DOCUMENT RESUME

ED 359 203 TM 019 829

AUTHOR Boone, William; Svec, Michael
TITLE Objective Assessment of Preservice Teachers.
PUB DATE Jan 93
NOTE 11p.; Paper presented at the Annual Meeting of the AETS (January 29, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Measurement Techniques; Models; *Objective Tests; Preservice Teacher Education; *Research Methodology; *Social Science Research; Student Evaluation; *Test Construction; Test Results; Test Use; Theories
IDENTIFIERS *Preservice Teachers; *Stochastic Analysis

ABSTRACT

Characteristics of objective measurement and how to develop an objective instrument are discussed. Objective measurement can be thought of as that type of measurement in the social sciences that parallels the measurement that takes place in science. The following are aspects of scientific measurement that should be transferred to the measurement of individuals: (1) evaluating only one variable; (2) building the instrument around a theory; (3) continuous calibration of the measurement; (4) reporting of measurement errors; and (5) the ability to work equally well in a number of situations. All of these points must be considered in designing a measurement instrument. It is expected that everyone will use the instrument in the same manner. Unexpected responses must be used to improve measures, and for this reason, errors must be accurately reported. Other results from objective measurement using a stochastic model are discussed, including the implications of non-linearity. (SLD)

Objective Assessment of Preservice Teachers

Dr. William Boone and Michael Svec

Program in Science Education
School of Education 3068
Indiana University
Bloomington, IN 47405
(812) 856-8132

Paper Presented at AETS Annual Meeting

January 29, 1993

Introduction:

What does one mean by "objective measurement"? Briefly, objective measurement can be thought of as that type of measurement in the social sciences which parallels the measurement that takes place in science. What are some aspects of scientific measures that should be transferred to the measurement of individuals?

* In scientific measures great care is taken to evaluate one variable. For example a volt meter should only measure voltage. An amp meter should only measure amps. The same care that is taken in the design of lab instruments and should be applied to the design of "science education" measurement instruments.

* In the sciences a measurement instrument is built around a theory. The theory is conceived and is used to fabricate the instrument. The same should be true in the design of social science measurement instruments.

* Measurement instruments in laboratories are continuously calibrated. Balances are checked, as are voltmeters, and the optics of telescopes. For a while turntables were manufactured with strobe lights so that the spin rate could be finely adjusted with each playing of a record.

* All good measurement devices report errors. Furthermore, the error for the reading of a device is not always constant throughout the range of readings which can be made. For example, the error for a 0-10 volt meter may be .1 volts at a 5 volt reading, but at 9 volts the error may be .15 volts.

* Another characteristic of useful measurement instruments is the ability to work equally well in a number of situations. In the case of voltage, it does not matter what voltage is being measured (whether it is in a kitchen, in South Carolina, or California). If the device can not measure at a number of sites it is not very useful.

Discussion:

All of the brief points of the introduction describe a common characteristic of powerful measurement instruments. These characteristics should be taken into consideration when one designs and evaluates measurement instruments in the social sciences. The remainder of this paper will present a detailed discussion of the points raised in the introduction. Much work needs to be done to improve the measurements made with tests and attitudinal surveys, however, by bearing in mind those assets of scientific measurement devices- great progress should be made.

Designing a Measurement Instrument

A measurement instrument must be theory driven. It can be your personal theory or another person's theory, but there must be some basis to the instrument.

Secondly, the theory should point to one variable for measurement. What one variable is to assessed? Next, once a variable is considered, what questions would sample parts of the variable?

Preparing for the Redesign of the Measurement Instrument

Once a measurement device is designed to evaluate a variable (i.e. attitudes towards science) the concern for the meaning of the "variable" should not be shelved. After the design of the instrument and before data collection one should be able to predict respondents' answers. For example, in the case of an attitudinal instrument, which items will be the most "easy to agree with", and which items are the most "difficult to agree with"? In the case of a multiple choice test- which items will be the most difficult to solve and which will be the most easy? If this sort of checking is not done before an evaluation, it is difficult to fully evaluate the data.

