

# Creating Usable Assessment Tools: A Step-by-Step Guide to Instrument Design

\*\* draft \*\*

Susan L. Davis and A. Katherine Morrow

**Susan L. Davis**

Center for Assessment & Research Studies

James Madison University

davissl@jmu.edu

**A. Katherine Morrow**

Indiana University-Purdue University Indianapolis

annmorro@iupui.edu

## TABLE OF CONTENTS

<b>Introduction</b> .....	<b>4</b>
Question 1: what is it that you want to measure? .....	5
Question 2: Why are you developing this instrument? .....	6
Question 3: How do you want to measure this construct? .....	6
Question 4: Who will be taking the test? .....	7
Question 5: What are the conditions of measurement? .....	7
<b>Worksheet 1. – Defining your purpose</b> .....	<b>9</b>
<b>Defining your Construct</b> .....	<b>10</b>
Breadth vs. Depth.....	11
Writing objectives.....	12
Some Sample Objectives .....	13
<b>Worksheet 2. Defining your construct</b> .....	<b>15</b>
<b>How to search for available instruments</b> .....	<b>16</b>
Commercial Instruments.....	16
Non-commercial instruments.....	17
<b>Worksheet 3 – Search for Available instruments</b> .....	<b>19</b>
<b>Creating the Test Blueprint</b> .....	<b>20</b>
<b>Worksheet 4. The Test Blueprint</b> .....	<b>22</b>
<b>Writing the items</b> .....	<b>23</b>
Item STEMS .....	25
Response Options.....	28
Finishing Touches.....	31
<b>Review of Items</b> .....	<b>33</b>
Selecting the item reviewers .....	33
Determining what to look for.....	33
Backwards translation.....	33
Back to the blueprint.....	33
<b>Reliability: a little more than clockwork</b> .....	<b>36</b>
Defining reliability.....	37
Why do we need reliability? .....	37
How to measure reliability.....	38
Coefficient of equivalence (Alternate-form Reliability).....	38
Stability coefficients (test-retest Reliability) .....	39
Internal consistency Reliability.....	40
When to use each coefficient .....	41
Impacts on reliability .....	42
How to interpret your reliability .....	42
<b>Validity: Definition and Introduction</b> .....	<b>44</b>
Important points regarding validity .....	44
Conducting a validity study .....	45
Validity evidence based on the content aspect .....	46
Validity evidence based on the construct aspect.....	46
Validity evidence based on consequences .....	47
Who is responsible for validation? .....	47

<b>Item Analysis</b> .....	<b>49</b>
Item difficulty .....	49
Item discrimination .....	49
Item bias .....	50
Modifying and re-piloting items .....	50
<b>Ethical Issues in Testing</b> .....	<b>51</b>
Common ethical practices .....	51
<b>Reporting</b> .....	<b>53</b>
Norm-referenced test .....	53
Criterion-referenced tests .....	53
<b>Other Types of Tests</b> .....	<b>55</b>
Affective scales .....	55
Performance measures .....	55
<b>Final Note</b> .....	<b>56</b>

## INTRODUCTION

Assessment and program evaluation have been a part of higher education for a number of years now. More recently, legislators, regional and specialized accrediting bodies, and other stakeholders are asking colleges and universities to demonstrate the effectiveness of their academic and co-curricular programs. Assessment of student outcomes is a responsible way for professionals in higher education to respond to the stakeholders. Campus professionals embark on the task of assessment with a variety of experience, enthusiasm, and expertise. We believe this manual will be beneficial to the novice as well as the veteran practitioner conducting assessment.

Assessment of student learning and developmental outcomes, like any large project, is more manageable when broken down into smaller steps. One of the steps most frequently asked about is that of instrument design. This manual was developed to assist those who are interested in developing assessment instruments yet have not had training in measurement theory. The process of instrument development in and of itself can be somewhat overwhelming. However, we believe this manual describes the process in a logical fashion so that crafting an instrument will be an achievement rather than a source of frustration. We also believe that a well crafted instrument will allow the user to obtain worthwhile information that can be used in decision making processes.

We provide a number of different examples in an attempt to describe the instrument design process fully. Throughout the manual, we will use a running example featuring Dr. Carson, a statistics professor at State University. Dr. Carson has been commissioned by his department head to assess the statistics knowledge of sophomores at State University. All students at State U. are required to take an introductory statistics course as part of their general education requirements by the second semester of their sophomore year. We hope that by following this continual example, you will better understand the instrument design process as it unfolds. Additionally, we provide worksheets that can be completed by the manual user. These worksheets can be found at the end of each section and can assist the user while he/she is crafting an instrument.

## DEFINING YOUR PURPOSE

Often, when people foresee the idea of developing an instrument they feel one of two emotions: fear or excitement (sometimes a little of both). The fear comes from engaging in a process they have never attempted before and the excitement comes from the idea of having an instrument to measure exactly what they want. Often, those interested in developing an instrument are already writing items in their mind before they even have any kind of plan for their design on paper. On other occasions, the state of fear is evident, as these same people cannot imagine tackling the huge task of developing an entire instrument. Taking time to define the purpose of developing the instrument can help make the project become more manageable as well as lay out a clear plan for action. By following the process laid out below and answering each of the five questions in the following worksheet, you will have a clear, coherent idea of your purpose and mission.

### Question 1: WHAT IS IT THAT YOU WANT TO MEASURE?

First, you need to broadly define what it is you want to measure. Specifically, you need to identify the construct of interest. The specifics of your construct will be defined in a later section. Most often, the topic of interest will fall into one of two categories:

- ✓ A type of cognitive **achievement**, either a knowledge or skill (e.g. math skills or knowledge of American history)
- ✓ A type of **affective trait** (e.g. motivation, interest in math)

The concept of achievement can be broken into knowledge and skills. Tests of knowledge measure an individual's understanding and comprehension of facts, concepts, and principles. Tests of skills involve testing the application of this knowledge to problems or situations (Haladyna, 1997). An example of this distinction with respect to our statistics example would be:

- ✓ Knowledge item – What is the difference between a median and a mean?
- ✓ Skill item – Given the set of test scores of 55, 89, 74, 68, 92, 73, 85, and 66; compute the mean.

This manual will focus primarily on the development of the instruments designed to measure achievement of knowledge or skill. A later section covers some of the procedural differences when developing a measure of an affective construct vs. a measure of an ability.

Dr. Carson has decided to develop a measure of basic college-level statistics including knowledge of terms, procedures, and methods, and the skills involved in computing basic statistics.

## **Question 2: WHY ARE YOU DEVELOPING THIS INSTRUMENT?**

This is a very important question to answer at the beginning of this process. There are several reasons why instruments measuring achievement are created:

- ✓ To assess learning from a particular course or subject area.
- ✓ To assess the effectiveness or outcome of a program
- ✓ To assess the level of student knowledge in comparison to a particular competence

There are other possible reasons why an instrument would be developed; the above list is by no means all-inclusive. It is important to define your purpose to justify the time and effort that will be put into this process by yourself and others.

In our example, Dr. Carson is developing his instrument to measure general statistics achievement to assess the effectiveness of the statistics portion of the general education program at State University.

He is developing this instrument because he has been commissioned by the head of his department to identify the strengths and weaknesses of the statistics education that all students must take as a requirement of State U.

## **Question 3: HOW DO YOU WANT TO MEASURE THIS CONSTRUCT?**

### *Format*

The most common type of instrument is a selected response format. This format is relatively easy to administer and easy to score. Despite the positive benefits of the selected response format, many researchers are exploring the option of performance assessment. Some common examples of performance assessments include having students responding to an essay prompt, playing a piece of music on an instrument, or performing a science experiment in front of a group of raters. Due to the complexity of design and evaluation of performance assessments, this manual will primarily focus on the design issues involved in selected response instruments (see Stiggins, 1987 for a procedural review of performance assessment development). Some considerations for performance assessments will be discussed in a later section of this manual.

### *Medium*

Traditionally, tests are designed to be administered on paper; however, advances in technology allow computerized versions of tests to be created. The use of computers in test administration has led to the development of adaptive tests as an alternative to the traditional standardized administration of items as a set. The computer has also added the capabilities for including some alternative item types (e.g. items with audio or video files, items that allow for internet searches). Despite the sophistication of the computer as a testing medium, the same principles presented in this manual for developing a test must be followed to ensure that the results are reliable and valid.

Dr. Carson has considered the options presented above and has decided that because he wants to test all the students at his University, he will want to create a selected-response exam that will be administered on paper. He hopes that as the administrators at his University see the test in use they will be inclined to give him funding to design and administer a computer-based test.

#### **Question 4: WHO WILL BE TAKING THE TEST?**

Defining the target population of a test is extremely important at the outset of instrument development for several reasons. First, as we will discuss later, all pilot studies, reliability and validity evidence will be used to describe the appropriateness of the test for a particular population. Second, if a test is sold or made available for public use, the intended audiences must be identified. Just as you wouldn't give medicine for an adult to a young child, it is wrong to give a test designed for one population to members of another. Examples of well-defined populations could be:

- ✓ An assessment of U.S. history knowledge designed for college students.
- ✓ An assessment of grammar knowledge for high-school seniors.
- ✓ A measure of self-confidence for teenage rape victims

Dr. Carson has defined his target population as all college students because he hopes to measure the statistics knowledge of all students – including those who have taken no classes in this area to those who have taken several.

#### **Question 5: WHAT ARE THE CONDITIONS OF MEASUREMENT?**

This is the final question in determining the purpose of your instrument. You need to clearly delineate how the test will be used. Specifically, will it be low-stakes or high-stakes? Clearly, any time someone takes the time and effort to develop, administer, and score a test it is high-stakes to someone. However, it is important to consider what implications this decision has for those who are taking the test. With low-stakes tests (as are often found in university assessments), motivation can be a concern and something that should be considered during the development of the instrument. With tests that are of high-stakes for the students other issues arise such as security, examinee anxiety, and cheating. Secondly, how long do you want the test to be? Some test developers have the freedom to make the test as long as they feel it needs to be, however, others are restricted by a time limit depending on how the test will be administered.

From the administration at State U., Dr. Carson has been told that his new statistics test will be of low-stakes for the students taking it as it will be given to all sophomores, not just those enrolled in a statistics class. The test will be administered during a required testing session that sophomores attend at the end of the school year.

Now it's your turn – answer each of these questions in Worksheet 1 pertaining to the instrument you wish to develop.

*For further reading on this, see:*

Embretson, S. (1985). *Test Design: Developments in psychology and psychometrics*. Orlando: Academic Press.

Haladyna, T. (1999). *Developing and validity multiple-choice test items*. Mahwah: Lawrence Erlbaum Associates.

