide the test, the reliability coefficient that the procedure yields is not unique. In addition, the test must be able to be divided into two equal halves. Another issue is that the Spearman-Brown formula assumes that the variances between the two halves are equal,[67] which may not always be a valid assumption.

2.5.2.3.2 Coefficient Alpha and Kuder-Richardson-20

A method that provides a unique result for the internal consistency of a test is needed. Two of the methods developed to address this issue are the coefficient alpha and the KR-20. Both of these methods look at the covariances of the items.

The coefficient alpha is the more general of the two and can be used with non-dichotomously, as well as dichotomously, scored items. Dichotomously scored items are scored as either right or wrong. The coefficient alpha can be calculated using the following formula

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_t^2}\right)$$

where k is the number of test items, σ_t^2 is the total test variance and σ_i^2 is the variance for item i.

The Kuder-Richardson 20 (KR-20) was developed to handle only dichotomously scored items. The formula looks very similar to the coefficient alpha and is given by

$$r_{tt} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{\sigma_t^2}\right)$$

where k is the number of test items, σ_t^2 is the total test variance and p_i is the proportion of students that answer item i correctly.

Most multiple-choice tests are dichotomously scored which is why the KR-20 tends to be the most widely used in the physics education literature. Both the coefficient alpha and the KR-20 yield similar results, but both methods yield reliability coefficients that are lower than that produced by the split-halves method.

2.5.2.4 Factors Affecting Reliability

There are four main factors that can affect the reliability of a test. These are:

1. Length of the test
2. Discrimination of items
3. Difficulty of items
4. Range of ability of group

The procedures in determining the reliability of the test depend on the variance of the test itself as well as the individual items; the larger the variance the higher the reliability. Thus, the more items on a test that can discriminate well increase the variance and thus the reliability. As we will see, the optimum average difficulty level

[see section 3.5.4.1] that will produce the maximum discrimination is 50%. Difficulty values above or below the 50% level will affect the discrimination and, in turn, the reliability of the test. Finally, the ability range of the students will also affect the reliability. If the group is fairly homogeneous in ability, the range of scores and corresponding variances will be small, which will lower the reliability. Optimally one would like to develop a test that can spread the scores as much as possible to gain the largest variance and thus largest reliability.

2.5.2.5 Acceptable Values of Reliability

Having discussed what reliability is and how to calculate it, we now need to ask the question of what values of reliability are appropriate. The answer of course will depend on the purpose of the test. Table 5 shows one guideline for determining the appropriateness of the reliability coefficient.[68]

Table 5: Guideline for Reliability Coefficients

0.95 – 0.99	Very high, rarely found
0.90 – 0.95	High, sufficient for measurement of individuals
0.80 – 0.90	Fairly high, possible for measurement of individuals
0.70 – 0.80	Okay, sufficient for group measurements, not individuals
Below 0.70	Low, useful only for group averages and surveys

Since the purpose of most conceptual multiple-choice tests is to identify areas of difficulty and evaluate teaching, values of 0.70 or above are acceptable.

2.5.2.6 Standard Error of Measurement

Not only do we want to be able to assess how well the test scores are consistent but we would like to estimate how close the obtained score is to the true score. To do this, we can calculate the standard error of measurement. This value gives an estimate of the standard deviation produced for an individual taking a large number of random parallel forms of the test.[69] The standard error of measurement (SEM) depends on the reliability of the test (r_{tt}) and the standard deviation of the test scores (σ_t).

$$SEM = \sigma_t \sqrt{1 - r_{tt}}$$

When the SEM is large, the uncertainty in the individual's true score is also large. When the SEM is small, the uncertainty is small, which means the more certain we

are of the score.[70] It is common to talk about a confidence interval for the true score. For instance, we can be 68% confident that the true score lies within one SEM. Other confidence levels are shown in Table 6.[71] In determining the confidence intervals, the distribution of scores for a single individual is assumed to be normal.

Table 6: Confidence Levels for True Score

Confidence level	±SEM
68%	1.0
90%	1.64
95%	1.96

2.5.2.7 Example Reliability Study

The focus of this section will be on applying these ideas to our thin lenses test. As a reminder, we are designing a conceptual multiple-choice test to ascertain the effects of a new curriculum on students' understanding of thin lenses. Our thin lenses test has been designed to be appropriate for high school and algebra- and calculus-based university courses including two-year colleges. We have incorporated known misconceptions reported in the literature as distractors. We have developed a single version of the test which will be dichotomously graded.

