Psychometrics 101: Test Development

Barbara Foster, Ph.D.
UT Southwestern Medical Center
Dallas, TX

ABSTRACT

For nearly a century, the development and use of tests were more or less restricted to the educational and psychological arenas. During the 70s and 80s a groundswell of testing was developed and used in the Human Resources arena, to help match positions with candidates.

Particularly in the last ten years, there has been a dramatic increase in the use of tests within the medical field. Quality of Life Measures and Pain Indices are two commonly known ones. Measures to determine the amount of exercise and food ingestion over time are constantly being developed and upgraded.

In recent years, research studies have used various tests as part of their research protocol. In some cases, the researchers have little if any knowledge of the science of test development.

Experts within the field of test development are called Psychometricians. The name reflects the roots from the fields of Education and Psychology. Years of study are required to obtain a Ph.D. in psychometrics and the body of knowledge is large, and growing constantly.

The purpose of this presentation is to acquaint SAS users with key issues within the field. With the expansion of tests into almost every area of life, an overview of the process may be helpful.

INTRODUCTION TO NUMBERS

The concept of measurement has been thoroughly debated and to some extent is still debated. For example, we can count the number of windows in a building. There is no measurement involved in determining the number of windows. It could be 3, 10, or 50 but it cannot be 3.5 or 10.78 or 50.222. However, if we want to determine the width of a window then we enter the realm of measurement. There must be an agreed upon unit—in this case a ruler or the equivalent. The ruler or measuring instrument must not change over time. For example, a spring can be marked into increments of distance and used as a measuring device. Unfortunately, the distance between the marks may change as the spring is used. A length of elastic can certainly be marked to show inches or even feet, yet elastic stretches so that the increments might not be the same from one reading to another.

Different types of measurement exist. The word "measure" brings to mind a situation such as time, weight, speed, and other continuous types of number with a zero point and going to some large number, sometimes infinity. More importantly, any specified interval size will be the same quantity regardless of where the interval falls, i.e. the pound difference between 25 and 30 on the scales is the same pound difference between 125 and 130. These numbers are known as RATIO type numbers.

What type of number is temperature? On the Celsius and Fahrenheit temperature scales there are no "zero" points. Yet the amount of difference between 55° and 65° is the same amount of difference as between 80° and 85°. These numbers are known as INTERVAL data.

Some times all the information the measure provides is order. The words "Small", "Medium", and "Large" are this type of measure. The difference between small

and medium may be the same as the difference between medium and large, but there is no guarantee of that. These numbers are known as ORDINAL data.

And there is a lot of information that is not even rank ordered. Despite what some think there is no inherent order to gender or race. Colors represent another example. We could code blue as 1, green as 2, red as 3, and black as 4. But we could also code blue as 17, green as 10, red as 8, and black as 15. The numbers have no meaning in and of themselves, but are used only as labels. These numbers are known as NOMINAL data.

Numbers are assigned to *aspects* of objects, not to the objects themselves[1]. One can measure the length, volume, or color of a box but not the box itself. One can measure the weight, gender, eye color, level of depression, or level of mathematical ability of a person but not the person himself or herself. When a given aspect is studied, all other aspects are ignored.

Within the field of Psychology or Educational Psychology, measurement issues become more difficult. Because of the complexity of humans, "[w]hat is relevant to a measure can be determined only within an implicit or explicit theory about the phenomenon one wishes to study." What a number "means" depends upon the theory from which it is derived. For this reason, users of psychological tests should be well trained in the theory and use of those instruments.

Different types of tests are developed for different purposes. Everything we measure uses some type of measuring instrument and collects some type of number or label. But the focus of this presentation is on psychological tests and dichotomously scored (true/false) licensing or certification exams.

---

[1] Pedhazur, Elazar J. & Liora Pedhazur Schmelkin. 1991. *Measurements, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Publishers, Hillsdale, NJ. p. 16.

OVERVIEW

The following stages in test development are not exclusive and most of these stages could be broken into smaller units[2] for more stages or merged into fewer stages.