Haladyna, T. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Needham Heights, MA: Allyn & Bacon.

Miller, P.W. & Erickson, H.E. (1990). *How to Write Tests for Students*. National Education Association: Washington, D.C.

Reckase, M. (1996). Test Construction in the 1990s: Recent approaches every psychologist should know. *Psychological Assessment*, 8(4), 354-359.

Stiggins, R. (1992). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice*, 12.

Wood, D.A.(1960). *Test Construction Development and Interpretation of Achievement Tests*. Columbus, OH: Charles E. Merrill Books, Inc.

**WORKSHEET 1. – DEFINING YOUR PURPOSE**

1) What is it that you want to measure?

2) Why are you developing this instrument?

3) How do you want to measure this construct?

4) Who will be taking the test?

5) What are the conditions of measurement?

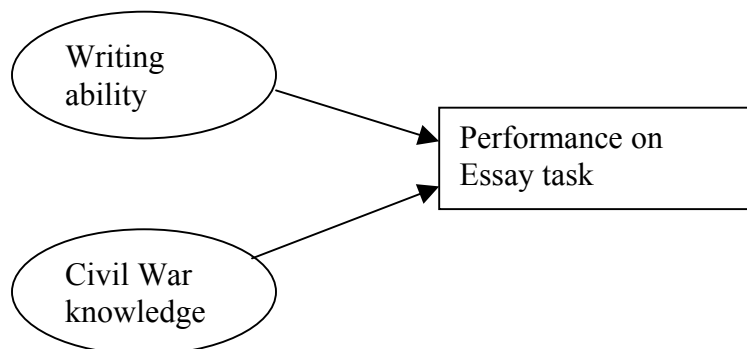
## DEFINING YOUR CONSTRUCT

Now that you have determined what it is you want to measure, the next step is to articulate clearly what is included within the realm of your construct of interest. Constructs are often latent variables. A latent variable is an immeasurable trait including things like cognitive ability to self-esteem (DeVellis, 1991). For example, scientific reasoning is a latent construct; you cannot look at a person and tell what their level of scientific reasoning is. Therefore, you have to design a measure to assess this type of construct. If the scale is designed properly, the items (i.e., indicators) will tap into a person's cognitive ability and the scores should reflect an examinee's level of ability. If the construct is not well defined, it is extremely difficult to write good items (Spector, 1992). Sometimes constructs can be more abstract (e.g. critical thinking) while others are clearer (e.g. knowledge of United States History from 1850-present). The process of defining the construct includes specifying:

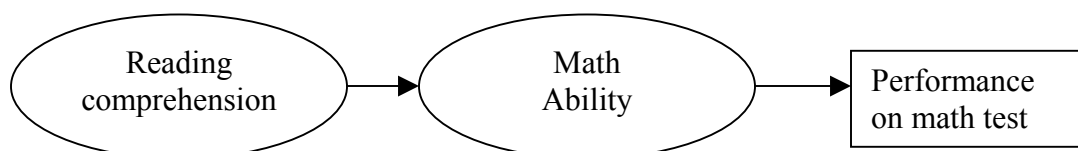
- ✓ What is included within the construct
- ✓ What is not included within the construct

There are several ways to approach the task of defining the construct. The method you choose to use will most likely depend on what type of construct you wish to measure. For example; if you are measuring something like critical thinking, you might want to search the literature for definitions of critical thinking. However, if you are interested in a construct such as U.S. History knowledge, you might search for information on what amount of history knowledge your population should know. Whichever method you choose, be prepared to find some overlap but also some difference of opinion. For example, one researcher may use the definition of critical thinking that includes being able to read one side of an argument and then write a well thought out debate to that argument. Another researcher might disagree and say that writing is not a component of critical thinking. The definition you decide to use is your choice. Often, people select the definition that is most prominent in the literature or the one most is more applicable to their particular situation.

It may be the case that you have more than one construct that you are trying to measure with your instrument. If so, they should all be defined, and more importantly, the relationships between them should be established. For example, if I wanted to measure students' ability to write an essay on the ramifications of the Civil War, I would be measuring more than one construct. Specifically, I would be evaluating their writing ability and knowledge pertaining to the outcome and effects of the Civil War. To illustrate how these constructs relate, we can draw a path diagram as outlined below:



As is indicated, both constructs (knowledge and ability) will have a direct influence on the students' performance on the essay task. In a different scenario, the constructs may not both be directly related to the observable measure. For example, you might want to administer a test to assess math ability but you want to use items that are in the form of "story problems." Therefore, students' ability to solve the math problem would be affected by their reading comprehension. If this was the case, your path diagram might look like:



## BREADTH VS. DEPTH

Included in this decision is determining the coverage your test will have of the overall construct. Specifically, we are referring to the breadth vs. depth argument. Do you want your instrument to measure knowledge in a wide variety of areas or do you want it to cover one smaller area in great detail? For example, if your test of U.S. history was designed for breadth you might have a few items on the American Revolution, a few on the settlement of the West, a couple on the Vietnam War, etc. However, if you designed your test for depth, you might have items that focused specifically on colonial U.S. History. Again, this is part of defining your construct.

Dr. Carson had previously defined his construct as basic college level statistics. To determine what particular statistical skills fall within this construct, he has sought out the help of the statistics faculty at State University. They have provided him with lists of the topics they covered in their basic level statistics classes from which he was able to delineate the types of statistical skills that students should have.

Dr. Carson determined that within his construct the following topics were included:

- Populations & Samples
- Variables & Data
- Measures of Central Tendency
- Measures of Variation
- Sampling Distribution
- Hypothesis Testing
- Student's t Distribution
- Confidence Intervals

Topics within statistics that were not a requirement of the basic level classes and things he did not want to include within his test were:

- Probability
- Straight-line Model

## WRITING OBJECTIVES

After you have in your mind what it is you want to measure and what this construct includes and what it does not include, the next step is to identify learning objectives. Simply, the objectives are a set of statements that list what it is the department or program is trying to do (Erwin, 1990). Often, test developers mistakenly focus the objectives on what the test is supposed to measure. Rather, the objectives should clearly define what students should be able to do as a result of the learning experience. These defined abilities should be measurable by the testing method that has been chosen. Often, if you are assessing the success of a program, the program directors may have previously defined objectives detailing the desired outcomes from their program.

It is important to state the objectives up front to ensure that you are using the proper assessment method. For example, a learning objective could be: ‘After taking this chemistry class, a student will be able to conduct a test of the acidity of a liquid’. In this case, you would want to use some form of performance assessment where the student is judged on their ability to carry out the experiment successfully. In a different situation, you might just be interested in determining if students know what types of liquids are considered acidic; this type of objective lends itself to using a selected-response assessment.

The process of writing objectives involves summarizing the topics that were previously defined as part of the construct by putting these topics into observable knowledge and skills. As far as the wording of the objectives, Kubiszyn & Borich (1996) offered some guidelines for writing clear objectives. The main point of their position is the phrasing of the objectives. First, objectives should contain action verbs; whatever actions you would like to see the examinee demonstrate (i.e. define, describe, demonstrate, illustrate, compare, contrast, explain, etc.). Second, the objectives should describe some form of observable behavior. For example, the objective:

“Students should understand basic-level geometry”

is not an observable behavior. You cannot directly assess an understanding of a topic area. However, you could use the following as an objective:

“Students should demonstrate understanding of angle calculation”

Finally, it is important that objectives are written in terms of the test-taker. Do not list what you want the test to do or what you want to see as an answer, rather, list an action that will be demonstrated by the student. This includes using the terms student or test-taker in each objective.

When writing objectives it is necessary that care is taken not to make them too broad or too narrow. For example, an objective that is too broad would be:

“Students will demonstrate an understanding of the importance of historical literature.”

This is very unclear. Importance of the literature to what? What do you mean by historical literature? Similarly, it is important that objectives are not too narrow:

“Students will demonstrate an understanding of the importance of Machiavelli’s *The Prince* to the political world in the early 1500’s.”

Instead, objectives should tap into an area of knowledge or skills that encompasses several smaller topics that questions can be formulated for. For example:

“Students will be able to delineate the importance of Greek philosophy on modern social science.”

### **SOME SAMPLE OBJECTIVES**

These objectives are from the Association of College and Research Libraries (ACRL) regarding student Information Literacy. The full explanation of Information Literacy and these standards is available at:

<http://www.ala.org/ACRLTemplate.cfm>

ACRL Information Literacy learning objectives:

**Standard One:** The information literate student determines the nature and extent of the information needed.

**Standard Two:** The information literate student accesses needed information effectively and efficiently.

**Standard Three:** The information literate student evaluates information and its sources critically and incorporates selected information into his or her knowledge base and value system.

**Standard Four:** The information literate student, individually or as a member of a group, uses information effectively to accomplish a specific purpose.

**Standard Five:** The information literate student understands many of the economic, legal, and social issues surrounding the use of information and accesses and uses information ethically and legally.

From his list above, Dr. Carson worked with the faculty from the statistics department to formulate a list of objectives. Several of these objectives are listed below.

- ✓ Students will be able to interpret the results of a hypothesis test
- ✓ Students will be able to compute confidence intervals
- ✓ Students will be able to present analysis results clearly
- ✓ Students will be able to write clear hypotheses

Now it is your turn to specify the details of your construct. Use Worksheet 2 to specify the details of your construct and write objectives of your project.

*For further reading on constructs, see:*

Bollen, K. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53, 605-634.

DeVellis, R. (1991). Scale development: Theory and applications. Applied Social Research Methods Series, Vol. 26. Newbury Park: SAGE publications.

Erwin, T. D. (1990). Assessing Student Learning and Development. San Francisco: Jossey-Bass.

Haladyna, T. (1997). Writing Test Items to Evaluate Higher Order Thinking. Needham Heights, MA: Allyn & Bacon.

Kubiszyn & Borich. Measuring Learning Outcomes. In *Educational Testing & Practice*.

Spector, P. (1992). Summated rating scale construction. Newbury Park: SAGE Publications.

## WORKSHEET 2. DEFINING YOUR CONSTRUCT

1. What is your construct of interest?

2. Specifically, what components of the construct do you want to include?

3. Are there parts of the construct that you do not want to assess in your instrument?

4. Objectives – take the components from step 2 and from them, formulate objectives to delineate what it is you want to measure.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.



## HOW TO SEARCH FOR AVAILABLE INSTRUMENTS

Once you have defined your purpose and your construct it is appropriate to begin making choices about the test you will use. First, you may want to consider using a test that already exists. There are numerous instruments available to educators that measure a variety of academic skills and abilities with so many tests out there why do we need to consider developing our own test? It is important to carefully examine each existing relevant instrument to ensure it is psychometrically sound and is the appropriate tool for the job. While this manual focuses mainly on developing instruments, there is no sense in “reinventing the wheel” when an appropriate instrument already exists.