Since we have developed a single thin lenses test, we will want to consider both time sampling and content sampling as part of our reliability study. We will have our test sites administer the test the week before and week after students study thin lenses. For most classes, the time delay will be approximately 2 weeks. We will calculate the correlation between the pre- and post-administration of the test as well as the KR-20. We choose the KR-20, as the test will be graded as right or wrong. We will at a minimum need to calculate the KR-20 for each group of students (HS, algebra-based, and calculus-based) as these groups are not equivalent, as well as an overall value for the test combining all three groups. We will also report the standard error of measurement.

2.5.3 Validity Studies

As part of the field testing procedure, the validity of the test is established. Recall that validity relates to the inferences that can be made about the scores that are obtained. Previously, we discussed the content-related validity of the test which is evaluated early in the test development process. Criterion-related and construct-related validity

are evaluated during the field testing portion of the process after the test has been shown to be reliable.

2.5.3.1 Criterion-related Validity

Criterion-related validity evidence provides an understanding of the relation of the test results to other significant measures. There are two types of criterion-related validity: 1) predictive validity and 2) concurrent validity. Predictive validity evidence indicates how well an individual's performance on the criterion, such as final course grade, compares to their performance on the test under validation. Predictive validity answers questions such as "Can the FCI be used to predict final course grades?" Concurrent validity evidence indicates how well an individual's performance on the criterion at this time and their performance on the test under validation compare. Concurrent validity answers questions such as "Can the observational results from the Reformed Teaching Observation Protocol (RTOP)[72] be replaced by a self-report teacher survey?" The main difference between these two is the time at which the criterion is evaluated. For assessing predictive validity the test is given and then later compared to the criterion. For assessing concurrent validity the criterion is evaluated at the same time as the test is given. Criterion-related validity is often reported as the correlation between the criterion and the test; the higher the correlation, the more valid the inference.

2.5.3.2 Construct-related Evidence

Construct-related validity is associated with an understanding of the nature of the characteristic(s) being measured and the consequences of the uses and interpretations of the results. Linn and Gronlund provide an excellent description of construct. They state

> Whenever we wish to interpret test results in terms of some individual char-acteristic [*sic*] (e.g., reading comprehension, mathematics problem-solving ability), we are con-cerned [*sic*] with a **construct** [*sic*]. A construct is an individual characteristic that we assume exists in order to explain some aspect of behavior. Mathematical reasoning is a construct and so are reading comprehension, understanding of the principles of electricity, intelligence, cre-ativity [*sic*], and such personality characteristics as sociability, honesty and anxiety. These are called constructs because they are theoretical constructions that are used to explain per-formance [*sic*] on an assessment. When we interpret assessment results as a measure of a particular construct, we are implying that there is such a construct, that it differs from other

> constructs, and that the results provide a measure of the construct that is little influenced by extraneous factors. Verifying such implications is the task of construct validation.[73]

There are a number of different ways to collect evidence of construct-validity. Three of the methods are: 1) intervention studies, 2) differential population studies, and 3) related measures studies. Intervention studies attempt "to demonstrate that examinees respond differently to the measure after receiving some sort of treatment."[74] One would expect that scores on the FCI would increase after students have been given instruction on forces. Differential population studies attempt to demonstrate that different populations of students score differently on the test. One would expect that English majors would score differently than physics majors on the FCI. Related measures studies look at both the positive (convergent validity) and negative (divergent validity) correlations between different measures. One would expect a positive correlation between the FCI and Force and Motion Conceptual Evaluation (FMCE), as they purport to measure similar constructs, force and motion. However, one would expect a negative or zero correlation between the FCI and DIRECT, as they purport to measure different constructs, force versus electricity.

Another method of collecting evidence for the construct validity of a test is to perform a factor analysis. A factor analysis is a "refined statistical technique for analyzing the interrelationships of behavior data."[75] A factor analysis simplifies "the description of data by reducing the number of necessary variables, or dimensions."[76] There are statistical packages available on-line that can perform factor analysis of data.[77] Even with the aid of such statistical packages, factor analysis is *not* a trivial technique.

