MEASURING IN THE AFFECTIVE DOMAIN

Presented by:

Larry E. Miller, Professor
Kattlyn Wolf, Ph D Graduate Student
Department of Human and Community Resource Development
208 Agricultural Administration Building
2120 Fyffe Road
The Ohio State University
Columbus OH 43210-1027
Phone: (614) 292 – 6909
Fax: (614) 292-7007
Email: miller.103@osu.edu
wolf.375@osu.edu

Presented to:

Professional Development Seminar Participants
Association for International Agricultural and Extension Education
E.A.R.T.H. University, Costa Rica
March 14, 2008
Measuring in the Affective Domain

Objectives:
At the conclusion of this seminar, you will be able to:
1. Define an attitude.
2. Describe the two major techniques used for measuring attitudes.
3. Describe the procedures to follow in developing instruments to measure attitudes.

Defining Attitudes
(For this seminar, attitudes will = perceptions, = opinions; = beliefs).
We’ll use the word attitudes.
"An attitude is the affect for or against a psychological object" Thurstone, 1931
evaluation of
like or dislike of
positiveness or negativeness toward (NOT feelings about)

Assumptions about Attitudes
Attitudes are predispositions to respond.
Attitudes are persistent over time.
Attitudes are susceptible to change, but not easily.
Attitudes produce consistency in behavior.
Attitudes are directional (Summers, 1970).

ATTITUDINAL SCALES
The terms attitudes, opinions, perceptions and beliefs will be used synonymously.
Kerlinger notes that measuring any of these constructs is similar. Please note that attitudes cannot be directly measured; therefore, we must construct a scale to measure them (they are latent factors). “Measurement scales are collections of items combined into a composite score, and intended to reveal levels of theoretical variables not readily observable by direct means, and are often referred to as scales” (DeVellis, 2003, pp 8-9). Scales consist of “effect-indicators” – items whose values are caused by an underlying construct -- latent variable (DeVellis, 2003, p 10). An index, on the other hand, is used to describe “cause-indicators” that it, items determine the level of a construct. [A construct is an abstraction – name – created to denote something that cannot be directly observed, e.g., love, justice; leadership.] An indicator to describe a political candidate’s appeal might be comprised of items about geographical residence, family size, physical attractiveness, ability to inspire campaign workers, and potential financial resources (DeVellis, 2003, p 10). He continues by noting that items of an index do not share any common cause, but they share an “effect” – increasing the likelihood of election in the campaign (p 10). We are most often interested in attitudinal scales where items do comprise a construct, a latent variable, an effect-indicator, or a domain.

Types of Attitudinal Scales
- Thurstone (equal-appearing interval)
- Likert (summated rating)
- Guttman (cumulative)
- Semantic Differential

**General Criteria for Attitude Statements**
- Items should be a series of statements; not questions.
- Avoid statements that refer to the past rather than to the present.
- Avoid statements that are factual or capable of being interpreted as factual.
- Avoid statements with multiple interpretations.
- Avoid statements that are irrelevant to the psychological object under consideration.
- Avoid statements that are likely to be endorsed or not endorsed by almost everyone.
- Select statements that are believed to cover the entire range of the affective scale of interest.
- Keep the language of the statements simple, clear, and direct. Avoid words that may not be understood by the respondents.
- Statements should be short; rarely exceeding 20 words.
- Each statement should contain only one complete thought.
- Statements containing universals such as *all, always, none, and never* often introduce ambiguity and should be avoided.
- Words such as *only, just, merely*, and other of similar nature should be used with care and moderation.
- Statements should be in the form of simple sentences rather than compound or complex sentences.
- Avoid the use of double negatives.
- Statements should not contain contractions.
- Avoid double-barreled statements.

**THURSTONE-TYPE SCALES** (Equal-appearing interval scales)

Louis Leon Thurstone, 1887 -1955

Thurstone proposed the equal appearing interval scale and his name is frequently associated with it. The understanding of the development procedure is essential to the construction of an equal appearing interval scale, and a summary of the procedures would include:

1. Construct a pool of items
- literature
- expert opinion
- personal experience
- pilot interviews

2. Secure a group of judges: for 100 to 150 items, one might secure 40-60 judges; know topic; own attitudes unimportant

3. Judges might be instructed as follows:
   “Note that there is only one item per card (slip) and that each item is numbered. The statements are in no particular order, but range from extremely favorable to extremely unfavorable. Before beginning your task, please read through all of the items carefully.”

PROCEDURES
A. Judges would be told to sort the items into 11 piles (A-K).

“Put the most favorable items into pile “A”. The items that are a little less favorable are put into pile “B”. Continue in this manner through pile “K”, which will contain the most unfavorable items. Your agreement or disagreement with the items is to be disregarded. Do not attempt to get an equal number into each stack.”

[Such a process is called a Q-Sort. Remember this technique because it can also be used to create “weightings” for items should one desire to do so.]

B. Return to the researcher.
   The “piles” might be made, for example, into old shoe boxes or manila envelopes.

C. Record on the back of the card/slip the letter corresponding to the pile into which the judge placed the item.

D. Record the frequencies for each item as shown on the following table:

E. Record proportions and cumulative proportions, if you plan to determine the scale and “Q” values graphically.
F. Calculate the scale and “Q” values for each statement.

<table>
<thead>
<tr>
<th>Item</th>
<th>f</th>
<th>p</th>
<th>cp</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>Scale</th>
<th>Q Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>.0</td>
<td>1</td>
<td>2</td>
<td>.0</td>
<td>1</td>
<td>2</td>
<td>.0</td>
<td>1</td>
<td>6</td>
<td>.0</td>
<td>3</td>
<td>6</td>
<td>.0</td>
<td>2</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.8</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>40</td>
<td>28</td>
<td>50</td>
<td>26</td>
<td>28</td>
<td>14</td>
<td>4</td>
<td>6.0</td>
<td>8</td>
<td>6.9</td>
<td>2.8</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>26</td>
<td>56</td>
<td>44</td>
<td>44</td>
<td>14</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>1.0</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The “Q Value” is a measure of the spread of the middle 50% of the judgments. The higher the “Q Value”, the more disagreement there was among judges. Thurstone regarded high “Q values” primarily as an indicator that a statement was ambiguous. To get Q, determine score at 25% and score at 75%; and subtract score at 25% from score at 75%. This can be done by constructing a cumulative proportion curve (Ogive curve) with the scores on the abscissa and the percentage or proportion on the ordinate. The resultant Ogive curve will typically form an “S” and is also known as an “S-curve”. One can also calculate the 25th and 75th percentiles from the table by extrapolation.

The “Scale Value” is the midpoint of the values; the median; the point (score) above and below which 50% of the cases fall. In this example, 50% would be somewhere between “F”(6) and “G”(7). On the cp, extrapolate between 6 and 7; finding 6.8. One could also “read” the score from the Ogive curve, if constructed, at the 50th percentile.

G. Constructing the Scale: Using items with the lowest “Q values”, select items with equal intervals, e.g.: 1.5, 2.0, 2.5, 3.0, 3.5 ... 10.5, 11. Arrange in random order on the instrument. Thus, the interval between them should seem the same to the subject: an equal-appearing interval scale.
H. Administering the Scale: Subjects are told to: “Check all items with which you agree.” For Example:

******************************************************************************

Attitude toward War

Put a check mark (√) next to those statements with which you agree.

( ) 1. Under some conditions, war is necessary to maintain justice.

( ) 2. The benefits of war rarely pay for its losses even for the victor.

( ) 3. War brings out the best qualities in humankind.

( ) 4. There is no conceivable justification for war.

( ) 5. War has some benefits, but it’s a big price to pay for them.

( ) 6. War is often the only means of preserving national honor.

[etc.]

******************************************************************************

I. Scoring Examples

Subject #1 agrees with 5 statements with Scale Values of: 3.2, 4.5, 5.6, 7.2, and 8.9. The subject’s score is the middle term (5.6).

Subject #2 agrees with 4 statements: 4.5, 5.6, 7.2 and 8.9. Their score is extrapolated between the middle two: Score = 5.6 + (7.2 - 5.6)/2 = 6.4

These scores can be treated as interval in scale of measurement.

J. Reliability

The typical method of establishing reliability is to select two scales of 20 items each. Administer both scales to the same subjects. Correlate the results with a Pearson product moment correlation coefficient. The result would be a “Coefficient of Equivalency” and would be reported as such.

DISADVANTAGES OF THE THURSTONE SCALE

1. Laborious process
2. Difficulty in getting enough judges
3. Possible effects of the judges’ own attitudes on their classification of the statements
4. Reliability tends to be only “adequate”
5. Always a question of: “Would a different set of judges have produced a different result?”
6. Sometimes one can go through the laborious process and still not have an instrument because the scale values do not provide adequate dispersion for the creation of equal intervals.

ADVANTAGES

1. Easy for the respondents to answer
2. Easy for the researcher to score and analyze


1. For each statement, circle the number that indicates the degree of favorableness of the statement toward the attitudinal object:

<table>
<thead>
<tr>
<th>Extremely unfavorable</th>
<th>Extremely favorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9</td>
<td>1. Divorce should be encouraged for many unhappy people.</td>
</tr>
</tbody>
</table>

1 2 3 4 5 6 7 8 9     2. Divorce brings happiness to some people, unhappiness to others. (etc.)

For each statement, place a mark on the adjacent line indicating the degree of favorableness of the statement toward the attitudinal object.

<table>
<thead>
<tr>
<th>Extremely unfavorable</th>
<th>Neutral</th>
<th>Extremely favorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Divorce should be encouraged for many unhappy people.</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

(etc.)

For each statement, circle the letter that indicates the degree of favorableness of the statement toward the attitudinal object. An 11 (9, 7 or so forth) interval stencil or ruler is superimposed over each line, like clear vinyl or transparency film, to quantify each judge's rating.]

<table>
<thead>
<tr>
<th>Extremely unfavorable</th>
<th>Neutral</th>
<th>Extremely favorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Divorce should be encouraged for many unhappily married people.</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

(etc.)

LIKERT-TYPE SCALES (*Summated Rating Scales*)
Rensis Likert, 1903–1981

Likert-type (Summated Rating) Scale; Advantages (Scale name is always capitalized as it is a proper noun: his name.)

- Easy for respondents to complete, most people familiar with the scale
- Relatively easy to construct
- Most popular attitudinal measure
- Easy to score and analyze
- Each item considered to be of equal attitude value (weight) -- homogeneous items

Summary of Steps:

- Identify the attitudinal object and delimit it quite specifically.
- Compose a series of statements about the attitudinal object that are half positive and half negative and are not extreme, ambiguous, or neutral.
- Establish, minimally, content validity with the help of an expert panel.
- Pilot test the statements to establish reliability (Cronbach’s alpha/K-R 21).
- Eliminate statements that negatively affect internal consistency.
- Construct the final scale by using the fewest number of items while still maintaining validity and reliability; create a balance of positive and negative items.
- Administer the scale and instruct respondents to indicate their level of agreement with each statement.
- Sum each respondent’s item scores to determine attitude.