The first step in searching for available instruments is to look for a commercial or non-commercial instrument. A *commercial instrument* is one that is available for purchase from a publisher. Some instruments can be scored by the test administrator on site while others must be returned to the publisher for analysis. A *non-commercial instrument* is one that is in the public domain. That is, the instrument can be used without cost and is usually available by contacting the test author.

### COMMERCIAL INSTRUMENTS

There are 4 common places to search for a commercial instrument:

- *Mental Measurements Yearbook*
- *Tests: A comprehensive reference for assessments in psychology, education, and business*
- *Test Critiques*
- *Test Critiques Compendium: Reviews of major tests from the Test Critiques series*

*Mental Measurements Yearbook* consists of 15 volumes and is published by the Buros Institute of Mental Measurements, Lincoln, Nebraska. This text can be found in many university libraries in hard copy form or online in electronic format. *Mental Measurements Yearbook* provides information about the test itself, 1-3 critiques of the test, and contact information for the test publisher. The Buros Institute website is: <http://www.unl.edu/buros/>

*Tests: A comprehensive reference for assessments in psychology, education, and business* has gone through 4 editions and contains information about each test such as the population for whom the instrument is intended and the contact information for the publisher. This text does not contain critiques of the tests.

*Test Critiques* also contains several volumes and provides critiques to instruments identified in *Tests: A comprehensive reference for assessments in psychology, education, and business*. *Test Critiques* and *Tests* should be used together for more complete information.

*Test Critiques Compendium: Reviews of major tests from the Test Critiques series* is exactly what its title implies – a concise summary of *Test Critiques*.

*Tests*, *Test Critiques*, and *Test Critiques Compendium* can all be found in many university libraries.

## NON-COMMERCIAL INSTRUMENTS

Non-commercial instruments can be found in scholarly research publications. There are 5 common places to search for a non-commercial test:

- Research databases
- *Tests in Microfiche*
- Health and Psychosocial Instruments (HAPI)
- Directory of Unpublished Experimental Measures
- Measures of Psychological Assessment: A guide to 3,000 original sources and their applications

The above mentioned resources can be obtained in most university libraries. Research databases such as PsycInfo or ERIC are available at any university library. Different subject areas may have other searchable databases and may have much needed information about instruments and/or test content. When searching for an instrument it can also be helpful to locate published research that uses a particular instrument.

*Tests in Microfiche* provide information about instruments that have been cited in the literature.

*Health and Psychosocial Instruments (HAPI)* contains information about health, psychological, and social science instruments.

*Directory of Unpublished Experimental Measures* contains several volumes and can be found in many university libraries. This directory contains psychometric information about educational and psychological tests.

*Measures of Psychological Assessment: A guide to 3,000 original sources and their applications* was published in 1975, but still provides useful information about a variety of measures.

Once the search for instruments has taken place, one or more measures may be considered for use. A careful analysis of the instrument and all available information is needed. First, the instrument should be closely aligned with the construct(s) being measured to determine if the instrument is appropriate for the stated purpose. This means that the items on the instrument should be appropriate for the audience and aligned with the proposed outcomes. To ensure the test is right for your purpose consider the following questions:

- ✓ Is the instrument designed to measure your population?
- ✓ When was the instrument normed?
- ✓ On what population was the instrument normed?
- ✓ What is the reliability evidence of the instrument scores?
- ✓ Is there any validity evidence for using the instrument in the desired manner?

It is important to consider the norming population for an existing instrument. One advantage of using a pre-existing instrument is that you can avoid much of the initial test development work by using the existing score interpretation provided by the test developers. However, if the test was normed on a population different from your own, you need to consider how you would obtain your own norming data/score interpretation and if the test would be appropriate for your

population. In addition, if you choose to use the instrument, it is important to gather your own reliability and validity evidence as your audience may be not be similar the audience with whom the instrument was normed.

However, if the items do not appropriately cover the defined learning objectives or the population for whom the instrument was originally intended is substantially different from your own then the instrument should not be selected for use. It is also important to consider how the test is going to be used. We will discuss validity in detail in a later section of this manual. In the meantime suffice it to say that validity evidence is needed for any intended use of the test interpretations. We do not advise using an instrument with weak reliability and validity evidence.

If the existing instrument(s) are not right for your needs then it is time to begin crafting your own instrument. The good news is that you already well on your way!

Dr. Carson conducted a search of the literature and found two instruments he felt might meet his needs. The first instrument was a commercial statistics instrument designed for high school students who completed an elective statistics and probability class. When Dr. Carson closely examined the population for whom the instrument was designed he decided that instrument was not appropriate for his purposes.

The second instrument Dr. Carson found was published in the literature and was designed specifically for college students. Encouraged by this fact, Dr. Carson next examined the instrument items and tried to find connections between the items on the test and the objectives of the introductory statistics course at State U. He realized that the items on the existing instrument did not cover his objectives. Dr. Carson felt the test did not match with the components of the statistics classes that needed to be assessed and decided he should create his own instrument.

## WORKSHEET 3 – SEARCH FOR AVAILABLE INSTRUMENTS

<b>NON-COMMERCIAL INSTRUMENTS TO BE CONSIDERED</b>		
Title	Author	Reference
1.		
2.		
3.		
<b>COMMERCIAL INSTRUMENTS TO BE CONSIDERED</b>		
Title	Publisher	Reference
1.		
2.		
3.		
<b>INSTRUMENT INFORMATION</b>		
<p>Instrument Title: _____</p> <p>For whom was the instrument designed? _____</p> <p>When was the instrument normed? _____</p> <p>On what population was it normed? _____</p> <p>Evidence of reliability:</p> <p>Evidence of validity:</p> <p>How well do the items match to the desired objectives?</p>		

## CREATING THE TEST BLUEPRINT

After you have specified the construct of interest and have defined objectives that fall under this construct, the next step is to make a table that includes the final specifications of your instrument (Reckase, 1996). This is called the test blueprint. The use of a blueprint will ensure that all of the objectives that are considered important will be included and questions will only be written to serve as indicators for objectives that are included in the definition of the construct (Mehrens & Lehmann, 1984). An example of a test blueprint is shown below.

	<b>Portion of test (%)</b>	<b>Corresponding items</b>
Objective 1		
Objective 2		
Objective 3		
Objective 4		
Objective 5		

To complete the test blueprint you need to first determine how many items you wish to have corresponding to each subscale. This is only an estimate, based on how many items you think you will have on your final version of the test. As mentioned earlier, the number of items (or length of the test) will be dictated by the conditions of measurement. Determining the number of items that will be on each subscale is again only an estimate but will be affected by the amount of weight you want to put towards measuring each objective. By establishing this at the beginning, it will help to ensure that the test is properly balanced in emphasis (Hopkins & Stanley, 1981; Mehrens & Lehmann, 1984). While many instrument developers write their objectives to have equal importance, content area experts suggest that certain objectives need more questions than others to cover the content area. The ability to assign different weights to different subscales is an advantage of creating your own instrument over using one made by another organization (Mehrens & Lehmann, 1984).

Once the objectives and appropriate weights have been established, the item writing can begin. As each item is written, it is important to make note of which objective the item was created to measure. Assuring that the test creation follows this blueprint will add to the construct validity of the instrument. The importance and evaluation of validity will be discussed in a later section of this manual.

Dr. Carson has used the objectives that he created along with the statistics professors at State University to create his test blueprint.

<b>Objective</b>	<b>Portion of test (%)</b>	<b>Corresponding items</b>
1 - Students will interpret the results of a hypothesis test	30%	
2 - Students will be able to compute confidence intervals	15%	
3 - Students will be able to present analysis results clearly	25%	
4 - Students will be able to write clear hypotheses	15%	
5 - Students will be able to interpret the results of a t-test	15%	

Now it's your turn, take your new objectives and try to complete the Test Blueprint (Worksheet # 3).

*For further reading on this, see:*

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando: Harcourt Brace Jovanovich.

Reckase, M. (1996). *Test Construction in the 1990s: Recent approaches every psychologist should know*. *Psychological Assessment*, 8(4), 354-359.

**WORKSHEET 4. THE TEST BLUEPRINT**

<i>Objective</i>	<i>Portion of test (%)</i>	<i>Corresponding items</i>
1.		
2.		
3.		
4.		
5.		
6.		
7.		
8.		
9.		
10.		

## WRITING THE ITEMS

Finally, it is time to start writing the items for your instrument! Item writing has been described as an art form and as a science (Embretson, 1985). It is important to remember that not every item will be perfectly crafted after one attempt. Most items will go through several rounds of revision before they can be used, while other items may be discarded at one of the stages between creation and final use. However, don't let this process discourage you; write down any possible items you can think of. Don't be afraid to write a bad item; you will learn what makes a good item through practice. Often, problematic items can only be identified when they are piloted on a sample of the population for whom the instrument is created.

The first step in writing a test item is determining the structure (e.g., fill in the blank, true/false, direct question). The second step involves writing the item stem. Before even thinking about the response options, we should make sure that the stem is a straightforward, fair, and thoughtful question. The third step involves creating the response options. This step can be somewhat difficult; however, it is always easier when the question stem has been well formed.

There are many different guidelines that should be followed when writing items for an instrument. Numerous books and papers have been published specifically on this part of instrument creation alone. Several of these guidelines are summarized in this section.

Writing quality items is important in demonstrating that the test is only measuring what it is intended to measure. As the following guidelines will show, if item writing is not carefully monitored, extraneous factors can influence student performance on the items and in turn, performance on the test will not clearly indicate their ability.

### Item Structures

Some common item structures are: straight-forward questions, matching, or true false (with some caution). Item writers have their own opinions about what types of question stems are acceptable and which are not; however, we will present you with our opinion on which types seem to be the most coherent and clear. The types of items that we recommend to be the best typical question stem are:

Straight-forward question:

What is the most commonly broken bone in the human body?

True-false questions:

The fifth metatarsal is the most commonly broken bone in the human body.

Fill in the blank questions:

The \_\_\_\_\_ is the most commonly broken bone in the human body.

There are several other item types that can be written: partial sentences, double completions, etc. The type or types of items you choose to incorporate into your test is completely up to you. Some item types may be more appropriate for different age groups, for measuring different constructs, for different testing situations, etc. This decision is the responsibility of the test developer and items of varying types can be piloted to determine if they are appropriate for that specific purpose.