Stage I	Defining the test
Stage II	Preparing and formatting items
Stage III	Obtaining data
Stage IV	Item checks
Stage V	Reliability determination
Stage VI	Validity checks
Stage VII	Dissemination of results

Stage I of preparing a test consists of many key steps. Drs. Shultz and Whitney[3] outline the stage in seven steps.

Step 1: Specify type of measure
Step 2: Define domain and intended context of test.
Step 3: Type of questions to be included.
Step 4: Item format.
Step 5: Test to be administered to individuals or in a group setting
Step 6: Determine the appropriate length of the test
Step 7: Determine the appropriate difficulty or level of the items.

And all of that is before question writing begins! Once items are written and the test is prepared then there is the burden to show that the items are appropriate and have good test qualities.

Stage II consists of preparing items for the test. While every novice thinks this is the easiest part of the process, they soon learn differently when they are involved in the process. For high-stakes tests, the cost of obtaining a single good item could easily be $1500 or more! There is good reason for testing entities to be concerned about the integrity of their items.

[2] Althouse, Linda A. "Ten Steps to a Valid and Reliable Certification Exam". SUGI 25, SAS Institute, Inc. Carey, NC. Paper 244-25

[3] Shultz, Kenneth S. and David J. Whitney. 2005. Measurement Theory in Action: Case Studies and Exercises. Sage Publications, Thousand Oaks, CA. pp. 51-56.

Stage III requires that the completed test be given to a sizeable number of appropriate individuals. Other information on the test-takers may be collected at this time. In fact, other tests may be given at the same session. The results of these sessions are entered into a database.

Stage IV uses the data from the test administrations to study the test items. Poor items are either removed from the test or reworded.

Stage V processes the data to determine the internal consistency and/or reliability of the test. In other words, how reproducible are the results from the test likely to be?

Stage VI uses any additional information for validity checks.

Stage VII usually means to publish the results in an appropriate journal.

This presentation focuses on Stages IV and V.

ITEM ANALYSES

Items must first be submitted to subject-matter experts for approval. Individuals across the domain should also check the questions for phraseology, interpretation, and multiple content. Once approved, the questions are assembled into a test format and given to individuals not previously exposed to the items. Analyses are performed on the results of those administrations.

The most common analyses performed are:

- Item difficulty
- Item discrimination
- Distractor information (if applicable)
- Item-total correlation—ideally the total without the item.
- Item-Item correlations
- Standard deviations
- Factor Analysis to determine if the test is one-dimensional or has subscales.
- Number of items

Item difficulty is usually obtained for educational tests and is simply the percent of examinees that answer the item correctly. When the correct answer is scored 1 and an incorrect answer is scored 0 then item difficulty is easily obtained from SAS by either of these two procedures:

```
PROC MEANS sum; VAR first_item-last_item; run;
PROC FREQ; TABLES (first_item – last_item)/list missing; run;
```

If **PROC MEANS** is used then the sums must be divided by the number of items.

Item discrimination is an indication of how well the value on the item separates the test takers along the range of the construct. Different approaches are used. The most common, especially for dichotomous data is described next. The examinees are divided so that high scorers form a group and low scorers form a group. The proportion of low-scorers answering the item correctly is subtracted from the proportion of high-scorers answering the item correctly and this difference is called the item discrimination index.

Alternatively, the data may be divided into several groups covering the range of the total scores. The proportion answering the question correctly is determined within each group and graphed. The pattern of the graph indicates how well an item is discriminating across the range of ability.

Item-total correlations and item-item correlations are easily obtained from

```
PROC CORR nosimple; VAR total_score first_item – last_item;
```

Theoretically special correlation coefficients should be used for dichotomous items. Within the SAS macro library is a macro for calculating the biserial correlation--%macro biserial. However, studies have shown that the value obtained by the biserial correlation is the same as that obtained by doing a Pearson correlation, so little is gained by the extra

effort. The same is true for the phi correlation, used for correlating two dichotomous variables.