1. Remember, this is a summated rating scale!! Individual items do not stand alone for analysis, but are summated (See Appendices A & B). The researcher can either use the sum, or the average, across a number of reliable and valid items. To handle individual items, simply report the frequency of responses for each response category and this would produce a mode, but do not use means! Items comprise a sample to measure a construct, domain or latent variable. Like trying to describe a wisp of smoke, we would need numerous samples of all the molecules. Just one molecule would not suffice. For example, with an instrument to measure professor’s attitudes toward teaching, theory might indicate that attitudes toward teaching are comprised of four domains: (1) planning, (2) delivery, (3) self-reflection and (4) student evaluation. Each of these could be considered to comprise a domain. Essentially, one would develop an instrument to assess each by establishing validity and reliability. One would
establish “internal consistency reliability” for each set of items comprising a domain because each will produce one number per subject (operational definition of each variable/characteristic). One would NOT do one Cronbach’s alpha for the total instrument across all domains, but calculate one alpha for each domain!!

2. Likert-type scales are probably the most popular type of scale because subjects are often familiar with them. They can be used with many different types of populations and can be adapted for pictures as well as words for the scale response categories.

3. Relatively easy to construct.

4. Easy for the respondents to complete.

5. Easy to score and analyze.

6. Procedures to develop Likert-type scale statements:
   A. Write items; can be from literature, personal opinion, personal interview information, focus group interviews, etc.; typically use the fewest items possible but still must be valid and reliable.
   B. Make about ½ positive and ½ negative statements, without being too extreme and avoiding the middle ground which might constitute vagueness or ambiguity. All items should be of approximately the same intensity.
   C. Written as statements, not questions!
   D. Field test for “face validity.” Establish internal consistency reliability with a pilot test to provide evidence of the ability to summarize. This would mean to use the Cronbach α or K-R 21. As a minimum, a panel of experts should establish content validity (See Appendix A & C; and E for an example format to use).

7. Procedures to develop Likert-type scale response categories:
   A. Can use words, acronyms for words, numbers or pictures
   B. For agreement scales, I recommend the use of even-numbered scales with levels of agreement and disagreement. You might have 2, 3, 4, or 5 levels of agreement and the same for disagreement. While some scales you see in literature will have a neutral point, Summers, G.F. (Attitude Measurement, Chicago, Rand McNally, 1970) notes that “attitudes have a directional quality with positive or negative affectations” (p. 2). If you perceive that the subjects may not have a frame of reference from which to judge a certain statement, then create a Not Applicable (NA) column and record the frequency of such NA responses, but do not use the data in the summation; only report the frequency of this response. [Discussions about the desirability of using “undecided” or “no-opinion,” middle-scale-points (odd-numbered scales) have occurred for many years. On this issue, I am on the side of even-numbered-scales because of my constitutive definition of an attitude.]
One can also use anchored scales and the number of categories is not an issue, e.g., a frequency scale: 1= never, 2 = sometimes, 3= frequently, 4= very often, and 5= all the time.

C. Clearly label all response categories on the final instrument. See Dillman, D. (Mail and Telephone Surveys, The total design method) for suggested layout for the items and the total questionnaire. See Appendix D for an example of a properly constructed questionnaire.

**Scaling Summary**
Response scales vary. Recommend to use an even-number of response categories (no neutral category), and use a N/A response for agreement scales. Label all response categories. Since it is a summated rating scale, the scale of measurement of the sum is interval (or, you can use the mean). Never analyze by item. It reduces the scale of measurement to ordinal. One often sees t-tests calculated on individual items. This is **WRONG**! First, the t-test uses means, and means should not be calculated on individual items because the mean is used with interval/ratio data and the scale of measurement of an item is ordinal. Secondly, the t-test assumes that the scale of measurement of the data is interval/ratio and this would not be true.

**Anchored scales:** frequency, importance, perceived knowledge level, etc. Pictures, thermometers, etc., may be used as scales. Multiple scales per item may be used.

D. Certainty Method
This method was developed by Warren, R. et al. (Department of Rural Sociology and Anthropology, Iowa State University, Report No. 82, Ames, 1969) to address the problem of limited variability of data from a Likert scale. Imagine, if you had a four-point scale, the data can just vary from 1 to 4. As a researcher you might have an agreement scale as: Strongly disagree = 1, Disagree = 2, Agree = 3; Strongly agree = 4. (Dillman says to always have the most positive statements assigned larger numbers.) You would probably code these data as 1-4, for each item, for each subject. **However**, is the semantic space (meaning of the words) between each the equivalent of “1”? Warren et al. argue that the distance between strongly disagree and disagree (currently 1 point) may be greater than the semantic distance between agree and disagree (also currently a 1 point spread).

They suggest constructing scales asking the subjects to first agree (A) or disagree (D), and then indicate their level of “Certainty” for the answer on a scale of 1-5 with each number having a descriptor:

A
D
(Respondents circle a letter and a number for each item.)

The response categories then become, with scores below for transformed values, and scores below that being the “Certainty Scores”

D5  D4  D3  D2  D1  D/A  A1  A2  A3  A4  A5
1  2  3  4  5  6  7  8  9 10 11
0  3  5  6  7  8  9 10 13 16

Note that this gives strong levels of certainty greater values and the semantic space is not always “1” between the response categories. Other “Certainty Score” weightings could be devised other than 0-16.

E. Other Considerations

- Anchored scales: importance, frequency, etc.

- Multiple scales to items; e.g., competency lists with Importance and Current Knowledge each comprising scales

Example: **Multiple Scales**

For each Competency, please rate (1) the level of importance of to your work and (2) your current level of knowledge

<table>
<thead>
<tr>
<th>Importance</th>
<th>Competency</th>
<th>Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4</td>
<td>Tomato Varieties</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>1 2 3 4</td>
<td>Tomato Wilt</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>1 2 3 4</td>
<td>Pruning Suckers</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

(Etc.)

Using needs theory, with discrepancy gap analysis:

Educational Need = (I – K) (Mean I)
Produces 3 scales: I, K and Need with 2 measured (Can measure more.)

- For each subject, sum (or use the mean) their responses to all items on that domain to produce one number that is the operational definition of an attitude for them. If you do this across 85 subjects, then, to describe the group, one would calculate a mean of the means (grand mean) to describe the group’s attitude. If there are two groups and one is examining a “comparative hypothesis” about differences between the groups, then the grand mean of each group would be used in the test of significance, e.g., t-test for independent groups.

F. Analysis
The **Sum or Average** produced by individuals would be assumed to be at the **interval scale** of measurement. Response to any one item **would be ordinal** measurement, but we know we should not just analyze single items **The Likert is a summated rating scale**!

**Guttman Scales (Cumulative Scale):** Louis Guttman, 1916-1987

The Guttman Scale consists of a relatively small set of homogeneous items that are supposedly unidimensional, measuring one and only one attitude dimension. Such scales get their name from the cumulative relation between items and the total score of individuals. Items can be ordered in difficulty, complexity or value-loading so that to answer correctly or approve the last implies success or approval on all the preceding ones; or, to miss a middle item implies failure.

Procedure: (1) define the area of content, (2) select 4-6 statements relating to the area. The number of items to select is a “mystery”. Guttman says selection is based on intuition and experience. Statements selected should have some homogeneous content, said one author. Each statement is more intense, shows a more positive attitude, than the last: it is cumulative.

**Example:**

1. I would speak to an Asian on campus.
2. I would invite an Asian to dinner.
3. I would live next door to an Asian.
4. I would marry an Asian.

Each item has a ( _ ) YES or ( _ ) NO response category

**The Cornell Technique**

- Assume each statement is to have only two response categories, such as agree – disagree with weights of 1 and 0 assigned to the two response categories
- The statements are then given to a group of about 100 subjects and each statement is scored

<table>
<thead>
<tr>
<th>Statements</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>4</td>
</tr>
</tbody>
</table>
### Another Guttman Example:

We ask four children three arithmetic questions:

- a) $28/7 =$
- b) $8 \times 4 =$
- c) $12 + 9 =$

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Child</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Second Child</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Third Child</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Fourth Child</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

If we know a child’s total score, then we can predict his/her pattern, if the score is cumulative. Logic is that: if a child gets “a” correct, then likely they will get “b” and “c” correct. If child misses “a”, but gets “b” correct, then is likely to get “c” correct. If child misses “c”, child is not likely to get “a” or “b” correct. People could be ranked by this technique (unidimensional). Some books note that the Guttman is useful to examine small shifts or changes in attitudes. Guttman scales, as a rule, are highly reliable.

Disadvantages are that arbitrary standards are employed, procedures are laborious, and, after going through all the procedures, there is no certainty that a usable scale will result.

#### Semantic Differential (SD)

The SD has no standard scale and no standard concepts. The concepts and scales are meticulously selected by the researchers to meet the criteria of the study. Thus, the nature of the problem dictates the class and form of concepts to be utilized. The three elements of the SD are the (1) concept, (2) bipolar adjectives which delineates part of the concept, and (3) an unspecified space (or number of positions) between the pair of bipolar adjectives.

### CONCEPT

<table>
<thead>
<tr>
<th>Polar Term X</th>
<th>___ : ___ : ___ : ___ : ___ : ___</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polar Term Y</td>
<td></td>
</tr>
</tbody>
</table>

Bipolar terms have been factor analyzed (see above book) to produce evaluative, potency, and activity factors which can vary, meld, or be neutralized by the nature and design of the study. The evaluative factors in human judgment (e.g.: good __ __ __ __ __ __ bad) seem to command greater relative importance to the semantic space dimensions. The potency factors are associated with size, weight, toughness. The activity factors are concerned with quickness, excitement, and agitation. The authors state that these factors are not only factors which can vary meanings and cite stability, tautness, novelty, and perceptivity factors as examples. AIAEE members probably would use evaluative dimension most frequently.

#### Some Paired Adjectives for the Evaluation Dimension

- Good – Bad
- Nice – Awful
- Beautiful – Ugly
- Pleasant – Unpleasant
- Clean – Dirty
- Sweet – Sour
- Sacred – Profane
- Valuable – Worthless
- Good – Bad
- Nice – Awful
- Beautiful – Ugly
- Pleasant – Unpleasant
- Clean – Dirty
- Sweet – Sour
- Sacred – Profane
- Valuable – Worthless (Osgood et al., 1957)

The criterion for selecting scales:
1. "... is this factorial composite -- we usually select about three scales to represent each factor, these being maximally loaded on the factor and minimally on others".
2. the relevance to the concepts being studied.
3. their semantic stability for the concepts and subjects in a particular study.
4. is for the scales to be linear between polar opposites and pass through the origin.

Once the concepts you wish to measure are identified and the respective scales, then the scaling method is applied.

GRADUATE SCHOOL ADMISSION PROCESS
The polar adjectives should be randomly placed at the right or the left, end positions. Some authors recommend not having over 7 or 8 sets of bipolar adjectives under each concept as the respondents seem to lose their frame of reference to the concept when there is more than this.