Variables, Items, and Persons

What is the appearance of a unidimensional variable? How might it be manifested? To imagine the interplay between the definition of a single variable and the measurement of respondents, consider the figure 1 which is much akin to the number line.

```
                    Bob

     1     5   7 2    4    3            6
<-------|------|----|-|-----|----|-------------|------>

Less Likely                              Most Likely
to Agree With                            to Agree With
```

Figure 1

In this diagram the 7 items from a fictitious survey are presented. Note that they are aligned with a particular spacing from the left to the right. Those items on the "less likely to agree with" part of the scale are those items that were found to be least likely for survey respondent to agree with. Those items to the right side of the scale were those which respondents were most likely to agree with. The location of "Bob" plotted on the variable line helps indicate that Bob, from a probabilistic standpoint, is highly likely to "agree with" item 6 of the survey, and likely to "disagree with" the remaining survey items.

Does Everyone Use the Instrument in the Same Manner? How Important is this?

If a measurement instrument is well designed, and functioning correctly, then the spacing and ordering of these items should not change regardless of the individuals measured by the instrument. If there are great shifts in the location of items then one learns that individuals using the measurement instrument are not utilizing the device in the same way. Certainly, one can appreciate the necessity of all respondents using the measurement instrument in the same manner, by considering the common everyday ruler. When measurements are made with a ruler not only are the calibration of the ruler's marks trusted, but the assumption is made that if a number of individuals collect measures with a ruler then everyone uses the ruler in the same manner. If people differ in ruler measuring techniques the data is of little use- the same is true for surveys and tests which do not function equally well with all respondents.

Preparing for Measurement Devices that do not Measure Each Person in the Same Manner

Just as the prediction of item ordering is critical in the analysis of a measurement device- so too is a prediction and a concern for those few items on a survey which cause students to react in an unpredictable manner. What is meant by "unexpected"? If a measurement instrument is designed correctly, and truly measures one variable, then students (from a probabilistic standpoint) will predictably agree with some items and disagree with other items. The number of items a survey taker "agrees with" or "disagrees with" will be a function of their overall attitude. Figure 2 shows the response of Sue in which she answers unexpectedly to item 4. This is a case in which a respondent is not using an instrument as it was designed. To prepare for this possibility an attempt should be made to evaluate those items which might cause an unexpected response before the data is analyzed. This makes the evaluator more aware and critical of the measurement device begin used.

Figure 2

```
                        Sue

<----1----3-----11-----2----10------8--6----7-  9---4--12--13-5->
     D   D     D      D    D      A A   A  A   D  A  A A

Likely to Disagree With                   Likely to Agree With
```

Using Unexpected Response to Improve Measures

The unexpected answer of Sue to item 4 suggests that there is something different about this item and/or this person. If only one person is answering in an unexpected way to this item, the evaluator learns something important about this person. In that case this item should probably be retained for the surveying of other individuals. However, in measuring this one person their response to this one item should be removed.

If a number of individuals react in an unexpected way to this item- these data suggest that the item is not functioning according to the theory by which the measurement instrument was designed. There may be many reasons for the poor functioning of an item-- the design of the item may be faulty or the item points to a misunderstanding in the theory used to design the evaluation instrument.

What to do when an item is causing a large number of individuals to react in an unexpected way? For measurements that are to be made with this instrument this item should be removed from the analysis. Finally, unexpected response by a very large number of individuals to one item may also be a sign that data may have been miskeyed and/or the answer key to a survey or test was misentered.

Errors of Persons

Up until this point the discussion has centered on critical ways in which a measurement device can be 1) built with a theory, and 2) improved with predictions made before (and while) data is being collected. Also it has been pointed out that it is important to remove items and people who are not using the measurement device as predicted.

Another aspect of measurement in the science that is often not carried forth to the social sciences is the accurate reporting of measurement error. Each person who completes a test or attitudinal questionnaire has a unique error which is a function of the number of items answered and the types of answers given to each item. For useful measurement to proceed such "person" errors must be reported. Consider figures 3a and 3b.

Figure 3a

```
   |--Bob--|     |----Jim----|
<---------------------------------------------->
         |----Sue----|
```

Negative Attitude   Positive Attitude

Figure 3b

```
      Bob              Jim
<-------------------------------------------->
             Sue

Negative Attitude                          Positive Attitude
```

Without error reporting it appears as if the ordering of students' attitudes is clear, however, when measurement error is reported it becomes apparent that Bob's attitude is not statistically different from Sue, but it is different from Jim's.