If there are a small number of items then the total_score should be calculated without the item being correlated with it. A simple macro performing this operation is shown in Appendix A. Experience has shown that with 100 items there is about a .01 difference between the correlation with the item (in the total score) and the correlation without the item. The difference increases as the sample size decreases.

There are item statistics and there are test statistics. The above discussion focused on ITEM statistics. The following TEST statistics are usually obtained:

- # Examinees
- Mean of the total scores
- Variance of the total scores
- SD of the total scores
- Skewness of the total scores
- Kurtosis of the total scores
- Min & max of the total scores
- Standard Error of Measurement
- Mean item-total
- Mean biserial correlation
- Max score (low group)
- N of low group
- Min score (high group)
- N of high group

The first seven statistics are easily obtained from:

```
PROC MEANS n mean stdev var sk kur min max; VAR total_score;
```

The standard error of measurement (SEM) is a measure of the amount of error in the scores and will be discussed further in the next section.

The last four numbers are used with Item discrimination to provide information on the groups used. The max score and N of each group can be obtained with code similar to:

PROC MEANS n min max; VAR high_scorers low_scorers;

Many of the above statistics are provided in a macro available from the SAS macro library called %Macro Item found at

http://support.sas.com/ctx/samples/index.jsp?sid=478.

RELIABILITY

Reliability is measurement precision and/or reproducibility. As opposed to the usual statistical situation, the reliability of a test is directly related to the number of items on the test and **not** to the number of subjects[4]. Many methods of determining a reliability coefficient have been developed over the last 75 years, based on different theories and different approaches attend to different sources of error[5]. One such set of errors are detailed by Shultz & Whitney[6].

1. Changes in examinees across time, known as stability.

A construct like intelligence is not expected to change a lot over reasonable time windows. An intelligence test should produce a result in October that is close to a result obtained in December from the same test. Stability is best checked by doing test/retest correlations.

Some constructs like Depression may be unstable and change from one time point to another. Stability must be checked by very narrow time windows or by using two or more forms designed to be equivalent to one another. Such equivalent tests are known as parallel forms.

---

[4] Nunnally, Jum C. and Ira H. Bernstein. 1994. *Psychometric Theory*, 3rd Ed. McGraw-Hill Series in Psychology, New York, NY.

[5] Pedhazur, Elazar J. & Liora Pedhazur Schmelkin. 1991. *Measurement, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Publishers, Hillsdale, NJ. p. 83.

[6] Shultz, Kenneth S. and David J. Whitney. 2005. Measurement Theory in Action: Case Studies and Exercises. Sage Publications, Thousand Oaks, CA. pp. 71-72.

2. Content sampling is consistency of items across test versions or within a given test.

Often several forms of a test will be developed and administered. There must be evidence that these forms are truly parallel and cover the same content domain at the same difficulty level. This is known as equivalence and involves giving more than one form to an individual and correlating the results on the two tests.

Another way to check content domain is to determine internal consistency of the test. Cronbach's alpha is used for this and can be accessed through **PROC CORR**. Note that internal consistency is not the same as unidimensionality, which will not be addressed in this presentation.

3. Inter-rater Consistency.

Particularly in psychological tests, a clinician asks the questions, notates the response, and assigns a number to the response depending upon perceived degree of difficulty. A good test must either be designed so there is little subjectivity in the ratings or adequate training must be provided so all raters are defining levels of difficulty in the same way. Obviously the inter-rater consistency must be checked. Cohen's kappa is usually used for checking inter-rater reliability and this statistic can be obtained in SAS from **PROC FREQ**. For multiple raters the Intraclass Correlation (ICC) is often used. A SAS macro for the ICC is also available in the SAS macro library.

Added to the above errors are just plain random errors. Many things can impact the results of a test. Perhaps at the test administration an examinee was feeling poorly or perhaps the barometric pressure was dropping. A repeat of the test would probably yield similar but not identical results. The standard error of measurement (SEM) addresses the degree of error expected within a test administration. The SEM also allows a confidence

interval to be placed around points of interest such as a cut-score. Appendix B provides the formula and code for this determination.