Rather lengthy directions must be given to subjects. For example:

The purpose of this study is to measure the meanings of certain things to various people by having them judge against a series of descriptive scales. In completing this questionnaire, please make your judgments on the basis of what these things mean to you. On each section you will find a different concept (in the middle) to be judged and beneath it a set of scales. You are to rate the concept on each of these scales in order. If you feel that the concept is very closely related to one end of the scale, you should place your check mark as follows:

<table>
<thead>
<tr>
<th>Fair</th>
<th>x:</th>
<th>:</th>
<th>:</th>
<th>:</th>
<th>:</th>
<th>Unfair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>x</td>
</tr>
</tbody>
</table>

If you feel the concept is quite closely related to one or the other end of the scale (but not extremely), you should place your check mark as follows:

<table>
<thead>
<tr>
<th>Strong</th>
<th>:</th>
<th>:</th>
<th>x:</th>
<th>:</th>
<th>:</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>x</td>
</tr>
</tbody>
</table>

If the concept seems slightly related to one side as opposed to the other side, then you should place your check mark as follows:

<table>
<thead>
<tr>
<th>Active</th>
<th>:</th>
<th>:</th>
<th>:</th>
<th>x:</th>
<th>:</th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>x:</td>
<td>:</td>
<td></td>
</tr>
</tbody>
</table>

The direction toward which you check, of course, depends on which of the two ends of the scale seem most characteristic of the thing you are judging.

IMPORTANT:

1. Place your check marks in the middle of the spaces, not on the boundaries at the colons.
2. Be sure you check every scale for every concept -- do not omit any.
3. Never put more than one check mark on a single scale.

Sometimes you may feel as though you have had the same item before on the questionnaire. This will not be the case, so do not look back and forth through the items. Do not try to remember how you checked similar items earlier in the questionnaire. Make each item a separate and independent judgment. Work at a fairly high rate of speed throughout the questionnaire. It is your immediate impressions, your immediate "feelings" about the items that we want. On the other hand, please do not be careless, because we want your true impressions.

Scoring is typically done by labeling the most positive adjective as a six (6) and the most negative as a one (1), and numbers proceed: 1, 2, 3, 4, 5 & 6. The mean of each subject on the scales under the concept constitutes their "attitude" score on this concept. For reliability, since this is a summated scale, an internal consistency reliability estimate (See Appendices) should be conducted, through a pilot test, on each concept to assure that the scales are additive before the questionnaire is used.

An advantage of the SD is that once directions are produced, they can be used with many concepts and scales. A second advantage is that the SD seems to be as valid and reliable as other techniques and, thus, questionnaires could be produced quickly.
disadvantage is that it takes the subject a long time to read the directions and they are not as familiar with this technique as the Likert scale.

Some References on Attitudinal Measures


Appendix A

Calculating, Interpreting, and Reporting Reliability (Internal Consistency) of Multiple-Item Instruments with Likert-Type Scaling

**Validity**: The essence of validity is truthfulness. Validity is concerned with this question: Are the data produced by an instrument true? Validity is an evaluative judgment of the extent to which an instrument measures what it purports to measure. Validity pertains to systematic (nonrandom) measurement error.

**Criterion-related validity**
How accurately an instrument predicts an accepted indicator (criterion) of a concept. Requires empirical investigation of the correlation between performance on the instrument and performance or behavior on the criterion. Limited usefulness in the social sciences; the more abstract the concept being measured the less likely there are indicator criteria against which performance on the instrument can be evaluated. Concurrent validity: Correlation between performance on the instrument and performance on the criterion concurrently. Predictive validity: Correlation between performance on the instrument and performance on the criterion at a future time.

**Content validity is:**
(1) the degree to which an instrument assesses the relevant aspects of the concept or domain the instrument purports to measure; (2) the adequacy with which a specified domain of content is sampled; representativeness of the items on the instrument to the domain of content for the concept being measured. Content validity is insured by the plan and procedures used to construct the instrument. Content validity cannot be determined by correlating performance on the instrument with an external criterion; the instrument itself is the criterion.

Content validity is appraised by subjective judgment, of experts in the domain being measured, of the adequacy with which relevant content has been sampled and the adequacy with which the content has been presented in the items comprising the instrument.

Appraisal of content validity is empirical in the sense that the evidence is judgment by knowledgeable persons regarding the content of the domain being measured and the procedures used to develop and use the instrument. [See Appendix E for an example: Note added by Miller]

Exploratory principal components analysis and common factor analysis can be used to identify the underlying constructs that an instrument measures.


Exploratory principal components analysis and common factor analysis can be used to identify (a) the number of different dimensions of the underlying construct that an instrument measures, (b) the items that constitute each different dimension, (c) how strongly each item defines its
relevant content dimension, and (d) the extent to which the identified underlying dimensions are related to one another.

**Face validity**

Concerns the degree to which an instrument “appears” to measure what it purports to measure. Is the extent to which an instrument “looks like” it measures what it is intended to measure. Face validity is a subjective judgment by persons who either administer or complete the instrument concerning whether the instrument appears, “on its face,” to measure what it is intended to measure. Face validity concerns judgments about an instrument after it is constructed; content validity is insured by the plan for determining content and the plan for constructing items. Face validity is an aspect of content validity concerned with an inspection of the final instrument to make sure that plans for insuring content validity are transformed into the final instrument.
Construct validity

Construct validity is determining whether an instrument actually measures the underlying construct it is intended to measure. The first step in establishing construct validity involves the formulation of an explicit definition of the construct that is purported to be measured. A second essential step is specifying the observable behaviors that are related to the construct being measured. Construct validity is established by investigating hypothesized relationships between the measurement of the construct and the observable behaviors related to the construct that have been identified. The establishment of construct validity is difficult; it involves extensive empirical investigations of relationships between measurement of the construct and observable behaviors. The degree to which it is difficult to establish construct validity of psychological variables is proportional to the degree the variable (construct) being measured is concrete or abstract.
Reliability: Defines the trustworthiness of an instrument. Reliability of the scores produced by an instrument pertains to consistency, repeatability, stability, dependability, and generalizability. Reliability pertains exclusively to random measurement error. Reliability is inversely related to the amount of random error in the scores produced by the instrument. Reliability is an empirical quantitative issue.

Test-Retest

Estimates reliability based on the correlation of scores from an instrument administered to the same respondents on two different occasions.
A major defect of test-retest is that the experience completing the first test will influence responses on the retest.
There is a tendency for test-retest to overestimate reliability due to respondents remembering their response on the test and repeating the responses on the retest.
Low correlation between performance on the test and performance on the retest may indicate that the underlying concept being measured has changed. The longer the time between measurements, the more likely the construct has changed.
Test-retest is frequently used in place of the parallel forms method of establishing reliability.

Parallel (alternative) forms

Estimates reliability from the correlation between scores on two versions of the same test administered to the same respondents.
Parallel forms method is superior to the test-retest method since it eliminates memory as a factor that tends to inflate the estimate of reliability.
A limitation of the parallel forms method is the difficulty in constructing alternative forms of an instrument that are parallel.

Split-half

Total items on the instrument are split in half and the scores on the two halves are correlated to estimate reliability.
The correlation between halves is statistically corrected (Spearman-Brown formula) to estimate reliability for the total instrument (Carmines & Zeller, 1979, p. 41).
A major problem with the split-half method is that the correlations between the halves will differ depending on how the number of total items is divided into halves.

Inter-rater reliability

Estimates reliability is calculating the extent to which two or more persons agree on their ratings or observations of phenomenon.
Estimate of reliability calculated as the proportion (or percent) of the ratings or observations for which there is agreement between the raters. [Intra-rater reliability is the agreement within an observer over time. Note added by Miller.]

Internal consistency
Internal consistency concerns the idea that single items on a multiple-item instrument converge to measure the underlying construct the instrument purports to measure. Cronbach’s alpha is an internal consistency estimate of reliability (Carmines & Zeller, 1979, p. 44). Cronbach’s alpha depends on the mean inter-item correlation among items and the number of items in the instrument.

When scale items are dichotomously scored (e. g. 0 = wrong; 1 = right), Cronbach’s alpha is equivalent to Kuder-Richardson 20 (KR 20) as an estimate of reliability. KR 20 is a special case of Cronbach’s alpha. Cronbach’s alpha is a conservative estimate of reliability. Reliability can never be lower than Cronbach’s alpha (Carmines & Zeller, 1979). Cronbach’s alpha requires only a single administration of an instrument. It is recommended that Cronbach’s alpha be computed for any multiple-item instrument (Carmines & Zeller, 1979).


<table>
<thead>
<tr>
<th>Criterion Rating</th>
<th>Exemplary</th>
<th>Extensive</th>
<th>Moderate</th>
<th>Minimal</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-item</td>
<td>Average of</td>
<td>Average of</td>
<td>Average of</td>
<td>Average below .10</td>
<td>Not reported</td>
</tr>
<tr>
<td>correlations</td>
<td>.30 or better</td>
<td>.20 -.29</td>
<td>.10 -.19</td>
<td>below .10</td>
<td>reported</td>
</tr>
<tr>
<td>Coefficient</td>
<td>.80 or better</td>
<td>.70 -.79</td>
<td>.60 -.69</td>
<td>&lt; .60</td>
<td>Not reported</td>
</tr>
</tbody>
</table>

“In the early stages . . . time and energy can be saved using instruments that have only modest reliability, e. g., .70” (Nunnally & Bernstein, 1994, pp. 264-265).

“As a general rule . . . reliability should not be below .80 for widely used scales . . . the most important thing to remember is to report reliability of the scale and how it was calculated” (Carmines & Zeller, 1979, p. 51).
Table 1
Examples of Researcher-Developed Multiple-Item Instruments with Likert-Type Scaling
Volume 41 (2000) *Journal of Agricultural Education*

<table>
<thead>
<tr>
<th>Construct Measured</th>
<th>Response Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitudes (perceptions) about . . .</td>
<td>1 = Strongly disagree; 2 = Disagree; 3 = Undecided; 4 = Agree; 5 = Strongly agree (Alternatives: 3 = Neutral, Uncertain, or No Opinion)</td>
</tr>
<tr>
<td>Self-assessed knowledge</td>
<td>1 = Know nothing; 2 = Know little; 3 = Know some; 4 = Know a lot</td>
</tr>
<tr>
<td>Level of knowledge/skill</td>
<td>1 = Don’t know enough to respond; 2 = Below average; 3 = Average; 4 = Above average; 5 = Expert</td>
</tr>
<tr>
<td>Level of competence</td>
<td>1 = Poor; 2 = Fair; 3 = Good; 4 = Strong</td>
</tr>
<tr>
<td>Expected impact from . . .</td>
<td>1 = Very unlikely; 2 = Somewhat unlikely; 3 = Unsure; 4 = Somewhat likely; 5 = Very likely</td>
</tr>
<tr>
<td>Impact of . . .</td>
<td>0 = None; 1 = Minor; 2 = Moderate; 3 = Major; 4 = Critical</td>
</tr>
<tr>
<td>Importance of . . .</td>
<td>1 = Not important; 2 = Somewhat important; 3 = Important; 4 = Very important</td>
</tr>
<tr>
<td>Need for . . .</td>
<td>1 = Not important; 2 = Sort of important; 3 = Important; 4 = Very important</td>
</tr>
<tr>
<td>Involvement in . . .</td>
<td>1 = Never; 2 = Seldom; 3 = Sometimes; 4 = Much of the time; 5 = Always</td>
</tr>
<tr>
<td>Extent term/phrase relates to . . .</td>
<td>1 = Not at all related; 2 = Related only some; 3 = Related; 4 = Strongly related</td>
</tr>
<tr>
<td>Faith in sources of information</td>
<td>1 = Very low; 2 = Low; 3 = Neutral; 4 = High; 5 = Very high</td>
</tr>
<tr>
<td>Acceptance of . . .</td>
<td>1 = Highly unacceptable; 2 = Somewhat unacceptable; 3 = Somewhat acceptable; 4 = Highly acceptable</td>
</tr>
<tr>
<td>Obstacles to . . .</td>
<td>1 = Insignificant; 2 = Moderately insignificant; 3 = Slightly insignificant; 4 = Slightly significant; 5 = Moderately significant; 6 = Significantly</td>
</tr>
<tr>
<td>Value of . . .</td>
<td>1 = None; 2 = Low; 3 = Moderate; 4 = High; 5 = Very high</td>
</tr>
<tr>
<td>Usefulness of . . .</td>
<td>1 = Not useful; 2 = Fairly useful; 3 = Useful; 4 = Very useful; 5 = Very much useful</td>
</tr>
<tr>
<td>Program quality</td>
<td>1 = Very low; 2 = Low; 3 = Average; 4 = High; 5 = Very high</td>
</tr>
</tbody>
</table>
Table 2