2.5.3.3 Example Construct Validity Study

The construct validity study for our thin lenses test will focus on the construct validity of the test. Does our thin lenses test actually measure the concepts associated with thin lenses or are we measuring something else? To help determine the answer to this question, we will conduct individual interviews with a small sample of students in each group (HS and university-level algebra-based and calculus-based) to determine the reasoning behind their answer selections and to ensure that the questions were being understood as intended. We will also conduct a factor analysis to see what factors account for the variation observed on the test. We will also examine the pre-post test results, as we would anticipate that the students' scores will increase after instruction on thin lenses. We will also examine the results to see if there is any gender bias associated with the test as a whole, as well as with individual test items.

2.5.3.4 Factors Affecting Validity

There are two types of errors that can affect the validity of a test. These are unsystematic and systematic. Unsystematic errors result from the unreliability of the test. Systematic errors are more numerous and include method of measurement, enabling behaviors, differential item effectiveness, administration errors, and comparative sample.[78] The method of measurement could include paper-and-pencil versus on-line. Enabling behaviors such as reading ability could negatively affect performance on any test if some of the students were non-English speaking students.[79]All items on the test should work in the same way for all test-takers. For example, McCullough has reported that at many institutions women perform more poorly on the FCI than men.[80] This difference in performance based on gender would be an example of differential item effectiveness. Administering the test in ways other than that indicated by the documentation would result in invalid inferences. Examples of administrative errors and comparative sample are giving only 10 items from DIRECT instead of all 29 or giving the test to students not in the comparative sample.

Some of these issues are rectified by conducting local reliability and validity studies to show that utilizing 10 items is equally reliable and valid as using all 29 items. Additionally, one could conduct a new comparative study to provide data for new student groups not included in the original comparative sample.

Alternatively as part of the validity process, one could examine the cultural validity of the test. Solano-Flores and Nelson-Barber define cultural validity as "the effectiveness with which science assessment addresses the sociocultural influences that shape student thinking and the ways in which students make sense of science items and respond to them." As the United States continues to become more culturally diverse, this type of validation will become increasingly more important.

2.5.4 Item Analysis

The purpose of the item analysis is to make certain that the items on the test are functioning well. If they are not, then the questions may require editing or reworking in order to function better. This section will discuss each of these calculations.

- Difficulty
- Discrimination
- Point-biserial correlation
- Examination of distractors

2.5.4.1 Difficulty

The difficulty (p_i or p-value) of an item is given by the number of students answering the item correctly divided by the total number of students. The p-value varies between 0 and 1. A value close to 1 indicates that most people answered the item correctly while a value close to 0 indicates that most people did not answer the item correctly. Ideally, items on the test should have an average difficulty of 0.5 so that the discrimination is maximized.[81] Mathematically, the difficulty is given by

$$p_i = \frac{\text{number answering item } i \text{ correctly}}{\text{total number taking test}}$$

The difficulty of an item can be affected by guessing. If many students are able to correctly guess the answer, the difficulty value will be arbitrarily high.

2.5.4.2 Discrimination

Discrimination examines how well the test as a whole or individual test items in specific distinguish between individual students. The Ferguson's delta is used to evaluate how well the test as a whole discriminates. The discrimination index and point-biserial correlation evaluate how well individual test items discriminate.

2.5.4.2.1 Ferguson's Delta

The Ferguson's delta is used to determine the discrimination power of the test as a whole. It does this by evaluating "how broadly the total scores of a sample are distributed over the possible range."[82] The authors of reference 82 provide a more thorough discussion of the Ferguson's delta formula. The Ferguson's delta varies between 0 and a maximum of 1 and assumes a rectangular distribution of scores.[83] An acceptable value is greater than 0.90. Ferguson's delta is given by

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - \left(\frac{N^2}{K+1} \right)}$$

where N is the number of students in the sample, K is the number of test items, and f_i is the frequency of cases at each score.

2.5.4.2.2 Discrimination Index

The discrimination index determines the discrimination power of individual test items. The discrimination index applies only to dichotomously scored items, those scored as right or wrong. In order to calculate the discrimination index for each item, it is necessary to divide the sample based on total test score. Two common approaches to dividing the sample are to divide the sample in half or to divide the sample into the upper 27% and the lower 27%.