Haladyna (1999) recommends avoiding use of any kind of true/false questions when writing items. However, we feel that if they are used sparingly and the quality of the items is still ensured, there is no reason why these items cannot be used. The final decision to use true/false items is completely up to the test creator; however, to help in this decision, Frisbie and Becker (1991) summarized the current literature on the advantages and disadvantages of using these items:

Advantages:

- 1) It is possible to sample a wide variety of the content domain thoroughly.
- 2) A great deal of information about achievement can be obtained in the testing time available.
- 3) Items are easy to construct
- 4) Several types and levels of learning can be tested
- 5) Responses can be scored quickly and accurately
- 6) Scores obtained can be quite reliable
- 7) The item presents a realistic task: Judge the truth of a statement
- 8) They are effective in minimizing the amount of reading required of examinees

### Disadvantages

- 1) Guessing can have a significant impact on scores; high scores are possibly due to guessing alone
- 2) They tend to be ambiguous statements
- 3) They fail to account for degrees of truth or shades of correctness in ideas.
- 4) They yield a smaller expected range of scores than some other item formats
- 5) Average item discrimination is lower than for some other item formats.
- 6) They can measure trivial, unimportant information.
- 7) Their use encourages the lifting of verbatim statements from textbooks.
- 8) Their use permits a correct response for knowing “what is wrong” without showing “what is right.”
- 9) They are susceptible to inferences due to response sets.

If a test developer decides to use true-false items, we suggest that they consider these ideas during the development of their test and in turn, use them sparingly and only when appropriate..

### ITEM STEMS

#### 1. Each item should be directly linked to one or two content areas

The test blueprint is necessary here. Each item should be designed to test one of the pre-set objectives as laid out in the test blueprint. The content should be restricted to those areas specified by the definition of the construct. The action that the item is requiring the examinee to perform should be one of the abilities laid out in the objectives.

For example, if one of your objectives for assessing biology knowledge is:

“Students will be able to define terms important to basic genetics”

An appropriate question would be:

What is a phenotype?

An inappropriate question would be:

Mary’s parents both have blue eyes and she also has blue eyes. Is this a result of her phenotype or genotype?

This question is not tapping the stated objective. Rather it is, an indication of whether or not the examinees could apply their knowledge of genetics terms to understanding the genetic makeup of a person.

**2. Items should not be too specific or too general**

The specificity of the questions that you write for your instrument is something that is preset in the stated objectives. When your objectives were written, you most likely considered what you wanted to assess and the desired level of cognitive ability you would expect your examinees to demonstrate. For example, if you were to assess psychology students at the end of their college career, you would not ask them what year Sigmund Freud died. However, you might ask a question that measures their understanding of Freud's theories of the unconscious.

**3. Be cognizant of what skills or knowledge base a particular item is measuring.**

Determine in the early stages of development if the test will include only single content questions or higher-order questions. Complex higher-order questions will most likely be included on an instrument designed to measure the integration of knowledge. However, these questions can be difficult to analyze. Therefore, some item writers would prefer to focus on a single content area for each question. For example, some items that focus on one skill or knowledge area would be:

1. If  $x + 3 = 16$ , what is  $x$ ?
2. Who was president of the United States during the Vietnam War?

Single skill questions can be much more complex than these examples, but what is important is to make sure that they don't require the use of different types of skills or knowledge. For example, if you had a question like:

The president of the United States during the Vietnam War made a huge impact on the U.S. foreign policy with his relations with foreign countries. What was the lasting impact of these policies?

This question not only requires you to know who was president at the time of the Vietnam war, but also, what his foreign policies were and what type of lasting effect they had on the U.S. foreign relations. Therefore, if a student gets this question wrong, a direct link cannot be made to which area of knowledge is lacking. This question would be more appropriate in essay format or broken down into smaller questions for a multiple-choice exam.

**4. Avoid opinion-based items**

While it seems obvious not to include opinion-based items, this flaw still frequently occurs. For example, if you were testing how well a student comprehends material from a particular source, make sure to identify the source so the examinee can clearly understand how to answer the question.

According to the chapter by Smith and Jones, what is the best type of exercise to improve cardiovascular flow?

Where an opinion based item would be:

What is the best type of exercise to improve cardiovascular flow?

### 5. **Avoid trick items**

Trick items can result from one of two things: the item writer trying to be tricky, or the item writer being tricky without realizing it. As a good rule of thumb, an item writer should never *try* to be tricky. This concept will be examined more closely under the topic of writing response options. However, items can appear tricky to examinees and not to the writer. An item writer can create a question with a typical examinee's thought process in mind. However, the examinee might actually become confused through that process and incorrectly respond to the item although they possess the content knowledge to answer the question. This is also a problem if the directions in the item stem are unclear in any way. Detecting this type of confusion will be discussed in the Item Review section.

### 6. **Include the central idea in the stem instead of the choice**

The stem of the item should make clear the question that is being asked. An examinee should be able to read the question and think of the correct answer before even looking at the response options. For example:

Unfocused stem: The mean of a set of scores

Focused stem: How is the mean of a set of scores calculated?

While both of these items can tap the same ability, the first item could have several correct answers. For example, all of the following would be correct responses to the unfocused question stem:

- Is a measure of the average of all the scores
- Is calculated by summing all the scores and then dividing by the number of scores
- Is a measure of central tendency

Whereas the focused stem question only has one possible correct answer. Focusing the item stems will help not only with making the items more comprehensible but also will help many test takers in trying to reason out the best answer.

### 7. **Avoid window dressing (excessive verbiage)**

Excessive wordiness is unnecessary and can lead to decreased item performance as examinees may get wrapped up in the reading and confuse the meaning of the question.

**8. Word the stem positively: Avoid negative phrasing and double-barreled items**

Negatively worded items should be used sparingly. Some different types of negatively worded items are:

Which of the following is NOT a measure of central tendency?

Of the characteristics listed below, which is NOT a characteristic of Schizophrenia?

While the rare occurrence of such items is not necessarily problematic, it can be for certain populations. For example, if your test is designed for a younger population, handling this negative distinction may be a problem. Even for adult populations, research has shown that items that are negatively worded are found to be more difficult than the same type of question without the negative phrasing.

More importantly, item writers should completely avoid writing items that are double-barreled. A double-barreled item is one that contains two or more negative phrases in the stem. An example of a double-barreled item is:

Which of the following is not an unobservable trait?

Notice that you have to read this stem at least twice to fully understand what the question is asking. Again, this added difficulty an indication that an item is measuring something other than the content domain that the item is referring to. Therefore, we recommend avoiding using items of this nature.

**RESPONSE OPTIONS**

When writing response options to a multiple-choice test, one must always include the correct response and at least one distracter option. Usually there will be more than one distracter unless the item writer is using the true-false format. The key to writing good test items is not only having a good item but also having good distracter options. If a novice can clearly identify the correct answer out of the list of options, the distracters may be considered 'weak'. However, strong response options create an item that differentiates between those who know the material and those who do not know the material.

- 1. Use as many distracters as needed, but keep it within a reasonable amount**  
What is the optimal number of response options? Again, test developers debate this point, but recommend the numbers of distracters range from 3 to 5. Distracters are only useful if they are viable answers to the question stem.
- 2. Make sure that only one of these choices is the right answer**

While this seems like an obvious point, often more than one response option exists that could be a correct answer to the question stem. While the item writer should consider this and choose the distracters carefully, initial item analysis can often provide clues to this problem. We stress the importance of having content experts serve as item reviewers to avoid this difficulty.

**3. Vary the location of the right answer according to the number of options**

Make sure all items do not have “c” or “b” as the correct answer and that the response options do not follow a pattern (e.g. a,b,c,d). This is usually checked by making sure there is a balance between the number of items that have the correct answer in the “a” option, the “b” option, etc.

**4. Place options in logical or numerical order**

This rule is one that will only increase the readability of the test and decrease the effect of extraneous factors, such as taking the time to search the response options. Response options can easily be put in a logical order. The following list of response options is listed in a logical order:

What is the square root of 81?

- a) 8
- b) 8.5
- c) 9
- d) 10

**5. Keep choices independent: Choices should not be overlapping**

If choices are overlapping, they can give a clue to the correct response. For example, in the following question response options A and B are overlapping and therefore can provide a clue how to narrow the selection of response options.

What areas of psychology is Freud best known for?

- a) The unconscious and psychotherapy
- b) Psychotherapy and behaviorism
- c) Experimental design and comparative psychology
- d) Socio-emotional development

**6. Keep choices homogenous in content**

Content of choices is something to keep in mind so that examinees can only get the answer correct if they actually know the content area. If the correct response option is not homogeneous with the rest it can stand out and clue the examinee, in turn, a distracter option that is not homogenous can be easily eliminated as a possible option. Consider the following example:

Who uttered the famous quote “Give me liberty or give me death?”

- a) Patrick Henry
- b) Nelson Mandela
- c) Some crazy revolutionary

d) George Washington

Obviously, response option “c” would stand out from the others as an obviously wrong choice, and therefore, not add anything to the test itself.

**7. Keep choice lengths similar**

Keeping the choices consistent in length and phrasing will minimize the number of clues to the correct response. Often item writers mistakenly phrase the correct response to be significantly longer than the distracters. This difference is an immediate clue as to the correct answer.

**8. Avoid using the choices such as none of the above, all of the above, or I don't know.**

These phrases can often confuse examinees when used as response options. Again, the goal of your instrument is not to determine the test-taking skills of the students, rather, their knowledge of the content area. Therefore, if it can be avoided, we do not recommend using these as possible response options.

**9. Phrase choices positively**

This rule follows the guideline of positively wording item stems. Negatively worded options can be confusing and make the item more difficult than intended. Additionally, one negatively worded response option can be a flag to a correct answer out of the list of options.

**10. Avoid various clues to the right answer.**

This rule summarizes some of the overarching points of cluing an examinee into a particular response option. For example, item writers should not use specific determiners (i.e. always, never), clang associations (e.g. same word is in the stem and one distracter that is not the correct answer), grammatical inconsistencies in the response options, pairs or triplets of answers, or ridiculous options.

**11. Make all distracters plausible**

All the response options should be possible answers to the stem questions. As mentioned previously, there is no point in having distracters that would not be considered by an examinee.

**12. Use common errors of students**

One of the best ways to write distracters for items is to use common incorrect responses from previous test takers. For example, when asked to calculate the mean of an item, students may often calculate the median or mode instead. The following question is an example of using this rule to write distracters.

What is the mean of the following numbers: 4, 5, 8, 6, 9, 3, 8, 5, 15, 7

- a) 7 --correct response
- b) 6 -- distracter (mode)
- c) 5 -- distracter (median)

## FINISHING TOUCHES

### 1. Edit and proof items

Never assume that the first draft of an item is going to be the final draft. Items often need several revisions. We are not saying this to discourage anyone from writing items but the process of item writing can be thought of as writing a paper; several drafts will be completed before the final version emerges. The editing that needs to be done by individuals other than the item writer will be discussed in more detail in later sections.