VALIDITY

Validity addresses the question, "Does the test measure what it purports to measure?" Nunnally & Bernstein[7] present 3 major meanings of Validity: "(1) *construct* validity—measuring psychological attributes, (2) *predictive* validity—establishing a statistical relationship with a particular criterion, and (3) *content* validity—sampling from a pool of required content." One can argue that all three meanings are part of *construct* validity.

Various approaches have been developed to estimate the various meanings of validity. One approach to construct validity is to compare groups known to be different on the construct. If the differences in the test results are consistent with the hypothesized differences then evidence exists for the validity of the test. Standard statistical tests are usually used for calculating differences across the groups.

An approach to predictive validity involves collecting information on the outcome and correlating the prediction with the actuality. Unfortunately in psychological tests and in licensing/certification exams there is either little external information available, or if available, is prohibitively expensive to collect.

The process for evaluating content validity involves comparing the new test to results from other tests. The correlations will show the similarities and dissimilarities of the new test to other tests. This is highly useful for psychological tests. A test developed

[7] Nunnally, Jum C. and Ira H. Bernstein. 1994. *Psychometric Theory*, 3rd Ed. McGraw-Hill Series in Psychology, New York, NY. p. 83.

for math anxiety must show that it is not a general anxiety test or that it is not capturing adolescent angst, neither of which have anything to do with math.

This presentation does not focus on Validity, which relies mainly on correlations and requires more thought and planning than computation.

OTHER ISSUES

Particularly in licensing/certification, new tests are developed for each testing cycle. There are reasons for this that we will not go into now. But such a situation requires that tests be fair—not only across the students taking the test at the moment but for students across the years. Even though the domain is well specified, the items used can be differentially difficult. For example, a test of mathematical ability may ask students to multiply 9*2. Next year's test may ask students to multiple 9*8. Are these two questions at the same difficulty level? Probably not. Standard equating practices have been developed to compensate for one test being easier or harder than another but these are beyond the scope of this presentation.

CONCLUSION

Because of what is involved in developing a good test, ANY change in a test negates the relationship to the original test until studies of the relationship are performed. Changes include but are not limited to translating the test into another language, dropping items, rewording items for a special population or to simplify the test, giving a test in a group setting instead of the originally specified standard individual format, or having the patient fill out the form rather than a clinician. When any changes are made to the test then the person or group using the test has the responsibility to redo a psychometric study

on the “revised” test. Even using the test with a different population requires that the reliability and validity be checked for use in the new setting and appropriately reported.

Test development is not an easy process, especially in our litigious society. As the numbers of tests increase and the areas in which tests are used increase, programmers may be involved in one or more of the stages of test development during their career. Having an overview of the process may help in the performance of those duties.

CONTACT INFORMATION:

Your comments and questions are valued and encouraged.
Contact the author at:

Barbara Foster, Ph.D.
UT Southwestern Medical Center at Dallas
Department of Family and Community Medicine
5323 Harry Hines Blvd.
Dallas, TX 75390-9067
barbara.foster@utsouthwestern.edu

APPENDIX A

```
%macro corrs (m);
                        proc sort data=dataset; by ID; run;

data total; set dataset; by ID;
if first.ID then total=0;

array q(30) item1-item30;  *in this example there are 30 items in the test;

do i=1 to 30;
      if i ne &m then total_score=total_score+q(i);
end;
run;

proc corr nosimple; var total_score item&m; run;
%mend corrs; run;
```

APPENDIX B

Standard Error of Measurement = $S_x\sqrt{1-r_{xx}}$, where S_x is the standard deviation of the total score and r_{xx} is the reliability estimate. A 95% confidence interval is created by adding to and subtracting from the score of interest the SEM multiplied by 1.96.

```
%macro SEM (score, s, r);

data sems;
     SEM_&score=&s*(sqrt(1 - &r));
     Upper_CI_&score= &score + (SEM_&score*1.96);          *upper CI;
     Lower_CI_&score= &score - (SEM_&score*1.96);          *lower CI;
run;

proc print data=sems;  run;

%mend SEM;
```