When dividing the sample in half, those students who scored above the mean are called high (H) while those who scored below the mean are called low (L). Doran points out that if there are a number of students who score at the mean they can either be eliminated from the calculation or assigned randomly to the high and low groups until both groups receive the same number.[84] The discrimination index is then computed using the following formula

$$D = \frac{H - L}{N / 2}$$

where H is the number in the high group and L is the number in the low group who answered the item correctly, and N is the total number in the sample.

When dividing the sample using the upper and lower 27%, the formula is given by

$$D = U - L$$

where U is the proportion of students in the upper group who answered the question correctly and L is the proportion of students in the lower group who answered the question correctly.

The discrimination index varies between -1 and 1. A negative discrimination index indicates that more students in the low group answered the question correctly while a positive discrimination index indicates that more students in the high group answered the question correctly. Typically, a discrimination index above 0.30 is considered acceptable.[85] The discrimination index is highly affected by the difficulty of the item.

2.5.4.2.3 Point-biserial Correlation

The discrimination index provides an estimate of how well an individual test item differentiates between those students scoring well on the test from those who do not. A related idea is the point-biserial correlation. The point-biserial correlation measures the correlation between the item's correctness and the whole test score.[86] The point-biserial correlation varies between -1 and 1, just as the discrimination index does. A

high positive value means that students who scored well on the test overall have a higher probability of answering the item correctly while a high negative value means that students who scored well on the test overall have a lower probability of answering the item correctly. Since the objective of any test is to have all items highly correlated with the total test score, the values for the point-biserial should be greater than 0.20.[87] Mathematically the point-biserial correlation (r_{pbs}) is given by

$$r_{pbs} = \left(\frac{\bar{x}_{correctly} - \bar{x}_{whole\ test}}{\sigma_{whole\ test}} \right) \sqrt{\frac{p_i}{1 - p_i}}$$

where $\bar{x}_{correctly}$ is the average total score for those students who answered item i correctly, $\bar{x}_{whole\ test}$ is the average total score for the whole sample, $\sigma_{whole\ test}$ is the standard deviation of the total score for the whole sample, and p_i is the difficulty index for item i.[88]

2.5.4.3 Examining Distractors

Since the distractors of the test should be designed to attract students who are not as knowledgeable, it is important to examine how the distractors of the test function. If a particular alternative had no students choosing it, it may require editing or replacing. If a particular item had a low point-biserial correlation, it might be that the distractors are too attractive and are also foiling those students who do understand the material being tested. This would require rewording of the distractors or interviews to find out why students who performed well overall on the test were drawn to that particular distractor instead of the correct answer.

2.5.4.4 Summary

The item analysis is performed to determine how each individual test item is functioning and how it contributes to the total test score. To summarize, the average difficulty of all items should be approximately 0.50 to ensure maximum discrimination. The items on the test should have an average discrimination index of 0.30 and an average point-biserial correlation of at least 0.20. Items that do not meet these minimum values should be examined more closely and revised and retested with a small sample of students to ascertain if the revisions were effective. If extensive revisions become necessary, then the item analysis and the reliability and validity studies will need to be performed again using a new sample of students. If

revisions are minor, then the test is ready to be standardized by collecting comparative samples.

2.5.4.5 Comparative Samples

Two other important factors that affect the quality of a test are the comparative samples that are established, as well as how well the test is tailored to the intended population. These factors have been discussed previously in relation to the determination of both the reliability and validity of the test. A test that has been designed for use with high school students would not necessarily be appropriate for use with university students; new comparative samples as well as reliability and validity studies would need to be conducted before use with a new population.

Once the reliability and validity studies and item analysis have been completed, comparative samples are ready to be collected. In collecting comparative samples, thousands of students are given the test. These students are in the target population. Often, it can be difficult in physics to get enough test sites that the data collected for the reliability and item analysis are pooled together to form the comparative sample.

Comparative samples should be representative of the population from which the sample comes. This can sometimes be difficult to achieve, as gaining enough test sites at the university level or high school level can be difficult. The comparative samples should also be up-to-date. Salvia and Ysseldyke suggests the following guidelines for comparative samples; the maximum lifetime for comparative samples of ability tests is 15 years while that of achievement tests is 7 years.[89] Most physics and astronomy tests would be categorized as achievement tests so the lifetime of a comparative sample would be 7 years.