### 2. Keep vocabulary simple for the group of students being tested

This is another way that the test specifications made earlier can help in the development of the test. Earlier, you defined a target population for whom the test was designed. Somewhere in your definition of the population, you specified what ability or age level your test was going to be made for. Keeping this specification in mind will help determine at what level the items are written. For example, make sure your items are not too difficult to read for your target population, this will ensure that you are not measuring reading/vocabulary ability at the same time.

### 3. Use correct grammar, punctuation, capitalization and spelling

This rule does not need extensive interpretation. Simple errors such as these in the test can be distracting and indicate that there was little attention to detail by the test designer.

### 4. Minimize the amount of reading in each item

Consider your time limit for the test. The longer it takes for each item, the fewer items you can have on your test. Therefore, unless the purpose of the item is to test reading comprehension, make sure to keep the wordiness to a minimum.

### 5. Make sure the items are independent of each other

A common mistake in item writing is to somehow embed the answer to one question in the context of another question. This often happens when you have multiple item writers or sample from a large pool of items where overlapping items are not indicated prior to assembling the pool. Students who are “testwise” often use strategies to search out these overlaps and increase their score. However, these items are not only measuring student’s cognitive ability but also, their ability to remember the content of all the previous questions as they progress through the test.

*For further reading on writing items, see:*

DeVellis, R. (1991). *Scale development: Theory and applications*. Applied Social Research Methods Series, Vol. 26. Newbury Park: SAGE publications.

Embretson, S. (1985). *Test Design: Developments in Psychology and Psychometrics*.

- Orlando: Academic Press.
- Frisbie, D. & Becker, D. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education*, 4(1), 67-83.
- Haladyna, T. (1999). Developing and validity multiple-choice test items. Mahwah: Lawrence Erlbaum Associates.
- Haladyna, T. (1997). Writing Test Items to Evaluate Higher Order Thinking. Needham Heights, MA: Allyn & Bacon.
- Miller, P.W. & Erickson, H.E. (1990). *How to Write Tests for Students*. National Education Association: Washington, D.C.
- Wood, D.A.(1960). *Test Construction Development and Interpretation of Achievement Tests*. Columbus, OH: Charles E. Merrill Books, Inc.

## **REVIEW OF ITEMS**

A review of the items will allow for an outside opinion regarding the construct representation of the items and validity of the inferences that are to be made from the items (DeVellis, 1991). This review is extremely important in the process of item development and should be conducted each time significant modifications are made to the items. The following section outlines selecting item reviewers, determining what to look for, conducting a backwards translation, and reviewing the test blueprint.

### **SELECTING THE ITEM REVIEWERS**

The items should be reviewed by someone who is experienced in item writing as well as a content expert in the area of interest. While several reviewers are recommended, it might be a good idea to limit the initial feedback that you get so that it is a manageable amount.

### **DETERMINING WHAT TO LOOK FOR**

DeVellis (1991) lays out the process for expert review very nicely. First, experts should rate the relevance of each item as an indicator of the construct. If an item is not relevant to the construct being measured, it should be rewritten or eliminated. Second, the reviewers should rate the clarity and conciseness of each item. As will be discussed later, problems with item clarity can lead to problems with reliability. Recall, you are not trying to measure an examinee's ability to decipher a complex item, rather, you are measuring their knowledge of the construct at hand. Finally, DeVellis (1991) recommended having the experts point out any aspects of the construct that you have failed to cover adequately. This again will help increase the construct validity of your scale.

### **BACKWARDS TRANSLATION**

The backwards translation exercise can greatly enhance the evidence for construct validity of the instrument. This exercise involves content experts reading each item and then providing their opinion of which objective or subscale the item measures (Dawis, 1987). A well written item clearly will be indicative of one of the learning objectives outlined for the particular assessment. The expert's determination is then compared with the initial assignment created by the test developers. If there are discrepancies between the original assignment and that of the item reviewers, the item could be problematic. It is recommended that at least two content experts review all items so there sufficient comparison.

### **BACK TO THE BLUEPRINT**

After the item review and backwards translation are completed, it is a good idea to revisit the test blueprint. At this point, you might want to re-evaluate your content coverage (% of the test covered by each subscale) based on the feedback from the review process. The test developer should evaluate all of the feedback from experts and make the final determination for any modifications. It is extremely important to document any modifications, additions, or deletions that are made to the test at this point. If items have to be reassigned, do so and make sure your original construct representation is the same as you defined it in the beginning. If it is not, you can either change your content

coverage based on the feedback from the reviewers or write new items to reestablish the subscale representation. If you only have one subscale, having experts rate the relevance of each item to the construct (as described above) can replace this exercise.

*For further reading on this topic, see:*

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, 34, 481-489.

DeVellis, R. (1991). Scale development: Theory and applications. Applied Social Research Methods Series, Vol. 26. Newbury Park: SAGE publications.

Haladyna, T. (1997). Writing Test Items to Evaluate Higher Order Thinking. Needham Heights, MA: Allyn & Bacon.

## PILOTING THE ITEMS

Piloting the items is a very important step. Up until this point, items have been developed based on a thorough understanding of the construct, reviewed by experts, and prepared for use. However, until the items have been actually tested by a sample of the target population, their utility cannot be accurately determined.

The proper sample must be selected for the pilot test administration. In order to make this selection, it is helpful to reflect back to the original purpose of the instrument. The sample chosen for the pilot testing should be drawn from the target population you identified on Worksheet 1, question number 4. If a sample of the target population is not available, the selected sample should be as close to the targeted population as possible. While most consider “pilot sample” to indicate a small sample of the population, you should ensure that your sample is not so small that subject variance will affect your results (DeVellis, 1991). In addition, a small sample might not be representative of the entire population. It is a good idea that you make sure that there is nothing unique about your sample that is not characteristic of the entire population. For example, if you were designing a test for psychology students and you used a night class as your pilot sample, you cannot assume that students who take a class at night are the same as the majority of the psychology students who are enrolled for daytime classes. The students in the night class may differ from the students in the day class in age, work experiences, or life situations.

There are several ways in which items can be pilot tested. First, the standard administration (worksheet 1, question 5) could be implemented to fully understand how the items would function in the desired setting. It is recommended that this type of pilot testing be conducted regardless of the conditions of measurement. However, to get a more complete picture of how the items are functioning, additional methods of pilot testing are recommended.

For example, a think-aloud could be conducted with several members of the target population. A think-aloud involves having members of the target population attempt to answer each of the items while voicing their thought process to one of the test developers. This process allows the test developers to understand how examinees will perceive each of the items and what mental process they use to arrive at the proper solution.

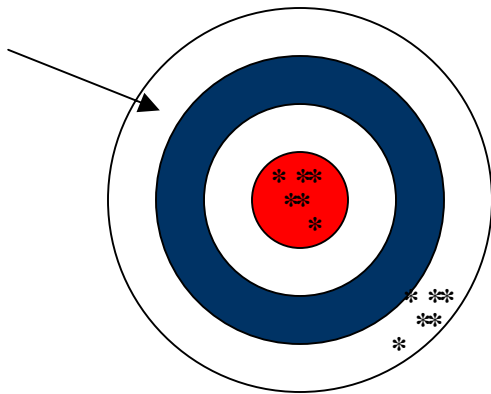
*For further reading on item piloting:*

DeVellis, R. (1991). *Scale development: Theory and applications*. Applied Social Research Methods Series, Vol. 26. Newbury Park: SAGE publications.

## RELIABILITY: A LITTLE MORE THAN CLOCKWORK

An easy way to conceptualize reliability and its relationship to validity is to think of a bathroom scale. If I step on the scale in the morning the scale will read a certain weight. If I step off and step back on again, the scale should read the same weight. I can repeat this process, and each time I should see the same weight. That consistency over measurements describes reliability. However, if each time I step on the scale it reads 20 pounds that is an invalid measure. I might consistently get the same 20 lb reading at each weighing (reliable), but not an accurate measure (valid).

Traub and Rowley (1991) also offer some excellent pragmatic examples of consistency or reliability such as a car that starts repeatedly or an employee who completes their work as expected. They do point out that reliable does not have to mean perfect. For example, a baseball player whose batting average is .350 is considered to be reliable even though they only get a hit 35 times in 100 at bats. There are many ways in which measures need to be stable: across time, over different forms of a test, across items, etc.



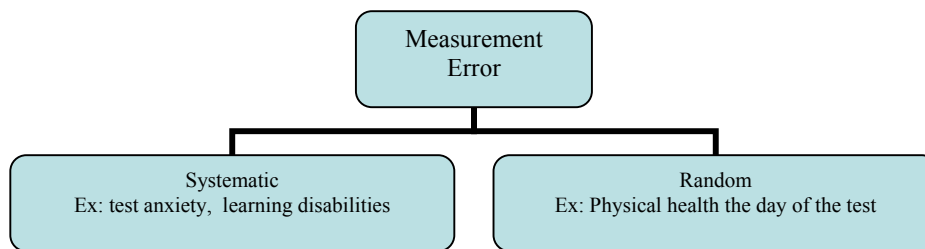
A more visual example of reliability and validity involves a bull's eye within a larger target.

To have a reliable shot one must hit the target in approximately the same place multiple times. However, reliability is not a sufficient condition for validity. For example, the target has been hit a number of times in the outer white ring. This represents reliability, but not validity.

Validity is represented by the multiple shots in the shaded center of the target. The shots are in approximately the same location in the appropriate area of the target. Validity is sufficient to imply reliability.

Consider another example Regarding the level of historical knowledge possessed by a student. This level is also referred to as true score, trait, achievement, universe score, or ability. It is not possible to measure the exact level of historical knowledge for a student. However, we attempt to obtain the best possible measure of that knowledge. Ideally, if we repeatedly measure the ability of this student we would get the same score. However, this ideal is not realistic. For example, an instrument may contain items that are not lucid causing the student confusion and therefore they answer incorrectly. Perhaps the student is not feeling well on one occasion and their physical condition prevents them from doing their best on the exam. Other factors including distraction in the classroom, construction outside the window, test anxiety, or lack of concentration can contribute to an inaccurate measure of the level of historical knowledge. These factors, among others, compose what is known as *measurement error*.

Measurement error can be divided into two categories: random and systematic. Random measurement error includes chance or unpredictable events that may occur during one administration of the test but not another (Friedenberg, 1995). Examples of random error include illness on the day of the test, distractions inside or outside the testing room, and an individual's motivation on the day of the test. Systematic measurement error includes consistent events that occur during a test administration that impact the test score yet are unrelated to the construct being measured (Crocker and Algina, 1986). Examples of systematic error include test anxiety and test-wiseness.