<table>
<thead>
<tr>
<th>Articles</th>
<th>No.</th>
<th>%</th>
<th>Reliability Coefficients Reported</th>
<th>Analysis Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10.3</td>
<td>None</td>
<td>None</td>
<td>Single item analysis exclusively</td>
</tr>
<tr>
<td>14</td>
<td>48.3</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Single item analysis exclusively</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.4</td>
<td>Cronbach’s alpha: Total scale and/or subscales Test-Retest: selected single items</td>
<td>Single item analysis exclusively</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.4</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Single item analysis primarily; Some summated (composite) score analysis</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>13.8</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Single item analysis and Summated (composite) score analysis</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>13.8</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Summated (composite) score analysis primarily; Some single item analysis</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7.0</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Summated (composite) score analysis exclusively</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Single-Item Analysis Versus Multiple-Item (Composite) Analysis**

Single items usually have a low degree of relationship with the particular construct being measured; single items may relate to constructs or attributes other than the constructs being measured.

*It is very unlikely that a single item can fully represent a complex underlying construct that is being measured.* [Emphasis added by Miller]

Single-item measures lack precision because they cannot discriminate among fine degrees of a construct or attribute.

Single items tend to be less valid, less accurate, and less reliable than multiple-item composites.

Single items have considerable measurement error; single items are prone to random error.

The unreliability of single items “averages out when scores on a number of items are summed to obtain a total score, which then frequently is highly reliable” (Nunnally, 1978, p. 67). Aggregation: A fundamental principle in statistics and measurement. “Through aggregation the random sources of error that contaminate each observed score have a chance to cancel out and leave standing a better estimate of the true score” (Strube, 2000, p. 30).
Because single items “provide only a single measurement, the social scientist rarely has sufficient information to estimate their measurement properties. Thus, their degree of validity, accuracy, and reliability is often unknowable” (McIver & Carmines, 1981, p.15).

“With a single measure of each variable, one can remain blissfully unaware of the possibility of measurement error, but in no sense will this make inferences more valid . . . . I see no substitute for the use of multiple measures of our most important variables ....” (Blalock, 1970, 2. 111).

Whenever possible, social scientists should employ multiple-item measures of theoretical concepts” (McIver & Carmines, 1981, p. 16).

Example A: Reliability of Single Items versus Summated (Composite) Scores for a Cluster of Items


Table 4
Test-Retest Agreement of Responses to Single Likert-Type Items (n = 30)

<table>
<thead>
<tr>
<th>Item a</th>
<th>Proportion of Responses with Agreement Between Test and Retest b</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITEM01 Inadequate wages/salary</td>
<td>.70</td>
</tr>
<tr>
<td>ITEM02 Cost of work/family needs</td>
<td>.67</td>
</tr>
<tr>
<td>ITEM03 Attitudes state administrators</td>
<td>.57</td>
</tr>
<tr>
<td>ITEM04 Administrator’s work/family policies</td>
<td>.60</td>
</tr>
<tr>
<td>ITEM05 Resentment among employees</td>
<td>.60</td>
</tr>
<tr>
<td>ITEM06 Lack of role models</td>
<td>.40</td>
</tr>
<tr>
<td>ITEM07 Organizational culture</td>
<td>.47</td>
</tr>
<tr>
<td>ITEM08 Lack of interest of administrators</td>
<td>.37</td>
</tr>
<tr>
<td>ITEM09 Lack of ability of administrators</td>
<td>.53</td>
</tr>
<tr>
<td>ITEM10 Too many meetings</td>
<td>.43</td>
</tr>
</tbody>
</table>

a Response scale: 1 = Not at all; 2 = To a slight extent; 3 = To a moderate extent; 4 = To a great extent; 5 = To a very great extent.
b Sixteen days between test and retest.
Table 5
Internal Consistency: Summated (Composite) Scores (n = 181)

<table>
<thead>
<tr>
<th>Cluster and Items</th>
<th>Item-total Correlation</th>
<th>Cronbach’s alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1 – Organizational Issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITEM03 Attitudes of state administrators</td>
<td>.79</td>
<td>.91</td>
</tr>
<tr>
<td>ITEM04 Administrator’s work/family policies</td>
<td>.74</td>
<td>.91</td>
</tr>
<tr>
<td>ITEM05 Resentment among employees</td>
<td>.58</td>
<td>.92</td>
</tr>
<tr>
<td>ITEM06 Lack of role models</td>
<td>.77</td>
<td>.91</td>
</tr>
<tr>
<td>ITEM07 Organizational culture</td>
<td>.82</td>
<td>.91</td>
</tr>
<tr>
<td>ITEM08 Lack of interest of administrators</td>
<td>.86</td>
<td>.90</td>
</tr>
<tr>
<td>ITEM09 Lack of ability of administrators</td>
<td>.78</td>
<td>.91</td>
</tr>
<tr>
<td>ITEM10 Too many meetings</td>
<td>.57</td>
<td>.93</td>
</tr>
<tr>
<td>Inter-item Correlations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean = .59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum = .34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum = .82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cronbach’s alpha = .92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 2 – Financial Issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITEM01 Inadequate wages/salary</td>
<td>.73</td>
<td>–</td>
</tr>
<tr>
<td>ITEM02 Cost of work/family needs</td>
<td>.73</td>
<td>–</td>
</tr>
<tr>
<td>Inter-Item Correlation = .73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cronbach’s alpha = .84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| a Clusters of items identified by principal components analysis with oblique rotation. |

Reliability of single items measuring demographic characteristics and behavior

Single questions about personal attributes (demographic characteristics) and behaviors result in “very little measurement error” (Salant & Dillman, 1994, pp. 87-90). (See Table 6)
### Table 6
Test-Retest Agreement for Single Items: Demographic Characteristics (n = 30)

<table>
<thead>
<tr>
<th>Demographic Characteristic</th>
<th>Proportion of Responses Agreement Between Test and Retest $^a$</th>
<th>Correlation coefficient $^r$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Categorical Variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>.97</td>
<td>–</td>
</tr>
<tr>
<td>Ethnic origin</td>
<td>1.00</td>
<td>–</td>
</tr>
<tr>
<td>Level of education</td>
<td>1.00</td>
<td>–</td>
</tr>
<tr>
<td>Martial status</td>
<td>1.00</td>
<td>–</td>
</tr>
<tr>
<td>Employment status</td>
<td>.97</td>
<td>–</td>
</tr>
<tr>
<td>Position</td>
<td>.94</td>
<td>–</td>
</tr>
<tr>
<td>Program specialization</td>
<td>.97</td>
<td>–</td>
</tr>
<tr>
<td><strong>Metric Variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (year born)</td>
<td>.93</td>
<td>.95</td>
</tr>
<tr>
<td>Hours per week worked</td>
<td>.71</td>
<td>.88</td>
</tr>
<tr>
<td>Years employed</td>
<td>.83</td>
<td>.99</td>
</tr>
</tbody>
</table>

Note. Data from Clutter (1998). $^a$ Sixteen days between test and retest.
Guidelines for Calculating, Interpreting, and Reporting Internal Consistency Reliability Estimates

“It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees. Thus, authors should provide reliability coefficients of the scores for the data being analyzed . . . Interpreting the size of observed effects requires an assessment of the reliability of the scores” (Wilkinson and the Task Force on Statistical Inference, 1999, p. 596).

Cronbach’s alpha as an estimate of the reliability (internal consistency) of summated scores on a multiple-item instrument provides no information about the reliability of scores of single items comprising the instrument.

When Cronbach’s alpha is reported for the total instrument or for subscales of items comprising the instrument, the computed reliability estimate pertains exclusively to summated (composite) scores for all items on the instrument or for clusters (subscales) of multiple items.

When Cronbach’s alpha is reported for the total instrument or for subscales of items comprising the instrument, the researcher has defined the variable or variables (summated scores) that are appropriate for reporting and interpreting the constructs measured.

If the analysis and interpretation is exclusively in terms of single items comprising the instrument, evidence of the reliability of the scores for single items must be reported. When analysis and interpretation is exclusively by single items, the reporting of Cronbach’s alpha as an estimate of reliability is meaningless. The content of single items can be used to “name” the constructs measured by composites of single items. When summated (composite) scores are calculated for the total multiple-item instrument or for subscales of items comprising the instrument, the researcher should report the distributions of summated scores and interpret the “meaning” (in terms of the construct measured) of the various values of the summated scores. (See the histograms reported on page 11.) Item analysis of multiple-item instruments (item-item correlation matrix and item-total correlations) can be used to make decisions about items that should be retained or deleted from an instrument or a subscale of the instrument. [See Spector (1992, pp. 29-46) and Table 5, page 10].

Exploratory principal components analysis and common factor analysis are useful techniques for identifying the underlying constructs (clusters of items) measured by a multiple-item instrument with Likert-type scaling.
References


Appendix B

Calculating, Interpreting, and Reporting Cronbach’s Alpha Reliability Coefficient for Likert-Type Scales

Joseph A. Gliem
Rosemary R. Gliem

Presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, The Ohio State University, Columbus, OH, October 8-10, 2003.

Abstract
The purpose of this paper is to show why single-item questions pertaining to a construct are not reliable and should not be used in drawing conclusions. By comparing the reliability of a summated, multi-item scale versus a single-item question, the authors show how unreliable a single item is; and therefore it is not appropriate to make inferences based upon the analysis of single-item questions which are used in measuring a construct.

Introduction
Oftentimes, information gathered in the social sciences, marketing, medicine, and business, relative to attitudes, emotions, opinions, personalities, and description’s of people’s environment involves the use of Likert-type scales. As individuals attempt to quantify constructs which are not directly measurable they oftentimes use multiple-item scales and summated ratings to quantify the construct(s) of interest. The Likert scale’s invention is attributed to Rensis Likert (1931), who described this technique for the assessment of attitudes.