2.6 Distribution

Once the final version of the test has been shown to be reliable and valid, it is ready to be distributed. In physics, we typically provide raw scores or percentages instead of *z*-scores* or percentile ranks. Sometimes the comparative sample is the same as the sample used to perform the item analysis and reliability and validity studies. Other times the comparative sample is collected over the course of several semesters and pooled together. Typically information about the test is disseminated through journal

* "A *z*-score or **standard score** is a dimensionless quantity determined by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation." [90]

articles describing the different aspects of the test or is the bulk of a doctoral dissertation.

Information about the results from the item analysis along with the percentage of students choosing each alternative for a given item are presented in table format for the overall sample and for each sub-sample. This information provides the test user with detailed information about all aspects of the test items. The development of the test, its intended uses, reliability studies, validity studies, item analysis, comparative sample, and administration instructions are all described.

2.7 Test Security

Hopefully, you have seen that the development of a conceptual multiple-choice test is very time intensive. Although test developers want their test to be used by others, it should be no wonder now why test developers are so concerned about the security of their work. Most test developers password protect their electronic versions of the test that they make available on the internet and ask for students not to be given copies of the test nor the answers. Once the answers are made available to one group of students, the answers quickly spread to other students and the test becomes useless. Since many of the physics and astronomy tests have been written up in journals such as the *American Journal of Physics*, it is common for the test developer to ask that a new first page of the test be used which does not utilize the actual name of the test so that students cannot surf the web to find the answers. When using a test already developed, please respect the amount of effort that went into its development and help ensure the test's integrity and security so that it can be remain useable for years to come.

3. Evaluating a Test

It is not always necessary to create your own conceptual multiple-choice test. Sometimes you need to be able to evaluate a number of tests to ascertain which test would be most appropriate for the population and purposes you need. Suppose you wanted to conduct a research study to determine if using a new textbook positively affects students' performance in the second semester of an introductory calculus-based university physics sequence. Part of your study design will use a conceptual multiple-choice test which covers electricity and magnetism. You must first select a test from those that are available. A search of the literature and internet yields the list shown in Table 7 along with the reliability, validity, discrimination, and target population information available for each test.

Table 7: List of Tests Available for Electricity and Magnetism

Test	Reliability coefficient	Evidence of validity	Discrimination	Target population
Conceptual Survey of Electricity and Magnetism (CSEM)[91]	0.75 (KR-20)	Content, Factor analysis	0.1 – 0.55	Algebra-based and Calculus-based
Diagnostic Exam of Electricity and Magnetism (DEEM)[92]	0.74 (coefficient Alpha)	Content	0.32 average	Calculus-based
Brief Electricity and Magnetism Test (BEMA)[93]	0.85 (KR-21*)	Not available	0.34	Calculus-based

*The KR-21 is an alternative form of the KR-20 which assumes all items have equal difficulty and does not require calculating the individual item variances.

Two other tests are also available: the Survey of Electricity, Magnetism, Circuits and Optics (SEMCO) and the Electromagnetics Concept Inventory (EMCI) but were rejected outright. SEMCO covers additional material (circuits and optics) and using only the questions related to electricity and magnetism would be inappropriate, as the reliability and validity of doing so has not been established. The EMCI is not acceptable, as it was designed for upper division electricity and magnetism.

Having reduced our list due to incompatibility between content and population, we need to turn our attention to the reliability, validity and discrimination. Examining each of the three tests for their reliability shows that all three tests have reliability values above 0.70 which is acceptable for group measurements. Since we want to determine how the class as a whole responds to the new text, these values are acceptable for our purposes. The CSEM and DEEM have evidence of content validity while the CSEM has also examined the construct validity via a factor analysis. All three tests seem to discriminate well, having an average discrimination over 0.30. At this point the three tests are fairly even in comparison.

All three tests have been designed and field tested with the population that we intend to test. A closer examination of the comparative samples may provide additional information which can help determine which test to use. The DEEM was administered to students at a single institution. BEMA was administered to

populations at two different institutions. The CSEM has been administered to a number of institutions including two-year and four-year universities. Because the CSEM has been administered to a more diverse population the results obtained by it are more likely to match our target population than the more selectively administered BEMA and DEEM.

In this case, there was more than one test available for consideration. Based on an examination of the reliability, validity, discrimination, comparative sample, and target population, the CSEM appears to best match our proposed study.