### DEFINING RELIABILITY

Once the concept of reliability is clearly understood, it is important to state a formal definition of reliability. The most appropriate place to find a definition of reliability is in the *Standards for Educational and Psychological Testing (Standards)*. The authors write, “Reliability refers to the consistency of such measurements [tests or scales] when the testing procedure is repeated on a population of individuals or groups” (AERA, APA, & NCME, 1999, p. 25)

“Reliability refers to the consistency of such measurements [tests or scales] when the testing procedure is repeated on a population of individuals or groups”

Crocker and Algina (1986) also provide a useful definition and description of reliability. “The desired consistency (or reproducibility) of test scores is called *reliability*” (p. 105). “Reliability refers to the consistency of examinees’ relative performances over repeated administrations of the same test or parallel forms of the test” (Crocker and Algina, 1986, p. 127). Technically, “the *reliability coefficient* can be mathematically defined as the ratio of true score variance to observed score variance” (Crocker and Algina, 1986, p.116).

### WHY DO WE NEED RELIABILITY?

Crocker and Algina (1986) remind us that in order for an instrument to be useful it must be reliable. That is, if a measure is used to make decisions and that measure is not reliable then it follows that the decisions were not based on sound information. Therefore those decisions are ill-informed and may not be sound. The *Standards* also addressed this issue and caution that reliability coefficients along with the method used to estimate them

should be reported. This charge directs test developers to examine the evidence of reliability in as much detail as possible.

Reliability is not an all or nothing prospect; meaning that we should not think of scores as having or not having reliability. Rather the level of reliability lies along a continuum from a small level of reliability to a large level (Traub and Rowley, 1991). The reliability coefficient provides information on how much measurement error there is and where the sources of error may be. Whenever possible we hope to minimize error so that the only differences we see in scores are the differences in true respondent ability.

It is important to clarify the terminology and correct use of reliability at this point. Tests aren't considered reliable; the scores they yield are considered reliable or unreliable (Traub and Rowley, 1991). Don't make the mistake of saying, "the reliability of the test..."!

### **HOW TO MEASURE RELIABILITY**

We report reliability in terms of variances or standard deviation (s.d.) of measurement errors (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 2000). It is also important to note that when we measure reliability coefficients we are actually calculating an estimate of the reliability because the exact reliability cannot be computed (Crocker & Algina, 1986).

There is no one "catch all" method for measuring reliability; instead one can consider the utility of each of the following four types of reliability coefficients:

- ✓ Coefficient of equivalence
- ✓ Stability coefficients
- ✓ Internal coefficients
- ✓ Generalizability coefficients

### **COEFFICIENT OF EQUIVALENCE (ALTERNATE-FORM RELIABILITY)**

- Alternate-form coefficients can be estimated when we administer two forms of the same test on one occasion.
- We try to construct two tests that are very similar in nature. If we could construct two tests that were exactly alike, then we would have parallel tests. In reality, parallel tests are not possible therefore we use alternate forms. Alternate forms should measure the same thing. We can compute a reliability coefficient for parallel tests, but an estimate of reliability for alternate forms of the test.

- The coefficient of equivalence can be estimated when we administer both forms of a test to the same group of people. To reduce fatigue, a short period of time can lapse between tests. However, the “break” should not be so long that the examinees would legitimately change in ability over that time. The two forms should be counterbalanced. The procedure of counterbalancing involves randomly assigning the order to test form administration to the test takers. Because not everyone receives the tests in the same order we minimize the effects of fatigue and practice. Counterbalancing allows for arranging the treatment effects in such a way that the effect of practice is minimized (Keppel, Saufley, & Tokunaga, 1992). The scores on the two forms should be correlated to get the coefficient of equivalence.
- A good rule of thumb for acceptable equivalence is a correlation of .8-.9. Ideally, the means, standard deviations, and standard errors of measurement will be close in value and should be reported as well.

The correlation between forms can be computed using your choice of software packages.

### **STABILITY COEFFICIENTS (TEST-RETEST RELIABILITY)**

- Stability coefficients are computed from administering one test on two occasions. To estimate the coefficient we administer the test and then administer the same test in a time period shortly after the original administration. We look at the correlation of the scores and determine the stability of that single instrument.
- A good rule of thumb for acceptable equivalence is a correlation of .8-.9 although slightly lower (.7) is acceptable.
- The question of how much time should lapse between administrations is often asked. That is, the meaning of “shortly after the original administration” often needs explanation. Unfortunately, there is no right or wrong answer. Ideally, allow enough time to minimize fatigue, especially if the instrument is lengthy. However, it is not appropriate to let too much time pass as respondents may change legitimately in their true skill or ability level over that time, thus impacting the reliability estimate (Crocker & Algina, 1986). Traub and Rowley (1991) suggest determining the amount of time between administrations on the test itself as the time needed to minimize memory and practice effects while reducing boredom and fatigue is different for each type of test. Despite our best efforts there may be a carry-over effect or influence of the first test administration on the second test administration (Allen & Yen, 1979)

## INTERNAL CONSISTENCY RELIABILITY

Internal consistency describes the relationship between items. The stronger the relationship between the items is, the stronger the internal consistency of the instrument. Also, if the scale is designed to measure a single construct then all items should have a strong relationship to that construct. If all items have a strong relationship to the construct then they should have a strong relationship with each other (DeVellis, 1991). Internal consistency coefficients are measured from a single administration of a test (Crocker and Algina, 1986).

Internal consistency coefficient = homogeneity of item content + quality of item

### Split-half method

To use the split-half method, divide the test in half to create 2 roughly parallel forms and find the correlation between halves. It is important when using the split half method to divide the test so the 2 halves are as parallel as possible. Some common ways to achieve this are to use odd-numbered items for one form and even-numbered items for the other, match content, random selection of items on both halves, rank items in order of difficulty, number the items and split odd- and even-numbered items onto the two forms.

One disadvantage of this method is that there is not one unique value for the reliability as the test can be divided into 2 halves in many different manners (Crocker & Algina, 1986). However, the solution to this problem is to consider using coefficient alpha as a way to estimate the internal consistency of a test. Additionally, the split-half method usually underestimates the reliability estimate for the test. To overcome this disadvantage one can use the Spearman Brown prophecy formula. For more specific information on this formula consult a measurement text such as Crocker and Algina, 1986.

### Coefficient alpha

One advantage of coefficient alpha is that it can be used with items that are dichotomously scored or have a range of scoring values. This method encompasses three different procedures, yet the same value is obtained through each procedure. The three procedures are Cronbach's alpha, Kuder-Richardson 20 (K-R 20), and Hoyt's analysis of variance. This type of reliability can be best understood in relation to split-half reliability. Several researchers conceptualize these forms of split-half reliability as the average of all possible split-halves one could create from a test. The formulae for each method shown below demonstrate the similarities and differences

between the methods. These methods of computing reliability are available in many statistical software packages.

Formula for Cronbach's alpha 
$$\hat{\alpha} = \frac{k}{k-1} \left( 1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_x^2} \right)$$

$\hat{\alpha}$  = reliability estimate  
 $k$  = number of items on test  
 $\hat{\sigma}_i^2$  = variance of item  $i$   
 $\hat{\sigma}_x^2$  = total test variance

The formula for KR-20 (for dichotomously scored items) is

$$KR_{20} = \left( \frac{k}{k-1} \right) \left( 1 - \frac{\sum pq}{\hat{\sigma}_x^2} \right)$$

$KR_{20}$  = reliability estimate  
 $k$  = number of items on test  
 $pq$  = variance of item  $i$   
 $\hat{\sigma}_x^2$  = total test variance

Hoyt's method is based on analysis of variance.

The formula for Hoyt's method is 
$$\hat{\rho}_{xx'} = \frac{MS_{persons} - MS_{residual}}{MS_{persons}}$$

$\hat{\rho}_{xx'}$  = reliability estimate  
 $MS_{persons}$  = mean square term for persons taken from the ANOVA summary table

$MS_{residual}$  = mean square term for persons taken from the ANOVA summary table

### WHEN TO USE EACH COEFFICIENT

It is important not to interchange the different reliability coefficients as generalizability coefficients provide different information than KR-20 or Cronbach's Alpha (AERA, APA, & NCME, 1999)

Because there are several different methods for estimating reliability, the question of which method to use is often asked. Two issues that should be taken under consideration are the characteristics of the instrument itself and practical issues of time, money, and resource availability (Henerson, Morris, & Fitz-Gibbon, 1987). Recall that reliability can be estimated through one or two test administrations. If you are not able to administer an instrument on multiple occasions then measures of internal consistency would be preferred. If you plan to administer 2 forms of a test then the alternate form estimates would be preferred. If your instrument contains a sufficient number of homogeneous items then the split half method may be preferable. Theoretically, the best procedure would yield an estimate that would be obtained if the forms were strictly parallel, however we know that is not possible (Crocker & Algina, 1986). Ultimately, the selection should be made based on how the scores will be used.

## IMPACTS ON RELIABILITY

Because reliability is an estimate there are several factors that impact that estimate. Being aware of these factors and the impact they have on our estimate will provide us with more information. Traub & Rowley (1991) remind us that the interaction between the test itself, the circumstances of the test administration, and the individual respondents combine to determine the reliability of the scores. These factors include the following components:

- Test characteristics
  - Test length – the longer the test the more reliable the score will be
  - Item type- objectively scored items contribute to more reliability
  - Item quality- lucid items contribute to more reliability as do items that discriminate
  - Item difficulty- a range of items from easy to difficult is needed to ensure discrimination among respondents of different abilities
- Circumstances of administration
  - Physical conditions-room configuration, temperature, noise level, lighting, etc. can impact reliability
  - Instructions- the lucidity of the instructions will impact the reliability as will the delivery of those instructions
  - Proctor- the test proctor may be a person who is motivating to students or is keenly aware of behaviors that impact reliability (e.g. cheating)
  - Time limit – on a speeded test some examinees will finish and some will not and therefore the reliability estimate may be artificially inflated on a speeded test
- Respondents
  - Homogeneity of group – a homogeneous group of respondents will have less variance among scores and thus a lower reliability estimate

## HOW TO INTERPRET YOUR RELIABILITY

Reliability is only one quality of a test. Recall that the group of respondents and the test administration itself can impact reliability (Traub & Rowley, 1991). It is important to calculate a reliability estimate for each test administration and investigate causes of consistent, low reliability.

The AERA, NCME, & APA *Standards* provide a great deal of information on reporting reliability evidence. Whenever a report is completed the reliability evidence should also be provided. It may be the case that practitioners will draft several reports for different audiences – an executive summary for the Dean or President and a complete report for the program director, but in any case reliability information should be easily obtained.