McIver and Carmines (1981) describe the Likert scale as follows:
A set of items, composed of approximately an equal number of favorable and unfavorable statements concerning the attitude object, is given to a group of subjects. They are asked to respond to each statement in terms of their own degree of agreement or disagreement. Typically, they are instructed to select one of five responses: strongly agree, agree, undecided, disagree, or strongly disagree. The specific responses to the items are combined so that individuals with the most favorable attitudes will have the highest scores while individuals with the least favorable (or unfavorable) attitudes will have the lowest scores.
While not all summated scales are created according to Likert’s specific procedures, all such scales share the basic logic associated with Likert scaling. (p. 22-23)

Spector (1992) identified four characteristics that make a scale a summated rating scale as follows: First, a scale must contain multiple items. The use of summated in the name implies that multiple items will be combined or summed. Second, each individual item must measure something that has an underlying, quantitative measurement continuum. In other words, it measures a property of something that can vary quantitatively rather than qualitatively. An attitude, for example, can vary from being very favorable to being very unfavorable. Third, each item has no “right” answer, which makes the summated rating scale different from a multiple-choice test. Thus summated rating scales cannot be used to test for knowledge or ability. Finally, each item in a scale is a statement, and respondents are asked to give rating about each statement. This involves asking subjects to indicate which of several response choices best reflects their response to the item. (p 1-2)

They identify the following: First, individual items have considerable random measurement error, i.e. are unreliable. Nunnally and Bernstein (1994) state, “… measurement error averages out when individual scores are summed to obtain a total score” (p. 67). Second, an individual item can only categorize people into a relatively small number of groups. An individual item cannot discriminate among fine degrees of an attribute. For example, with a dichotomously scored item one can only distinguish between two levels of the attribute, i.e. they lack precision. Third, individual items lack scope. McIver and Carmines (1981) say, “…it is very unlikely that a single item can fully represent a complex theoretical concept or any specific attribute for that matter” (p. 15). They go on to say, “The most fundamental problem with single item measures is not merely that they tend to be less valid, less accurate, and less reliable than their multiitem equivalents. It is rather, that the social scientist rarely has sufficient information to estimate their measurement properties; thus, their degree of validity, accuracy, and reliability is often unknowable” (p. 15). Blalock (1970) has observed, “… with a single measure of each variable, one can remain blissfully unaware of the possibility of measurement [error], but in no sense will this make his inferences more valid” (p. 111).

Given this brief background on the benefits of Likert-type scales with their associated multi-item scales and summated rating scores, many individuals consistently invalidate research findings due to improper data analysis. This paper will show how data analysis errors can adversely affect the inferences one wishes to make.

Data Analysis Errors with Likert Scales

Reporting Errors with Reliability Measures

While most individuals utilizing Likert-type scales will report overall scale and subscale internal consistency reliability estimates in the analysis of the data, many will analyze individual scale items. Table 1 shows the results of an analysis done by Warmbrod (2001) of The Journal of Agricultural Education, Volume 41 (2000). Volume 41 of the journal contained 44 articles of which 36 (82%) were quantitative. Of these 36 articles, 29 (66%) used researcher developed Likert scales for which internal consistency reliability measures need to be reported. The table shows the reliability coefficients reported and the analysis strategy used with the quantitative articles contained in the journal. As shown in the table, only two (7%) of the individuals correctly analyzed the data collected based upon the reliability measures reported. The majority of individuals correctly reported Cronbach’s alpha as the measure of internal consistency reliability, but then opted to conduct data analysis using individual items. This is particularly troubling because single item reliabilities are generally very low and without reliable items the validity of the item is poor at best and at worst unknown. This can be illustrated using a simple data set of actual data collected from a class of graduate students enrolled in a Winter Quarter, 2003, research design course. Cronbach’s alpha is a test reliability technique that requires only a single test administration to provide a unique estimate of the reliability for a given test. Cronbach’s alpha is the average value of the reliability coefficients one would obtained for all possible combinations of items when split into two half-tests.

Calculating Cronbach’s Alpha Coefficient for Internal Consistency

A single statement (item) was presented to each student and then this same statement was presented to the student 3 weeks later. A test-retest reliability coefficient was calculated on this individual statement (item) since individual items can not have a Cronbach’s alpha internal consistency reliability calculated. The statement presented to each student was, “I am pleased with my graduate program at The Ohio State University.” Students were asked to respond to the statement using a five-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). Figure 1 shows a scatterplot of student response for the first administration of the statement and for the second administration of the statement 3 weeks later. The test-retest reliability coefficient for this statement was .11. A multi-item scale was also developed and given to the same students to measure their attitude towards their graduate program. The multi-item scale is presented in Table 2.
Table 1. Reliability estimates and analysis strategies: Researcher-developed multiple-item instruments with Likert-type scaling

<table>
<thead>
<tr>
<th>Articles</th>
<th>Number</th>
<th>%</th>
<th>Reliability Coefficients Reported</th>
<th>Analysis Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>10.3</td>
<td>None</td>
<td>Single item analysis exclusively</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>48.3</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Single item analysis exclusively</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.4</td>
<td>Cronbach’s alpha: Total scale and/or subscales Test-Retest: selected items</td>
<td>Single item analysis exclusively</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.4</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Single item analysis primarily; Some summated score analysis</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>13.8</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Single item analysis analysis and summated score analysis</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>13.8</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Summated score analysis primarily; Some single item analysis</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.0</td>
<td>Cronbach’s alpha: Total scale and/or subscales</td>
<td>Summated score analysis exclusively</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Scatterplot of first administration and second administration; r = .11

Table 2. Multi-item statements to measure students’ pleasure with their graduate program at The Ohio State University

<table>
<thead>
<tr>
<th>Item</th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. My advisor is knowledgeable about my program</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2. It is easy to get an appointment with my advisor</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 3 shows the item-analysis output from SPSS for the multi-item scale of student attitude towards their graduate program. A description of the sections and related terms are as follows:

1. Statistics for Scale – these are summary statistics for the 8 items comprising the scale. The summated scores can range from a low of 8 to a high of 40.
2. Item means – these are summary statistics for the eight individual item means.
3. Item Variances – these are summary statistics for the eight individual item variances.
4. Inter-Item Correlations – This is descriptive information about the correlation of each item with the sum of all remaining items. In the example in Table 2, there are 8 correlations computed: the correlation between the first item and the sum of the other seven items, the correlation between the second item and the sum of the other seven items, and so forth. The first number listed is the mean of these eight correlations (in our example .3824), the second number is the lowest of the eight (.0415), and so forth. The mean of the inter-item correlations (.3824) is the r in the $\alpha = rk / [1 + (k -1) r]$ formula where k is the number of items considered.
5. Item-total Statistics – This is the section where one needs to direct primary attention. The items in this section are as follows:
   a. Scale Mean if Item Deleted – excluding the individual item listed, all other scale items are summed for all individuals (48 in our example) and the mean of the summated items is given. In Table 2, the mean of the summated scores excluding item 2 is 25.1.
   b. Scale Variance if Item Deleted – excluding the individual item listed, all other scale items are summed for all individuals (48 in our example) and the variance of the summated items is given. In Table 2, the variance of the summated scores excluding item 2 is 25.04.
   c. Corrected Item-Total Correlation – this is the correlation of the item designated with the summated score for all other items. In Table 2, the correlation between item 2 and the summated score is .60. A rule-of-thumb is that these values should be at least .40.
   d. Squared Multiple Correlation – this is the predicted Multiple Correlation Coefficient squared obtained by regressing the identified individual item on all the remaining items. In Table 2, the predicted Squared Multiple Regression Correlation is .49 by regressing item 2 on items 4, 5, 6, 7, 8, 9, and 10.
   e. Alpha if Item Deleted – this is probably the most important column in the table. This represents the scale’s Cronbach’s alpha reliability coefficient for internal consistency if the individual item is removed from the scale. In Table 2, the scale’s Cronbach’s alpha would be .7988 if item 2 were removed for the scale. This value is then compared to the Alpha coefficient value at the bottom of the table to see if one wants to delete the item. As one might have noted, the present scale has only 8 items where the original scale had
10 items. Using the above information, removing items 1 and 2 resulted in an increase in Cronbach’s alpha from .7708 to .8240.

f. Alpha – the Cronbach’s alpha coefficient of internal consistency. This is the most frequently used Cronbach’s alpha coefficient.

Table 3. Item-Analysis from SPSS Output

<table>
<thead>
<tr>
<th>Statistics for Scale:</th>
<th>Mean</th>
<th>Variance</th>
<th>Std. Dev.</th>
<th>Number of Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29.1042</td>
<td>30.8187</td>
<td>5.5515</td>
<td>8</td>
</tr>
<tr>
<td>Item Means:</td>
<td>Mean</td>
<td>Minimum</td>
<td>Maximum</td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td>3.6380</td>
<td>3.3125</td>
<td>3.9792</td>
<td>.6667</td>
</tr>
<tr>
<td>Item Variances:</td>
<td>Mean</td>
<td>Minimum</td>
<td>Maximum</td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td>1.0750</td>
<td>.7017</td>
<td>1.4109</td>
<td>.7092</td>
</tr>
<tr>
<td>Inter-Item Correlations:</td>
<td>.3824</td>
<td>.0415</td>
<td>.5861</td>
<td>.5446</td>
</tr>
<tr>
<td>Item-total Statistics:</td>
<td>Scale Mean</td>
<td>Scale Variance</td>
<td>Corrected Item-Total Mean</td>
<td>Corrected Item-Total Variance</td>
</tr>
<tr>
<td>Item 2</td>
<td>25.1250</td>
<td>25.0479</td>
<td>.6046</td>
<td>.4909</td>
</tr>
<tr>
<td>Item 4</td>
<td>25.7917</td>
<td>23.2748</td>
<td>.5351</td>
<td>.3693</td>
</tr>
<tr>
<td>Item 5</td>
<td>25.6667</td>
<td>24.6525</td>
<td>.4260</td>
<td>.4474</td>
</tr>
<tr>
<td>Item 6</td>
<td>25.2500</td>
<td>25.2128</td>
<td>.5134</td>
<td>.4587</td>
</tr>
<tr>
<td>Item 7</td>
<td>25.6250</td>
<td>22.9202</td>
<td>.6578</td>
<td>.5104</td>
</tr>
<tr>
<td>Item 8</td>
<td>25.7083</td>
<td>24.3387</td>
<td>.4473</td>
<td>.3116</td>
</tr>
<tr>
<td>Item 9</td>
<td>25.1250</td>
<td>23.9840</td>
<td>.6134</td>
<td>.5202</td>
</tr>
<tr>
<td>Item 10</td>
<td>25.4375</td>
<td>24.0811</td>
<td>.6432</td>
<td>.4751</td>
</tr>
</tbody>
</table>

Reliability Coefficients for 8 items

Alpha = .8240 Standardized item alpha = .8320

g. Standardized Item Alpha - the Cronbach’s alpha coefficient of internal consistency when all scale items have been standardized. This coefficient is used only when the individual scale items are not scaled the same. Cronbach’s alpha reliability coefficient normally ranges between 0 and 1. However, there is actually no lower limit to the coefficient. The closer Cronbach’s alpha coefficient is to 1.0 the greater the internal consistency of the items in the scale. Based upon the formula $\alpha = \frac{rk}{[1 + (k-1)r]}$ where $k$ is the number of items considered and $r$ is the mean of the inter-item correlations the size of alpha is determined by both the number of items in the scale and the mean inter-item correlations. George and Mallery (2003) provide the following thumb-rules: “$\alpha > .9$ – Excellent, $\alpha > .8$ – Good, $\alpha > .7$ – Acceptable, $\alpha > .6$ Questionable, $\alpha > .5$ Poor, and $\alpha < .5$ Unacceptable” (p. 231). While increasing the value of alpha is partially dependent upon the number of items in the scale, it should be noted that this has diminishing returns. It should also be noted that an alpha of .8 is probably a reasonable goal. It should also be noted that while a high value for Cronbach’s alpha indicates good internal consistency of the items in the scale, it does not mean that the scale is unidimensional. Factor analysis is a method to determine the dimensionality of a scale but is beyond the scope of this paper.