4. Item Response Theory

As one would expect, STEM education researchers have become more sophisticated in the design of conceptual multiple-choice tests since the FCI. More attention is now paid to the issues of reliability and validity. Just as STEM education researchers have become more sophisticated, so has the field of psychometrics. Item response theory (IRT) has yet to make significant inroads into the lexicon or tool box of most STEM education researchers.

Item response theory came about to rectify some of the shortcomings of classical test theory. In classical test theory, a student's ability is determined by the score on a particular test. The difficulty and discrimination of a particular item as well as the reliability and validity of the test are determined by the ability of a group of students; if the characteristics of the group changes so do these factors. This makes it difficult then to compare students across different tests and to compare items across different student groups. Additionally if the students who take a particular test are of different ability, then their scores will have different amounts of error, which is contrary to the assumption behind the standard error of measurement that it is the same for all individuals. In developing a test, it would be preferable to be able to predict how different groups or individuals will perform on a given item. In classical test theory the emphasis is on the test, while in IRT the emphasis is on the items.[94]

The underlying assumption of IRT is that the responses on a particular test are accounted for by a small number of latent traits. As Crocker and Algina write

> At the "heart" of the theory is a mathematical model of how examinees at different ability levels for the trait should respond to an item. This knowledge allows one to compare the performance of examinees who have taken different tests. It also permits one to apply the results of an

> item analysis to groups with different ability levels than the group used for the item analysis.[95]

IRT is useful in building tests, identifying potentially biased test items, equating scores from different tests or forms of the same test, developing tests which can discriminate at a particular level of ability, and in the development of "tailored testing" systems. The details for conducting analyses using item response theory will be left for future installments of ***Getting Started in PER***.

5. Conclusions

The purpose of this paper was to introduce the reader to the characteristics of a high-quality test. Five characteristics were discussed: reliability, validity, discrimination, comparative sample, and tailoring the test to the population with which it will be used. Reliability depends on the variance of the test scores, which in turn depends on the discrimination ability of individual test items. An item's ability to discriminate depends on how difficult the item is. The validity of the test can depend on how well the domain is covered on the test, which is dependent on the instructional objectives. The instructional objectives can be evaluated early in the test development process by a panel of experts, thus ensuring the content validity of the test. The construct validity is affected by how well the individual items on the test function and can be assessed after an administration of the test by asking a subset of the population to explain their reasoning behind their answer choices to individual questions. The point-biserial correlation values can be used to select the individual questions to be used during individual interviews. An item with a low point-biserial correlation indicates that students scoring well on the test overall are not scoring well on this particular item. As you can see, these five characteristics of high-quality tests are intertwined. A well-developed test will build these qualities into its development methods.

Multiple-choice tests are valuable in that they are objectively graded and can be given to many students in a relatively short amount of time. Well-developed tests can be designed to provide diagnostic information to the student and instructor regarding common misconceptions or other areas of difficulty. With this information in hand, instructors can make informed decisions about the course of action that is needed in their particular setting. This information can also be used to evaluate new endeavors to help students overcome these misconceptions through the use of new teaching methods or curriculum. As with evidence of validity, multiple-choice tests do *not* provide a definitive answer but provide evidence that when combined with other methods can provide a clearer picture of the status of student understanding in the classroom.

Acknowledgements

The author would like to thank Steve Robinson, Tom Foster, Kathy Harper, Charles Henderson and the anonymous reviewers for their insightful feedback which has greatly improved and focused this manuscript.