*For further reading on reliability:*

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Crocker, L.M. & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth: Harcourt Brace Jovanovich.

DeVellis, R. (1991). *Scale development: Theory and applications*. Newbury Park: Sage Publications.

Friedenberg, L. (1995). *Psychological testing: Design, analysis, and use*. Boston: Allyn & Bacon.

Henerson, M. E., Morris, L. L., & Fitz-Gibbon, C. T. (1987). *How to Measure Attitudes*. Newbury Park, CA: Sage Publications.

Keppel, G., Saufley, W. H., & Tokunaga, H. (1992). *Introduction to design & analysis: A student's handbook, 2<sup>nd</sup> edition*. New York: W. H. Freeman and Company.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications*. Thousand Oaks, CA: Sage Publications.

Traub, R.E. & Rowley, G. L. (1991). An NCME module on understanding reliability. *Instructional Topics in Educational Measurement Series*.

## VALIDITY: DEFINITION AND INTRODUCTION

Validity is a measurement topic that receives a great deal of attention and rightfully so. Validation is the key component in test development (Benson, 1998). The *Standards* define validity as, “the degree to which evidence and theory support the interpretation of the test scores entailed by proposed uses of tests” (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999; p. 9). Messick (1995) also provides us with a useful definition of validity “Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment” (p. 741).

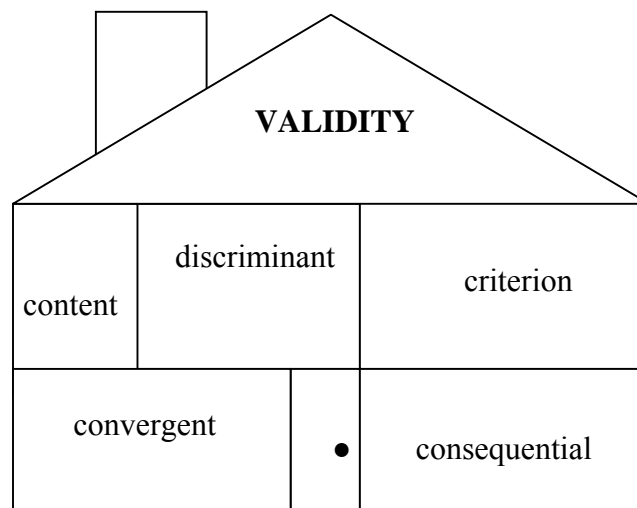
### IMPORTANT POINTS REGARDING VALIDITY

There are some general facts about validity that should be noted before moving on. Highly regarded professionals such as Messick (1995) and Benson (1998) have offered these thoughts on validity.

- ✓ Validity should be thought of as a unitary concept including different ways of gathering evidence that contribute to validity
- ✓ Validity is the meaning of the test scores. It is not a property of the test.
- ✓ Validation is a continuous process not completion of a finite number of discrete tasks
- ✓ Test scores should be accompanied by supporting validity evidence for each intended use

### Conceptual Example of Validity

It may be easy to conceptualize validity by comparing it to a house. A house is one unit, yet it includes different areas such as a bedroom, kitchen, living room, and bathroom all under one roof. Each room serves a purpose, yet by itself does not make a house. Similarly, validity is a unitary concept that is based on different kinds of evidence. Different aspects of validity evidence each serve a purpose, but together under one roof they form a validity argument.



We know that once a house is built the work is not finished. It seems that there is always a home improvement project waiting for the owner. These projects may include a routine task such as mowing the grass or larger projects such as painting, adding wallpaper or even major renovations or additions to the home. Likewise, once an instrument is designed, gathering validity evidence is not complete. Validity, like home improvement, is never really complete. Routine evaluations of test use are important to ensure that the instrument is still serving the purpose for your audience. If an instrument is going to be used for a different purpose or given to a new audience, then that requires major validation work. Certainly, a homeowner can fail to maintain their house with the result being an undesirable environment. A test user who fails to maintain validity evidence for their test uses will have an instrument that is not desirable for their purposes.

### **CONDUCTING A VALIDITY STUDY**

A validation process is both a science and an art because it requires both statistical evidence and rhetorical arguments (Messick, 1995.) It is important to approach a validity study with the right frame of mind. There is no “one size fits all” approach and the test user will need to consider validation as an ongoing project rather than a finite task. Validity evidence should be gathered during the instrument design phase as well as the pilot and implementation phases (Crooks, Kane, & Cohen, 1996).

#### **How to think of validity study**

Level or degree of information

Contributing to the body of validity evidence

#### **How NOT to think of validity study**

Checklist

One piece of evidence is sufficient

All or nothing

A validity study begins with understanding how the test scores are to be interpreted. That is, what the test purports to measure. It is important to have a framework by which we determine what the test is measuring (Crocker & Algina, 1986). It is not correct to consider “types” of validity as validity is a unitary concept, but there are different components or aspects to validity (just like a house has different types of rooms). Your goal is to obtain as much evidence as possible in as many aspects as possible.

#### **4 aspects of validity:**

- content
- construct
  - substantive
  - structural: structure of content domain and scoring structure
  - external: divergent and convergent evidence, criterion
- generalizability: scores generalize across domains, settings, groups
- consequential: value implications – intended and unintended consequences

#### **VALIDITY EVIDENCE BASED ON THE CONTENT ASPECT**

To gather content validity evidence one must analyze the relationship between the content of the instrument and the domain or objectives it is purported to measure. For example, if the domain of interest is algebra skills and the instrument contains a number of items related to geometry then the content validity evidence is weak.

In order to establish strong content validity it is important to have a well-defined domain of interest, employ content experts to match test items to the domain of interest, and analyze the results of that matching process (Crocker & Algina, 1986).

Other considerations to keep in mind when examining content validity evidence are content relevance, content representation, and technical quality of the content. Content relevance refers to whether the items are necessary to adequately represent the content the instrument is designed to measure. Content representation refers to whether the items on the instrument adequately represent the content the instrument is designed to measure and technical quality of the content refers to how the instrument content compares to recognized standards in the field as well as audience appropriate content. For example, if we want to measure if a student knows the factors leading to the US involvement in World War II it is not relevant that the student know the factors that lead the US involvement in World War I. Additionally, if we were to measure the statistical skills of graduate students studying educational measurement we might ask them to conduct a repeated measures ANOVA as opposed to asking them to compute the mean of a set of values.

#### **VALIDITY EVIDENCE BASED ON THE CONSTRUCT ASPECT**

Messick (1995) warns of two threats to validity – construct under representation and construct irrelevance. Construct under representation occurs when an instrument does not appropriately cover the domain of interest or the construct it is intended to measure. That is, test users would like to make conclusions or generalizations, but are unable to do so because that area of interest was not represented on the test. Construct irrelevance occurs

when an instrument taps into skills or knowledge beyond what it is intended to measure. That is, the instrument measures extraneous domains that it was not intended to measure. Construct irrelevance can cause scores to be invalidly low if the extraneous “noise” causes the test to be more difficult or it can cause scores to be invalidly high if the extraneous “noise” causes the test to be easier (Messick, 1995).

Constructs that we hope to measure are based on theory and the relationship between the construct of interest and other constructs forms what is known as a nomological net (Cronbach & Meehl, 1955). The relationships between the constructs offer possibilities for rich information including the formation and testing of rival hypotheses. Benson (1998) describes the process of obtaining construct validity evidence as iterative. That is, an instrument is purported to measure a particular construct and it is compared to other instruments that purport to measure the same thing. Additionally, it is compared to rival hypotheses. Based on the evidence found the theory may be refined and the process begins again.

### **Validity evidence based on the external aspects**

Discriminant validity is the relationship between your instrument and an instrument that measures different constructs (Angoff, 1988). You want the two instruments to be different because they are purporting to measure different things therefore a strong negative correlation or no correlation is ideal.

Convergent validity is the relationship between your instrument and an instrument that measures the same constructs (Angoff, 1988). You want the two instruments to be similar because they are purporting to measure the same things therefore a strong positive correlation between the scores is ideal.

### **VALIDITY EVIDENCE BASED ON CONSEQUENCES**

Consequential validity is a term first introduced by Messick. Consequential validity includes the intended and unintended and predictable and unpredictable consequences of the test use. For example, if we create a test that is used for making decisions about graduation status, it is important to understand the consequences associated with the decisions made using the test scores. Some consequences of a test are predictable and some are not predictable only to come to our attention after the test has been administered. For example, a university history department is engaged in assessment of its general education course offerings and requires that all sophomores participate in testing. The intended consequence of the assessment is to provide formative program evaluation information. An unpredictable consequence may be that a student who has yet to declare a major is fascinated with the items on the test and realizes the history is her passion and she declares history as her major. We would be remiss as test developers and test users if we did not evaluate the effects of our tests following their administration (Shepard, 1997).

### **WHO IS RESPONSIBLE FOR VALIDATION?**

According to the *Standards*, both the test user and the test developer are responsible for providing validity evidence. The test developer should provide information about their

idea of the intended use when they produce the test manual. In addition, the test developer should continually update their manual to reflect new validity evidence and empirical studies involving the instrument. The test user should corroborate that evidence for their intended population and if they choose to use the test for a different purpose, then they must collect validity evidence for that purpose.

To read more about validity:

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W.H. (1988). Validity: An evolving concept. In Wainer, H. & Braun, H.I. (Eds.) *Test Validity*. Hillsdale, NJ: Erlbaum.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice* 17. 10-17.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. San Francisco: Harcourt.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* (52) 4. 281-302.

Crooks, T.J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education* 3(3). 265-285.

Messick, S. J. (1995). Validity of Psychological Assessment. *American Psychologist*. 741-749.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issue and Practice*. 5-24.

## ITEM ANALYSIS

When examining the results of a test it is important to conduct an analysis at the item level as well. An item analysis includes computing the item difficulty, examining the items for possible bias, modifying the items, and re-piloting items.

### ITEM DIFFICULTY

The item difficulty is the proportion of examinees who answered the item correctly (Crocker & Algina, 1986). Because item difficulty is a proportion this parameter can take on values from 0 to 1. If one were analyzing test data in a statistical software package, the item responses could be recoded as 1 for a correct response and 0 for an incorrect response. The mean of each item would then represent the item difficulty as shown below.

$$\text{Item difficulty} = \frac{\text{\# of examinees who answered the item correctly}}{\text{\# of examinees who responded to the item}}$$

An item difficulty that is too high or too low can be a cause for concern. An item difficulty that is too high might indicate that the item was too easy; if all examinees are able to answer the item correctly it might not add any important information to the test users. An item difficulty that is too low would indicate an item that is extremely difficult. In this case it is important to examine the item to determine if the item is difficult because of the item content, not an extraneous factor that is influencing performance.