Conclusions
When using Likert-type scales, it is imperative to calculate and report Cronbach’s alpha coefficient for internal consistency reliability for any scales or subscales one may be using. The analysis of the data then must use these summated scales or subscales and not individual items. If one does otherwise, the reliability of the items is at best probably low and at worst unknown. Cronbach’s alpha does not provide reliability estimates for single items.

References


Appendix C

Calculating Reliability and Validity

Measurement Error

Measurement error can be defined as all the systematic effects which operate to bias recorded results given that a result was obtained from the sampling unit. Some of these might be (1) questions not clearly stated, (2) ambiguously stated questions, (3) instructions not clear, (4) tendency of respondents to give socially acceptable answers-reactive effects, (5) respondents do not possess the correct information, and/or (6) respondents deliberately lie. As related to these examples, let us discuss validity, reliability, and suitability.

In research, it is important to have findings and conclusions based on data that are both relevant and accurate. Relevant data are valid data and indicate that the measuring procedure did what it claimed to do. Thorndike and Hagen suggest that when the question of validity is addressed “we are inquiring whether the test measures what we want it to measure, all of what we want it to measure, and nothing but what we want it to measure.”

In contrast, accurate data are reliable data and are the result of an instrument that consistently measures whatever it measures. Other terms for reliability are dependability, stability, consistency, predictability and accuracy. Validity and reliability can both be explained in terms of measurement error. There are two types of measurement error, systematic and random. Let us look at systematic error, first.

Validity. Validity refers to the systematic error in measurement; are we measuring what we think we are measuring or is there some systematic error involved? Remembering Thorndike and Hagen’s definition of validity, an example of an invalid test might be a 4th grade mathematics test written at an 8th grade reading level. The test is not only measuring math ability but also reading ability, thus, the test would have validity problems.

Educational researchers need to be especially carefully when designing measurement instruments because the constructs in educational research; like achievement, attitude, motivation, creativity and aptitude; cannot be directly measured but must be inferred from representative measurement. Validity, then, refers to how well the measured indicators (representing the desired construct) really measure what they are supposed to measure.

Questions to ask when considering validity might include: Does this instrument really contain a good representation of the desired content? What psychological or underlying constructs are being measured? Does it measure other characteristics as well? Could I use this test to make useful predictions? Is it appropriate for the chosen population? Does it look like it is measuring what it claims to measure? The researcher needs to address each of these questions to determine the validity of the measuring instrument. Remember that the question of validity of an instrument is specific to the situation and purpose so an instrument deemed to be valid in one setting cannot be assumed to be so in a different setting.

There are four types of validity: content validity, criterion-related validity, construct validity and face validity. These are not independent of each other but do address specific aspects of an instrument.

Content validity refers to the representativeness of the items on the instrument as they relate to the entire domain or universe of content being measured. In other words, to have content validity, the
measure must adequately sample the domain of content the researcher claims it measures. See Appendix E for a sample form to use with a panel of experts.

Content validity cannot be represented numerically but is determined subjectively by a thorough examination of the instrument by a panel of expert judges. Many times, a field test can also be conducted with a population similar to the proposed population to help with content validity. For example, an instrument designed to measure attitudes of nine to eleven year olds could be given to a like group. The group might not only answer the items but respond to what the items actually say by writing comments in blank spaces next to each item. The researcher can use the results of the field test to reword items and make them more appropriate and understandable which in turn will increase validity.

If the researcher is interested in the predicting capability of the instrument, then criterion-related validity is extremely important. Criterion-related validity refers to the relationship between scores on the instrument in question and scores on another measuring instrument or other external variable referred to as a criterion. If scores on the criterion are obtained at the same time as scores on the original instrument, then the specific validity term to be used is concurrent validity. Conversely, if the criterion is a future measurement then predictive validity is in question. In either case, the criterion must be relevant (reflects the characteristic in question), reliable (consistent measurement) and free from bias [no factors influence if other than performance on the criterion] (Ary, et al.10). To determine concurrent validity, a validity coefficient can be obtained by correlating performance on the original instrument with performance on the criterion. To determine predictive validity, obtain scores using the original instrument but do not disclose them. At a later time, obtain criterion scores and then correlate the performance on both.

Construct validity answers the question “What does the instrument really measure?” Examples of constructs are anxiety, intelligence, motivation and attitudes. Performance on a test can be interpreted in terms of the specific targeted construct if there is construct validity. There is both a logical and an empirical approach to construct validity. Examining the items of the instrument logically to determine if elements represented are those elements that “make up” the construct is one approach.

Using an empirical method to test construct validity, elements purported to be indicators of the construct should be highly correlated. If they are not, then the conclusion can be that the instrument lacked construct validity and the theory and/or instrument should be reworked.

Validity coefficients can be used as indicators of construct validity. Indicators of the same construct should converge or come together by producing a high validity coefficient. For example, scores on a mathematics test should correlate highly with grades in mathematics. If this is the case, then the math test would have convergent validity.

Oppositely, the expectation would be that a construct valid instrument would produce scores that differ from an indicator measuring a different construct. An example would be that there would be a low correlation between a mathematics test and a reading comprehension test if the tests were construct valid. The tests would be judged to have discriminant (divergent) validity because they discriminated between constructs.

Factor analysis is another statistical method to determine construct validity. Common factor analysis can be used to determine (1) the number of constructs being measured, (2) which constructs are contributing to performance on which test and (3) the amount of variance in test scores accounted for by which constructs. All of this is done by examining intercorrelations of test scores. Construct validity is
certainly the most comprehensive type of validity since it includes content and criterion-related validity as well, and should not be overlooked when designing an instrument.

**Face validity** is the simplest kind of validity to determine and is as important if not more so than the other types, depending on the use of the instrument (Cascio24). Face validity refers to the appeal and appearance of the instrument. Does the instrument “look like” it is measuring what is supposed to? Sometimes instruments can elicit biased or incorrect responses because respondents do not take the task seriously because of the lack of face validity. An instrument can be judged to be face valid by the researcher.

An example of a reading test and its uses serves as quick review of validity. (Figure 1)

**FIGURE 1**

**VALIDATION OF A READING TEST FOR DIFFERENT PURPOSES**

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Type of Validation</th>
<th>Questions to be Asked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement test in 6th grade reading</td>
<td>content</td>
<td>How well does the test sample what has been learned?</td>
</tr>
<tr>
<td>Aptitude test to predict performance in 7th grade reading</td>
<td>criterion-related predictive</td>
<td>How well does the test predict achievement in 7th grade reading?</td>
</tr>
<tr>
<td>Diagnostic test to identify reading problems</td>
<td>criterion-related concurrent</td>
<td>How well does the test diagnose current reading difficulties?</td>
</tr>
<tr>
<td>A test to measure reading comprehension</td>
<td>construct</td>
<td>How well does the test measure reading comprehension? Do data support hypotheses about reading comprehension?</td>
</tr>
</tbody>
</table>

(Ary, Jacobs, Razavieh, 1985, p. 224.)

Remember, validity refers to systematic error in measurement. The better a test or instrument measures what it claims to measure, the less systematic error there is and the more valid the instrument is judged to be. An important point to remember is that validity is specific to the particular purpose of the instrument. An instrument constructed for more than one purpose or population may be valid in one instance but invalid in another.

**Reliability.** The second type of error in measurement is random error. Random error occurs by chance and in no systematic pattern, thus is cannot be predicted. The degree to which random error occurs in measurements indicates the relative reliability of the measuring instrument and data produced: the less the random error, the higher the reliability.
If something is deemed reliable, we can depend on it. Performance is consistent and predictable over time and use. Image a bull’s-eye representing a target we are trying to “hit” with a measuring instrument. If the measurement were reliable, the bull’s-eye would look like Target A. All the shots would be clustered together. The shots are consistently located in the same place. (However, not necessarily in the center.)

Target B would represent not only a reliable instrument but also a valid one because all the shots are clustered together and in the center. The shots consistently hit what was aimed for. Target C represents a measuring instrument that is neither reliable nor valid. These targets illustrate an important point -- an instrument cannot be valid if it is not reliable. It can, however, be reliable (as in target A) but not valid. We will look closer at this concept later when reviewing the relationship between reliability and validity.

To better understand the concept of reliability, let us review some basic principles of measurement.

1. Any observed score (X) on an instrument is comprised of a true scores (TS) and some unsystematic error (E). Under perfect conditions, the observed score would equal the true score.

\[ X = TS + E \quad \text{or} \quad TS = X - E \]

In addition, since each individual score is comprised of a true score and error, then

2. In any set of scores, there is observed score variance (\( \sigma_X^2 \)) which is comprised of true score variance (\( \sigma_T^2 \)) and error variance (\( \sigma_E^2 \)). Thus,

\[ \sigma_X^2 = \sigma_T^2 + \sigma_E^2 \]

3. Reliability can be theoretically defined as the ratio of true-score variance to the observed score variance in a set of scores.

\[ r_{XX} = \frac{\sigma_T^2}{\sigma_X^2} \]

In simpler terms, “reliability is the proportion of the variance in the observed scores that is free of error.” (Ary, et al., 228)

\[ r_{XX} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \]

The reliability coefficient (\( r_{XX} \)) can range from 0 to 1.0 representing the situation where the entire measurement is error and 1 representing the total absence of error. Viewing these equations it is easy to see the theoretical relationship between the reliability coefficient and random error variance. In reality, there can be many sources of random error in measurement which reduce reliability. Errors in scoring, errors due to guessing, errors due to day-to-day fluctuations in conditions and errors due to changes in individuals all serve to lower reliability. It is generally accepted that when measurement occurs there will always be some random error. Reliability procedures are concerned with minimizing random error by increasing the precision and consistency of the measuring instrument.
We can call an instrument reliable to the extent that after repeated measures, individuals’ scores remain in the same order for each administration or measurement. Any variation is error. One method used to measure reliability, then, is to administer an instrument to the same group twice and correlate the two sets of scores. This “test-retest reliability” yields a coefficient of stability [a Pearson product moment correlation].

When using a coefficient of stability of a measure of reliability, it is important to wait an appropriate length of time between administrations. Anastasi\(^25\) suggests that the length of time should be greater than zero but less than six months.

When reporting a coefficient of stability, always list the time interval between administrations and any significant experiences that may have intervened in the measurements. Also describe the conditions of each measurement to account for measurement error due to poor lighting, loud noises and the like.

Another method for determining reliability is the parallel form reliability procedure that yields a coefficient of equivalence (again, a Pearson r correlation coefficient). Because an instrument contains only a sample of possible items, it should be possible to construct parallel forms of an instrument. Parallel forms should contain the same number of items, contain items of equal difficulty, and have means, variances and interrelations with other variables that are not significantly different from each other (Gulliksen\(^26\)). In other words, forms are parallel if it does not matter which one you use.