---

[1] H. T. Hudson and C. K. Hudson, "Suggestions on the Construction of Multiple-choice Tests," Am. J. Physics **49**, 838 (1981).
[2] G. J. Aubrecht and J. D. Aubrecht, "Constructing Objective Tests," Am. J. Physics **5**, 613 (1983).
[3] D. F. Treagust, "Development and Use of Diagnostic Tests to Evaluate Students' Misconceptions in Science," Int. J. Sci. Educ. **10**, 159 (1988).
[4] R. J. Beichner, "Testing Student Interpretation of Kinematics Graphs," Am. J. Physics **62**, 750 (1994).
[5] M. Fuhrman, "Developing Good Multiple-Choice Tests and Test Questions," Journal of Geoscience Education **44**, 379 (1996).
[6] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction*. (NSTA, 1980).
[7] E. E. Ghiselli, J. P. Campbell, and S. Zedeck, *Measurement Theory for the Behavioral Sciences*. (W. H. Freeman and Company, 1981).
[8] A. J. Nitko, *Educational Tests and Measurement: An Introduction* (Harcourt Brace Jovanovich, Inc., 1983).
[9] W. A. Mehrens and I. J. Lehmann, *Measurement and Evaluation in Education and Psychology* (Holt, Rinehart and Winston, Inc., 1991).
[10] W. Wiersma, *Research Methods in Education: An Introduction* (J. B. Lippincott Company, 1969).
[11] W. A. Mehrens and I. J. Lehmann, *Using Standardized Tests in Education* (Longman, 1987).
[12] W. J. Popham, *Modern Educational Measurement :A Practitioner's Perspective*, 2nd ed. (Prentice Hall, Englewood Cliffs, N.J., 1990).
[13] T. M. Haladyna, *Developing and Validating Multiple-Choice Test Items* (Lawrence Erlbaum Associates, 1994).
[14] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Wadsworth Group/Thomson Learning, 1986).
[15] D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory," Phys. Teach. **30**, 141 (1992).
[16] *Ibid*.
[17] A version of Ref. 15 revised in 1995 by I. Halloun, R. Hake, E. Mosca, and D. Hestenes is available in E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).

[18] E. Mazur, "Farewell, Lecture?" Science **323**, 50 (2009).
[19] Hestenes, Ref. 15, p. 142.
[20] D. F. Treagust, "Development and Use of Diagnostic Tests to Evaluate Students' Misconceptions in Science," Int. J. Sci. Educ. **10**, 159 (1988).
[21] W. A. Mehrens and I. J. Lehmann, in *Measurement and Evaluation in Education and Psychology* (Holt, Rinehart and Winston, Inc., Chicago, 1991), p. 128.
[22] P. Tamir, "Justifying the Selection of Answers in Multiple Choice Items," Int. J. Sci. Educ. **12**, 563 (1990).
[23] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Methuen & Co., New York, 1986), p. 1.
[24] Beichner, Ref. 4, p. 753.
[25] R. J. Beichner, private communication.
[26] A few of the websites that list conceptual multiple-choice tests are http://www.ncsu.edu/PER/TestInfo.html; http://www.flaguide.org/tools/tools_discipline.php; http://www7.nationalacademies.org/bose/Libarkin_CommissionedPaper.pdf; and http://www7.nationalacademies.org/bose/Reed_Rhoads_CommissionedPaper.pdf.
[27] Beichner, Ref. 4, p. 751.
[28] P. V. Engelhardt, PhD Dissertation, North Carolina State University, 1997.
[29] D. R. Sokoloff and R. K. Thornton, *Tools for Scientific Thinking* (Vernier Software, 1993).
[30] D. R. Sokoloff, R. K. Thornton, and P. Laws, *RealTime Physics* (Vernier Software, 1995).
[31] Nikto, Ref. 8, pp. 93-115.
[32] M. A. Lorber and W. D. Pierce, "Writing Precise Instructional Objectives" in *Objectives, Methods, and Evaluation for Secondary Teaching* (Prentice Hall, New Jersey, 1990), pp. 18-35.
[33] R. D. Knight, B. Jones, S. Field, et al., *Instructor Guide, College Physics: A Strategic Approach* (Pearson/Addison Wesley, Boston, 2007).
[34] Nikto, Ref. 8, p. 95.
[35] Aubrecht and Aubrecht, Ref. 2, p. 614.
[36] *Taxonomy of Educational Objectives Handbook I: Cognitive Domain*, edited by B. S. Bloom (McKay, New York, 1956).
[37] Popham, Ref. 12, pp. 49-52.
[38] *Ibid.*, p. 51.
[39] Fuhrman, Ref. 5, p. 381.
[40] See Refs. 5, 8, 12, and 13.
[41] R. L. Linn and N. E. Gronlund, *Measurement and Test in Teaching* (Merrill, Upper Saddle River, N.J., 2000), p. 203-214.
[42] T. Foster, private communication.