### ITEM DISCRIMINATION

The item discrimination is the amount that an item distinguishes between respondents of different abilities (Crocker & Algina, 1986). It is desirable to have items that discriminate, as we want students to demonstrate their ability and distinguish themselves from other students. A commonly used measure of item discrimination is the point biserial correlation. These correlations are provided through statistical software packages as a part of the reliability estimation procedures.

This correlation represents the relationship between how students perform on the item and how they perform on the entire test. Ideally, if a test is only supposed to measure one construct (e.g., history knowledge) the items should be related to one another. Typically, items that point biserial correlations of less than .3 are a cause for concern. Again, it is important to review these items to determine why they are different from the remaining part of the test.

Item response theory (IRT) also provides methods for computing the item difficulty and item discrimination values.

### **ITEM BIAS**

Item bias occurs when an item behaves differently for different groups of test takers. Item bias is not necessarily a negative feature of an item. There may be certain circumstances where we expect certain groups to score differently on an item. For example, a theory of parenting styles purports that first time parents discipline their child more harshly than experienced parents. Therefore, we expect first time parents and parents with at least one child to respond differently to questions about child discipline. However, in other situations bias creates an unfair advantage for some test takers. Crocker & Algina (1986) define an unbiased item as one that is affected by the same sources of variance for both groups and examinees of the same ability level in different groups will have the same distribution of irrelevant sources of variance. One of the methods of investigating item bias is through the use of IRT. A comparison of the item characteristic curves for the two groups will show if the item behaves differently for each group. However, if bias is shown it is up to the test developer to try and determine if there is a reason for that finding (Crocker & Algina, 1986).

### **MODIFYING AND RE-PILOTING ITEMS**

Once a complete item analysis is conducted the modification process can begin. If an item is found to be suspect you may try to examine it for common pitfalls such as lack of clarity, relationship to other items (the answer may be contained in another item), or appropriateness of the item stem and responses. It may be helpful to consult content experts and/or test takers for help when modifying the items. Once an item is modified it should be re-piloted and re-examined before it is used on the final instrument.

For more information on item analysis consult:

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Crocker, L. & Algina, J. (1986). *An introduction to classical and modern test theory*. Orlando, FL: Harcourt, Brace, Jovanovich

Hambleton, R., Swaminathan, H., Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

## ETHICAL ISSUES IN TESTING

Testing professionals have expectations of one another to uphold high standards while completing their jobs in the best manner possible. Test developers and test users should become accustomed to providing or requesting information that demonstrates how the test was constructed within the ethical guidelines. AERA, APA, & NCME (1999) address professional ethics in the *Standards*. These guidelines apply to any type of test, instrument, survey, or scale. *The Program Evaluation Standards 2<sup>nd</sup> edition* (1994) also provide practitioners a resource for conducting themselves in an appropriate manner. Additionally, survey researchers follow the *Code of Professional Ethics and Practices* which is published by the American Association of Public Opinion Research (McNamara, 1999). Professionals involved with computer based testing also have ethical guidelines outlined in *Guidelines for computer-based tests and interpretations* (American Psychological Association, 1986) and *Guidelines for computerized-adaptive test development and use in education* (American Council on Education Task Force, 1995).

Every individual makes decisions about how they conduct themselves professionally. However, there are very few people who would knowingly work with an individual whose behavior is unethical. It is important to follow the ethical guidelines and seek consultation for situations that call for actions that are questionable.

### COMMON ETHICAL PRACTICES

This list includes practices suggested from McNamara (1999):

- ✓ Inform all individuals in the instrument development process of any affiliations or interests that could pose a conflict. Remove yourself from any situation that could pose a conflict of interest.
- ✓ Be honest and forthright through all steps of the instrument development process.
- ✓ Follow all governing human subjects protection policies.
- ✓ Fully inform all participants about the research you are conducting.
- ✓ Maintain confidentiality when promised.
- ✓ Do not ask unnecessary questions.
- ✓ Ask personally sensitive questions only when it is dictated by the research question.

To read more about ethics in testing consult the following:

American Council on Education Task Force. (1995). *Guidelines for computerized-adaptive test development and use in education*. Washington, DC: American Council on Education.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Hyattsville, MD: Committee on professional standards and Committee on psychological tests and assessment.

McNamara, J. F. (1999). Ethical guidelines in survey research. *International Journal of Educational Reform* 2(1). 96-100.

Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards 2<sup>nd</sup> edition*. Thousand Oaks, CA: Sage.

## REPORTING

In the first worksheet of this manual, you defined your reason for creating your test. Now that it is finally time to use your test, it is important that you keep that initial purpose in mind. It is easy to administer a test and collect data from that administration. However, it is very important to remember to use the test only for the purposes for which they have been designed. There are two basic ways scores can be reported from a test: based on a norm or based on a criterion.

### NORM-REFERENCED TEST

A norm referenced test uses a method of score interpretation where test performance is judged in relation to the performance of others. The “others” could be the rest of the students who took the test at the same time, or a group from a different administration. Usually when test are norm referenced, the norming score came from a careful administration of the instrument to a group representative of the population. Crocker and Algina (1986) suggest the following steps for conducting a norming study as follows:

- ✓ Identify the population of interest (this is your target population)
- ✓ Identify the most critical statistics that will be computed from the sample data (mean, standard dev)
- ✓ Decide on the tolerable amount of sampling error (how large of a difference will you allow between your sample and population estimates of the parameter of interest)
- ✓ Devise a procedure for drawing a sample from the population of interest (sampling method – use the best available that follows the test administration procedure as set forth in the first worksheet)
- ✓ Estimate the minimum sample size required to hold the sampling error within the specified limits (several different methods are available for this – see Crocker & Algina, 1986).
- ✓ Draw the sample and collect the data
- ✓ Compute the values of the group statistics of interest and their standard errors (standard error - how confident can you be that your calculated statistics represent the population)
- ✓ Identify the types of normative scores that will be needed and prepare the normative score conversion tables (stanines, z-scores, percentile rank, etc.)
- ✓ Prepare written documentation of the norming procedure and guidelines for interpretation of the normative scores (this should be very detailed)

Following these steps will ensure that your norming data will be representative and reliable.

### CRITERION-REFERENCED TESTS

The second type of scores that can be used are criterion based. A criterion based test interprets scores based on an understanding of what each level of score means in terms of the construct being assessed. This is usually accomplished by setting one or more cut scores that represent different levels of proficiency. Setting cut or standard scores is a

very complicated process when the proper measures are taken to establish a valid distinction. Crocker and Algina (1986) suggest three different approaches to setting a cut-score standard:

1. Judgments based on holistic impression of the examination or item pool  
Individuals (who know what type of knowledge the target population should possess) examine the item pool and determine what percentage of items examinees should be able to pass based on the expected level of knowledge
2. Judgments based on the content of the individual test items  
In this method, judges examine each item and determine the likelihood that an examinee (who barely meets the competency) will get this item correct?
3. Judgments based on examinee's test performance  
This method has been used several different ways; however, the most common would be to administer the test to a group of students who should demonstrate the competence and a group who shouldn't (maybe they haven't taken the class yet) and compare the score distributions. Graphically, if histograms were used to plot the distributions of the two groups scores, the cut scores would be the middle of the overlap between the two groups.

The choice of score interpretation method should be determined by the test developer and user; however, one method is likely to be more appropriate than the other based on the purpose of the test. The important thing to remember is that the proper methodology should be followed to ensure accurate interpretation of the results.

For further reading on score interpretation, see:

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando: Harcourt Brace Jovanovich.

Friedenberg, L. (1995). *Psychological Testing*. Needham Heights: Allyn & Bacon.

## OTHER TYPES OF TESTS

While this manual has focused on creating multiple choice or selected response item tests, we did want to mention other types of instruments that you might want to develop and to briefly review some of the differences in the development of these types of instruments

### AFFECTIVE SCALES

Affective scales refer to instruments designed to measure attitudes, emotions, perceptions, personality traits, or opinions. These types of measures are commonly seen in survey research and psychological testing. The development of such measures is similar in many ways to the process described in this manual; however there are some differences. Item writing will no longer be based on learning objectives; rather, it will be based on the components of the construct as defined by the theoretical literature. Secondly, the review of items will obviously not focus on the scores and how they relate to other measures of ability, rather they will be assessed in terms of their relation to other constructs. As far as internal validation of the test, a technique called Structural Equation Modeling can be utilized to assess the structure of the items. This will help determine how the items are functioning together as a whole instrument. These are just a couple of the major differences in the process of development, for a further review of the process of creating affective scales see:

DeVellis, R. (1991). *Scale development: Theory and applications*. Applied Social Research Methods Series, Vol. 26. Newbury Park: SAGE publications.

Spector, P. (1992). *Summated rating scale construction*. Newbury Park: SAGE Publications.

### PERFORMANCE MEASURES

It is very common now to see performance or alternative assessment methods being used to measure student ability. The debate over which types of abilities can or cannot be measured by selected response items is a continuing conversation. Development of performance assessments can be somewhat different, however, the process is still driven by a thorough understanding of the underlying construct and what types of indicators will assess knowledge of that construct. Review of items such as these could include using Generalizability Theory to assess the reliability of such a measure. For further reading on the topic of performance assessment development, see:

Stiggins, R.J. (1987). Design and development of performance assessment. *Educational Measurement: Issues and Practice*, 4, 263-273.

## FINAL NOTE

Item writing is a difficult process; make no mistake about that. However, the benefits that can come from the satisfaction of completing this process and the product that will be created are worth the effort. We would offer some bits of advice for surviving this process that we have learned by being a part of this process.

- ✓ Make sure that there is both a content expert and a measurement professional involved in every stage of this process; these two roles can be played by the same person if it is appropriate.
- ✓ This process always seems to take longer than you expected; don't worry when you can always meet your pre-set timetable.
- ✓ Make yourself a schedule for each step and try to stick to it.
- ✓ Hold yourself accountable for keeping the initiative going; if you are working in a group, set goals that must be met in a timely fashion.
- ✓ Try to get feedback from as many experts (content and measurement) as possible during this entire process; take this feedback with an open mind and consider it constructive criticism
- ✓ Document your process at every step; keep accurate records of any change modifications, etc. that that you make to the instrument as it is a work in progress.
- ✓ When this process is completed, create a manual from all of your work including analyses, history of development, and research involving your instrument. This is helpful even if you don't have immediate plans for distributing your instrument.
- ✓ Finally, realize that this process is never complete; instrument development should be a continual effort as you use and learn more about your instrument.

We hope that using this manual and following the steps as they are laid out here will help make this an easier process. Again, the benefits that can result from creating your own instrument will greatly outweigh any effort or difficulty that is put into the process.