Again, a correlation is calculated from the two sets of scores obtained at the same time. When reporting a coefficient of equivalence, include the length of time between administrations (no more than a few days maximum) and also a description of possible intervening variables.

Note that this method provides a conservative estimate of reliability due to the sampling error present in the construction of two forms of an instrument rather than one.

A third type of reliability coefficient is the coefficient of stability and equivalency that is produced by correlating scores from parallel forms that are administered at different times. A pretest, posttest measurement would be a good example of when to use this type of correlation. By using a parallel form of the pretest for the posttest, (rather than the exact form) there would be less of a threat of the pretest contaminating results on the posttest. This coefficient will be the most rigorous test of reliability and can serve as a lower bound of reliability. Again, when reporting results, record the length of time involved between administrations and any suspected intervening variables.

A measure of reliability can also be obtained using a single administration of an instrument and is generally referred to as a measure of internal consistency. The logic is that splitting the instrument into halves and correlating scores from each half will produce a coefficient of equivalence similar to the parallel forms method. The basic difference is that this method cuts the length of the test in half when computing the correlation. Because of the shorter test, this method produces a very conservative estimate of reliability. There is a correction procedure available, however, called the Spearman-Brown correction formula. “Split-half” reliability, then, is obtained by (1) either systematically (odd-even items) or randomly splitting the instrument into halves and (2) correlating the two halves. Then, using the Spearman-Brown correction formula, the obtained coefficient can be corrected to produce a more accurate estimate of reliability for the true length of the instrument.

There are other internal consistency measures that do not require the researcher to split the instrument in half and score each half separately but instead assess the homogeneity of the items. Two sources of random error are reflected in the measures of reliability using homogeneity of items: 1. content sampling (as in split-half) and 2. heterogeneity of the behavior domain sampled. The more
homogeneous the domain, the less error; thus, the higher the reliability. Let’s examine two procedures for determining reliability of using homogeneity of items: Kuder-Richardson procedures and the Cronbach’s alpha procedure.

The **Kuder-Richardson 20** (KR-20) is calculated using the proportion of correct and incorrect responses to each item on the instrument. Remember, because of the logic of this method, the KR 20 can only be used when the items on the instrument are scored dichotomously, that is, either right or wrong. A second formula, the **Kuder-Richardson 21** (KR 21), is simpler and can be used when there are multiple response categories. It does, however, assume that all items on the instrument are of equal difficulty. If this is not the case, then reliability estimates produced by this procedure will be lower than the actual.

Both the KR 20 and KR 21 are essentially “splitting the test in half” in as many ways as possible and will yield a similar estimate to the split-half method after the Spearman-Brown correction has been used.

The last reliability procedure to examine is the **Cronbach’s alpha**. Like the KR 21, the Cronbach’s alpha can be used when items have multiple response categories. Either the KR 21 or the Cronbach’s alpha would provide an appropriate estimate of reliability for the example in Figure 2. The items on this instrument require responses to an attitude measurement that uses a Likert-type scale where individuals may receive a score from one to six depending on the response.

**FIGURE 2**

<table>
<thead>
<tr>
<th>LEVEL OF AGREEMENT</th>
<th>(circle your response)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18. I am compensated fairly for the job that I perform.</td>
<td>VSD SD D A SA VSA</td>
</tr>
<tr>
<td>19. I work harder but get rewarded less than most other employees who hold a similar position.</td>
<td>VSD SD D A SA VSA</td>
</tr>
<tr>
<td>20. The Ohio Cooperative Extension Service promotes its employees according to how well they perform.</td>
<td>VSD SD D A SA VSA</td>
</tr>
<tr>
<td>21. It seems that the harder I work, the less I am formally rewarded.</td>
<td>VSD SD D A SA VSA</td>
</tr>
<tr>
<td>22. Generally speaking, I am paid equitably for my performance on the job.</td>
<td>VSD SD D A SA VSA</td>
</tr>
<tr>
<td>23. I seldom get formally recognized when I do a good job.</td>
<td>VSD SD D A SA VSA</td>
</tr>
</tbody>
</table>

In descriptive research, the most widely used and appropriate reliability tools are the KR 21 and the Cronbach’s alpha because both can be determined using only one administration of the instrument and both assess multiple response items. The choice between the two is the researcher’s and would depend on which statistical package is being used to analyze the data.

For example, SPSS-X will provide the Cronbach’s alpha. Remember, the purpose of a test of homogeneity, as an estimate of reliability, is to produce an instrument (or portions of an instrument) with items that are all measuring the same domain. And a good way to determine this would be to use the Cronbach’s alpha and follow this procedure.

1. Design items that appear to measure the same domain.
2. Collect data using a pilot test.
3. Run Cronbach’s alpha procedure.
4. View correlation matrix.
   • no 0’s or 1’s
   • no negatives
5. Decide, by viewing correlations, which items appear to be measuring the same domain.
   • do all items correlate highly?
   • eliminate negatives, ones, and zeros
6. Viewing “alpha if item deleted” at bottom of the printout
   • if alpha can be raised by deletion, decide if reduction in number of items will hurt content validity
7. Rerun alpha and repeat procedure.

Some factors influencing reliability include:
1. The greater the number of items, the more reliable the test.
2. On the whole, the longer the test administration time, the greater the reliability.
3. The narrower the range of difficulty of items, the greater the reliability.
4. The more objective the scoring, the greater the reliability.
5. The greater the probability of success by chance, the lower the reliability.
6. Inaccuracy in scoring leads to unreliability.
7. The more homogeneous the material, the greater the reliability.
8. The more common the experiences of the individuals tested, the greater the reliability.
9. “Catch/tricky” questions lower the reliability.
10. Subtle factors leading to misinterpretation of the test item lead to unreliability.

Another question about reliability: “How high should reliability of a measurement be?”
Actually, there is no absolute answer to this question. The appraisal of any instrument must always be in terms of other instruments with which it is in competition. Thus, a high school mathematics test with a reliability coefficient of .80 would look relatively unattractive if tests with reliabilities of .85 to .90 were already available. On the other hand, a procedure for judging “leadership” that had a reliability of no more than .60 might look very attractive if the alternative were a set of uncontrolled ratings having a reliability of .45 to .50. Although we cannot set an absolute minimum for the reliability that is required to enable us to achieve specified levels of accuracy in describing an individual or a group, a test with relatively low reliability will permit us to make useful studies of and draw accurate conclusions about groups, especially groups of substantial size; but, quite high reliability is required if we are to speak with confidence about individuals. Nunnally suggested that: “What is a satisfactory level of reliability? The answer depends on how a measure is being used. In the early stages of research on predictor tests or hypothesized measures of a construct, one saves time and energy by working with instruments that have only modest reliability, for which purposes reliability of .50 or .60 will suffice.”

Remember, a reliability coefficient of .90 says that 90% of the variance in scores is due to true variance in the characteristic measured, which leaves only 10% of the variance due to error. Because our goal in descriptive research is to describe as accurately as possible, high reliability is imperative to good descriptive research. (See Appendices A & B)

**The Relationship between Reliability and Validity.** It was stated that there can be reliability without validity but not validity without reliability. Intuitively, that makes sense but let us look at measurement error to better understand just what the relationship between reliability and validity really is.
Remember that observed variance \( \sigma \) in a set of scores contains true variance and random error variance:

\[
\sigma_X^2 = \sigma_T^2 + \sigma_E^2
\]

Now, let’s look at what makes up true variance (\( \sigma_T^2 \)). True score variance (indicated by the measure of reliability) consists of relevant valid variance (\( \sigma_V^2 \)) and irrelevant but systematic variance (\( \sigma_I^2 \)). The full equation looks like the one in Figure 3.

**FIGURE 3**

\[
\sigma_X^2 = \sigma_V^2 + \sigma_I^2 + \sigma_E^2
\]

The observed score variance equals the relevant valid variance plus the irrelevant but systematic variance plus random error variance.

We can say then, that reliability is the proportion of true variance in a set of scores and validity is the proportion of true variance that is relevant to the purpose of the measurement procedure (Cascio). (Figure 4)

**FIGURE 4**

<table>
<thead>
<tr>
<th>reliability</th>
<th>relevant valid variance (validity)</th>
<th>irrelevant but systematic variance</th>
<th>random error variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed variance</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As random error decreases, reliability (true score variance) increases and, potentially, validity (relevant valid variance) increases also, but in reality, it could be either relevant or irrelevant variance that increases. Therefore, low random error alone does not guarantee high validity, only high reliability. And, high reliability is a necessary condition of validity but not sufficient alone.

**Suitability.** As related to suitability (a part of validity), consider the questionnaire that is sent to 9-12 age 4-H’ers, that contains questions written by a college student at a college reading level. See some problems? Questions may be well written, but can be above the reading level and abilities of the target population (9-12 year old 4-H’ers) to read and understand. There are means of determining the reading level. Suitability is actually a part of validity. If an instrument is not suitable, then it will not be valid. Nonetheless, it is an important idea to keep separate for those of us who gather data from persons with limited literacy. One way to approximate the level of education needed to read a given passage is to use the “Fog Index”:

1. Take a sample of 100 words. Divide the total number of words in the sample by the number of sentences. This gives the average sentence length (SL) of the sample;
2. Count the number of words with three of more syllables in the 100-word sample. Do not count words (a) that are capitalized, (b) there are combinations of short, easy words (like: bookkeeper, or butterfly); (c) that are verb forms...
made three syllables by adding “es” or “ed” (like: created or trespasses). This gives you the percentage of hard words (%HW) in the passage; (3) Calculate Fog Index:

# years of education = (SL + %HW) (.4)

One readability program is already in Microsoft Word. Below is the sequence of steps to set your Word program to do a “readability” check each time you do a spell check.

Calculating “Readability” with Microsoft Word

Open Word:
Click on “Tools”
Click “Options”
Click “Spelling/Grammar”
Click the box for “Show readability statistics”
Readability data will be shown after each spell check

Determining the Number of Items Needed to Obtain a Desired Reliability
(Assuming all items are good items)

\[ N = \frac{P^a (1-P^b)}{P^b (1-P^a)} \]

Where \( P^a \) = desired reliability
\( P^b \) = existing reliability
\( N \) = Increased (times) number of items needed

Example:

\( P^b = .60 \)
\( P^a = .80 \)
\( N = ??? \)

You now have 10 items.

\[ N = \frac{.8 (1-.6)}{.6 (1-.8)} = \frac{.32}{.12} = 2.66 \]

Therefore, you would need to increase the number of items by 2.7;
10 X 2.7 = 27 similar items needed to get a reliability of .80
Appendix D

**Likert Agreement Scale Example**

**Attitudes of Music Educators toward the Teaching of Language Reading Skills in the Music Classroom.**

**Begin Here.** Please respond to each of the following questions about the teaching of language reading skills in the music classroom. Circle the appropriate number from the following scale that accurately expresses your level of agreement.