[43] M. H. Ashcraft, *Human Memory and Cognition* (HarperCollins Publishers, USA, 1989), p. 141.
[44] T. Foster, private communication.
[45] Linn and Gronlund, Ref. 41, p. 76.
[46] *Standards for Educational and Psychological Testing* (American Psychological Association, Washington, DC, 1992) p. 9.
[47] P. V. Engelhardt and R. J. Beichner, "Students' Understanding of Direct Current Resistive Electrical Circuits," Am. J. Physics **72**, 98 (2004).
[48] Kline, Ref. 22, p. 6.
[49] J. Salvia and Ysseldyke and J. E. Ysseldyke, *Test* (Houghton Mifflin Company, Boston, 2001), p. 147.
[50] Linn and Gronlund, Ref. 41, p. 79; Mehrens and Lehmann, Ref. 11, p. 83; and Popham, Ref. 12, p. 97.
[51] F. Goldberg and L. C. McDermott, "Investigation of Student Understanding of Real Image Formation by a Converging Lens or Concave Mirror," Am. J. Physics **55**, 108 (1987).
[52] D. I. Dykstra, Jr. and W. S. Smith, "Doing Physics: Image Formation by Lenses" workshop presented at American Association of Physics Teachers, Boise, ID, (1993).
[53] Mazur, Ref. 17, pp. 174-188.
[54] Crocker and Algina, Ref. 14, p. 221.
[55] See website at http://listserv.boisestate.edu/archives/physlrnr.html.
[56] See website at http://www.aapt.org/.
[57] See website at http://www.compadre.org/.
[58] See website at http://www.compadre.org/per/.
[59] Crocker and Algina, Ref. 14, p. 322.
[60] Salvia, Ref. 49, p. 78.
[61] Engelhardt, Ref. 28, p. 82.
[62] A. Anastasi, *Psychological Testing* (Macmillan Publishing Company, New York, 1988), p. 109.
[63] Salvia and Ysseldyke, Ref. 49, p. 120.
[64] Ghiselli, Campbell and Zedeck, Ref. 12, p. 432.
[65] Crocker and Algina, Ref. 14, p. 138.
[66] *Ibid.*, p. 136.
[67]Wiersma, Ref. 10, p. 181.
[68] Doran, Ref. 6, p. 104.
[69] Kline, Ref. 22, p. 16.
[70] Salvia and Ysseldyke, Ref. 49, p. 135.
[71] *Ibid.*, p. 138-139.
[72] D. MacIsaac and K. Falconer, "Reforming Physics Instruction via RTOP," Phys. Teach. **40**, 479 (2002).

[73] Linn and Gronlund, Ref. 41, p. 83.
[74] Popham, Ref. 12, p. 108.
[75] Anastasi, Ref. 62, p. 154.
[76] *Ibid.,* p. 374.
[77] See for example http://linus.highpoint.edu/~atitus/assess/.
[78] Salvia and Ysseldyke, Ref. 49, pp. 155-157.
[79] Linn and Gronlund, Ref. 41, p. 84.
[80] L. McCullough, "A Gender Context for the Force Concept Inventory," a talk presented at American Association of Physics Teachers, in San Diego, CA, (2001).
[81] Doran, Ref. 6, p. 96.
[82] L. Ding, R. Chabay, B. Sherwood, et al., "Evaluating an Electricity and Magnetism Assessment Tool: Brief Electricity and Magnetism Assessment," Phys. Rev. ST. PER **2**, 010105 (2006).
[83] Kline, Ref. 22, p. 8.
[84] Doran, Ref. 6, p. 98.
[85] *Ibid.*, p. 99.
[86] Beichner, Ref. 4, p. 752.
[87] *Ibid*.
[88] Ding et al., Ref. 82, p. 010105-4.
[89] Salvia and Ysseldyke, Ref. 49, p. 115.
[90] http://en.wikipedia.org/wiki/Standard_score.
[91] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, et al., "Surveying Students' Conceptual Knowledge of Electricity and Magnetism," Phys. Educ. Res., Am. J. Phys. Suppl. **69**, S12 (2001).
[92] J. D. Marx, Doctoral dissertation, Rensselaer Polytechnic Institute, 1998.
[93] Ding et al., Ref. 82.
[94] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory* (Sage Publications, Inc., Newbury Park, CA, 1991), p. ix-9.
[95] Crocker and Algina, Ref. 14, p. 339-340.