<u>Error of Items</u>

The reporting of measurement error for test and survey items is just as important as it is for individuals. The error reported for items is dependent upon the number of persons answering items (not everyone may answer all items) and the part of the scale which the items occupies.

Figure 4a

```
             |--1--|   |--3--|
<---------------|----|----|---->
                  |--2--|
Disagree                      Agree
```

Figure 4b

```
                1       3
<---------------|----|----|---->
                     2
Disagree                      Agree
```

As was shown in figure 3, the reporting of measurement error has great implication for the interpretation and use of a measurement scale. Figure 4a shows the calibration of 3 test items with error bars, while figure 4b shows the items without error bars. Commonly researchers will claim (in terms of attitude) that item 3 is clearly above item 2 and item 2 is clearly above items 1, but when measurement error is taken into consideration this can be seen to be untrue.

Errors of all Persons not the Same

Not only must measurement errors be reported, but it is important to note that errors of persons will not all be the same, nor will all errors of items.

For Persons

If a student gets most of the items on a test correct, their measure (how able they are) will be quite high- however, the error of their ability estimate will be greater than those individuals who answered items in a mixed manner (e.g. half of the items correct). The reason for this pattern is simple to understand if one just considers the data. If John answers all of the items correctly on an Algebra test we know John knows a lot, but we do not know how much more he knows (thus there is great error in our measuring of John's ability). If another student (Sam) gets half the test items correct- we have a much more certain knowledge of what he knows in terms of Algebra. Thus the measurement error of Sam's ability is much smaller.

For Items

If an item on an exam is correctly answered by all of the students we learn that the question was quite easy for students- but one does not know with great accuracy how easy the item was, for few students had a differing reaction to the item. Thus the error for this item calibration (how easy or difficult the item is) will be large in comparison to an item that might have been correctly answered by half the students.

Other Gains From Objective Measurement Using a Stochastic Model

What other gains are there from utilizing objective measurement techniques to conduct science education measures?

1) Students must not answer all the items on a test or questionnaire. When data is missing it will only mean that the measurement error of the person will be greater.

2) One concern when comparing students over the course of many years is whether or not the same measurement scale is being used in the same manner (even if identical items are administered). By using objective measurement techniques and determining the spacing and calibration of items from a survey, one can anchor items at values which will define the same scale whenever a test or questionnaire is given. The best way to visualize this anchoring is to consider the marks on a ruler or a thermometer. When measures are taken with a particular ruler or thermometer, the location of centimeter marks or degrees is well understood and invariant. It is the invariance of the scale that allows useful measures to be made. This is why the ability to anchor a scale is so important.

3) The math behind the stochastic model corrects for the non-linearity of "test counts" (how many items are right or wrong), and the non-linearity of rating scales. What is meant by non-linearity? Consider a basic rating scale often used to collect data:

Strongly Agree	Agree	Disagree	Strongly Disagree

Usually a student's selection of "Strongly Agree" is counted as a 4, while an "Agree" is counted as a 3, "Disagree" as a 2 and "Strongly Disagree" as a 1. The reverse ordering (SD=4) can just as well be given. Now comes the mistake that many evaluators make- the labels "4", "3", "2", and "1" are considered measures, however, by doing so an implicit (an often incorrect) assumption is made that a jump in attitude from "Agree" to "Disagree" is the same as the jump in attitude as from "Disagree" to "Strongly Disagree". This is not necessarily the case at all. The psychometrician can not forget that the numbers "4, 3, 2, 1" are only labels that show what category was selected. The selections can not be immediately used to indicate a "known" spacing between categories. By calculating objective measures, a correction for the non-linearity of rating scales can be made.

Key Formulas from Best Test Design (Wright and Stone: Mesa Press, Dept. of Education, The University of Chicago, 5835 S. Kimbark Ave., Chicago, IL 60637).

logit=ln [(r/L)/(1-r)/L)]

r is number of items correct and L is the number of test items.

The above equation given the person ability.

logit=ln[[(N-S)/N]/[1-(N-S)/N]]
N= number of responses to an item
N-S is the number of incorrect responses top an item

The above equation gives the item difficulty

What is the model? it is the probabilistic Rasch model:

log (Pni/(1-Pni))=Bn-Di

Pni is the probability of person n getting item i correct

Bn is the ability of person n

Di is the difficulty of item i

Supplies: ruler, volt meter, amp meter, rocks

c:\u\article\carol.1