1 = Very Strongly Disagree (VSD)
2 = Strongly Disagree (STD)
3 = Disagree (DIS)
4 = Agree (AGR)
5 = Strongly Agree (STA)
6 = Very Strongly Agree (VSA)
? = Do not know (DNK)

<table>
<thead>
<tr>
<th>Example</th>
<th>VSD</th>
<th>STD</th>
<th>DIS</th>
<th>AGR</th>
<th>STA</th>
<th>VSA</th>
<th>DNK</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. The directions to this questionnaire are easy to follow.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>?</td>
</tr>
</tbody>
</table>

By circling the number 6, you are indicating that you very strongly agree with this statement, thus believing that the directions to this questionnaire are indeed easy to follow.

**Level of Agreement**

(Circle your response)

<table>
<thead>
<tr>
<th>VSD</th>
<th>STD</th>
<th>DIS</th>
<th>AGR</th>
<th>STA</th>
<th>VSA</th>
<th>DNK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>?</td>
</tr>
</tbody>
</table>

1. I teach language reading skills in my music classes.

2. My teacher preparation program did not provide instruction on how to teach language reading skills through musical content.

3. Every teacher, regardless of content specialty, should teach language reading skills.

626
Level of Agreement
(Circle your response)
VSD STD DIS AGR STA VSA DNK

4. My school district offers professional development training for teaching language reading skills in all content areas.

5. I believe that I have adequate knowledge of how to teach language reading skills in my music classes.

6. District administrators would encourage me to teach language reading skills in my music classes.

7. I am familiar with specific tools and strategies students can use to help with decoding and comprehending text.

8. Language reading skills should be taught in every class.

9. I am uncomfortable with teaching language reading skills in my music classes.

10. My principal supports the teaching of language reading skills in all content areas, including music.

11. Language reading skills cannot be taught in performance classes like band, choir, or orchestra.
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12. Language reading is a skill students need to be successful in my class.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>13. I would participate in district sponsored professional development training geared at increasing my knowledge of how to teach language reading skills in my music classes if offered in the future.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>14. I believe language reading skills can be taught in a general music class.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>15. I am unwilling to periodically suspend music instruction and devote a portion of my class time to the teaching of language reading skills.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>16. Language reading skills should only be taught by a language arts or reading specialist.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>17. I have consulted with other music specialists on how to best integrate the teaching of language reading skills into my music class.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>18. I would teach more language reading skills in my music classes if I were provided training.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>19. Students will benefit musically when language reading skills are integrated into the music curriculum.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
Appendix E

Example Letter/Form for a Panel of Experts to Establish Content Validity

Dear Dr. __________:

First of all, thank you very much for your participation in serving as one of the panel experts for examining the content validity of this instrument. The purpose of this study is to identify the roles of intangible resources associated with the management of agri-tourism enterprises. In this study, intangible resources are divided into two categories. They are competency-based and asset-based. Competency-based intangible resources were defined as collective attributes including knowledge, abilities, skills, and experiences that helped an agri-tourism enterprise gain sustainable competitive advantage (e.g., employee know-how, organizational culture, and organizational learning). Asset-based intangible resources are defined as long-lived assets without physical substances that are used in business (e.g., intellectual property rights, databases, and networks).

On the following pages, statements are included to examine competency-based intangible resources. Please rate each statement on two criteria: (1) the appropriateness of the statement in representing the competency-based intangibles, and (2) the clarity of the meaning of the statement. Please circle your response.

(1) Is this statement appropriate?
   Yes = Appropriate
   No = Not Appropriate

(2) Is the statement clear?
   Yes = Meaning Clear
   No = Meaning Unclear

If the statement is appropriate but unclear, please reword the statements on the blank line below the statement.

If you would like to add statement(s), please write them/it down.

If deleting statement(s) is necessary, please indicate the statement(s) on the blank line below the statement.

If you have any comments and recommendations concerning the statements, please feel free to provide your precious opinions on the blank line below the statement. Thank you very much.

**Competency-based Intangible Resources** (defined as collective attributes including knowledge, abilities, skills, and experiences that helped an agri-tourism enterprise gain sustainable competitive advantage):

A successful agri-tourism enterprise is able to:
1. Establish quality standards.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

2. Manage change.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

3. Carry out tasks in situations of uncertainty.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

4. Innovate.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

5. Provide quality services.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

6. Recognize the needs of customers.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

7. Build enterprise culture.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

8. Build enterprise traditions.

<table>
<thead>
<tr>
<th>Appropriate?</th>
<th>Clear?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

[etc., for all proposed items]
Measuring in the Affective Domain
Larry Miller and Kattlyn Wolf, OSU

For this seminar, attitudes will = perceptions, = opinions; = beliefs

Thus, we will consistently use the word “attitude” in our presentation.

Defining an Attitude

- “An attitude is the affect for or against a psychological object” Louis Thurstone, 1931
- evaluation of
- like or dislike of
- positiveness or negativeness toward
- (Do not say “feelings”)
Assumptions

- Attitudes are predispositions to respond.
- Attitudes are persistent over time.
- Attitudes are susceptible to change, but not easily.
- Attitudes produce consistency in behavior.
- Attitudes are directional (Summers, 1970).

An attitude is like a puff of smoke that we wish to describe. To do so, we must sample many molecules in the puff if we wish to describe it adequately. One molecule will not do!

Types of Attitudinal Scales

- Likert (summated rating)
- Semantic Differential

We will discuss these two, most common, scales but additional information is provided in the handout should you have an interest in the latter two.

- Guttman (cumulative)
- Thurstone (equal-appearing interval)
General Criteria for Attitude Statements

- Items should be a series of statements; not questions.
- Avoid statements that refer to the past rather than to the present.
- Avoid statements that are factual or capable of being interpreted as factual.
- Avoid statements with multiple interpretations.
- Avoid statements that are irrelevant to the psychological object under consideration.
- Avoid statements that are likely to be endorsed or not endorsed by almost everyone.
- Select statements that are believed to cover the entire range of the affective domain of interest.
- Keep the language of the statements simple, clear, and direct. Avoid words that may not be understood by the respondents.

Criteria for Attitude Statements (continued)

- Statements should be short; rarely exceeding 20 words.
- Each statement should contain only one complete thought.
- Statements containing universals such as all, always, none, and never often introduce ambiguity and should be avoided.
- Words such as only, just, merely, and others of similar nature should be used with care and moderation.
- Statements should be in the form of simple sentences rather than compound or complex sentences.
- Avoid the use of double negatives.
- Statements should not contain contractions.
- Avoid double-barreled statements: “My principal and my superintendent support my efforts at innovative teaching.”
Likert Scales: Advantages (summated)
Rensis Likert, 1903–1981

- Easy for respondents to complete, most people familiar with the scale
- Relatively easy to construct
- Most popular attitudinal measure
- Easy to score and analyze
- Each item considered to be of equal attitude value (weight) -- homogeneous items

Likert Scale Construction

- Identify the attitudinal object and delimit it quite specifically.
- Compose a series of statements about the attitudinal object that are half positive and half negative and are not extreme, ambiguous, or neutral.
- Establish (a minimum of ) content validity with the help of an expert panel.
- Pilot test the statements to establish reliability (Cronbach’s alpha) for each domain.
- Eliminate statements that negatively affect internal consistency.
- Construct the final scale by using the fewest number of items while still maintaining validity and reliability; create a balance of positive and negative items [Remember to reverse-code when summing].
- Administer the scale and instruct respondents to indicate their level of agreement with each statement.
- Sum each respondent’s item scores to determine attitude.
Likert Scale Instrument Construction

- Use the general criteria for attitude statements.
- Begin with non-threatening, easy items first; demographics last.
- Have clear instructions with an example.
- Anticipate data entry and analysis.
- Anticipate missing data on items.
- Use approved layout techniques (see Dillman).

Scaling of Statements

Response scales vary. Recommend to use an even number of response categories (no neutral category) and a N/A response for agreement scales. Label all response categories.

Since this is a summated rating scale, the scale of measurement of the sum or mean is interval. Never analyze by item. Scale of measurement of any one item is ordinal.

Anchored scales: frequency, importance, etc. (Odd # = OK)
Pictures, thermometers, etc., may be used as scales
Multiple scales per item may be used.
Greater range in the scales produce more variability in the data: 8 better than 6, 6 better than 4, etc. (Correlations work better.)
Likert Scaling

- Even Number of Response Categories
- Label all categories
- Use N/A if appropriate [No neutral/undecided]
- Frequency, Importance, etc. [Anchored]

Example: Multiple Anchored Scales
For each Competency, please rate (1) the level of importance of it to your work and (2) your current level of knowledge.

<table>
<thead>
<tr>
<th>Level of Importance</th>
<th>Level of Competency</th>
<th>Level of Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4</td>
<td>Tomato Varieties</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>1 2 3 4</td>
<td>Tomato Wilt</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>1 2 3 4</td>
<td>Pruning Suckers</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

(Etc.)

Educational Need = (I – K) (Mean I)
Produces 3 measures: I, K and Need with 2 measured
Can measure more.
Certainty Method

- Why consider it? Low variability in data with typical scales
- Procedure: Two responses from subjects:
  (1) agree/disagree and
  (2) level of certainty of response.
- Semantic space between categories is considered and weighted.
- Result: Greater variability in data.

Certainty Method

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The response categories, then, are transformed with scores below for transformed values, and scores below that being the "Certainty Scores".

D5  D4  D3  D2  D1  D/A  A1  A2  A3  A4  A5
1  2  3  4  5  6  7  8  9  10  11
0  3  5  6  7  8  9 10 11 13 16
(Or, other values assigned between points)
Semantic Differential

- Originally created by Osgood, et al., and not specifically designed, initially, for the purpose of attitude measurement.
- Has three primary domains (factors): (1) evaluation, (2) potency, and (3) activity.
- Uses pairs of opposite adjectives in describing a particular object/concept.
- For purposes of attitude measurement, a special form of the semantic differential consisting of entirely adjective pairs representing the evaluation dimension may be constructed.

Some Paired Adjectives for the Evaluation Dimension (See Book for more)

- Good – Bad
- Nice – Awful
- Beautiful – Ugly
- Pleasant – Unpleasant
- Clean – Dirty
- Sweet – Sour
- Sacred – Profane
- Valuable – Worthless
- Kind – Cruel
- Honest – Dishonest
- Fair – Unfair
- High – Low
- Fragrant – Foul
- Successful – Unsuccessful
- Reputable – Disreputable
- Tasteful – Distasteful

(Osgood, 1957)
Semantic Differential Procedure

- Identify the concept to be measured
- Generate a list of approximately 7 or 8 bipolar adjectives with an number of positions between each pair. (Subjects lose focus after 8)
- Administer the scale and instruct respondents to identify where, on the continuum between the two adjectives, their beliefs about the concept lie.
- The spaces or positions between the adjectives become categories with a numerical value (e.g. 1=unfavorable and 6=favorable) and responses are summed to determine attitude.

GRADUATE SCHOOL ADMISSION (Concept)

(Adjective Pairs with scaling)
Rough ___: ___: ___: ___: ___: ___ Smooth
Fair ___: ___: ___: ___: ___: ___ Unfair

(Add other sets of bipolar adjectives as desired)
**Strengths**
- Easy to construct
- Short and quick to administer
- Usually highly reliable
- Correlate highly with scores from Likert or Thurstone scales.

**Weaknesses**
- Long directions for administration can be a problem
- Blatant and transparent in their purpose and can invite bias or socially acceptable responses
- May fail to isolate the evaluative dimension.