Introduction to Scales

PSI's Core Values
Bottom Line Health Impact * Private Sector Speed and Efficiency * Decentralization, Innovation, and Entrepreneurship * Long-term Commitment to the People We Serve
Introduction to Scales

PSI Research & Metrics
2010

Contact Information

Hibist Astatke¹, Kate O'Connell², and Linda Kaljee³

1. Senior Researcher, PSI/Washington
2. Deputy Director, PSI Research & Metrics
3. Associate Professor, Wayne State University, Pediatric Prevention Research Center

For more information, please contact:

Hibist Astatke
1120 19th St. NW.
Washington, DC 20036

Telephone 202-572-4623
Email 202-785-0120
LEARNING OBJECTIVES

By the end of this chapter, the reader will be able to:

1. Understand what multi-item scales are and why we use scales in TRaC surveys
2. Understand the basic principals of scale reliability and validity
3. Understand differences between multi-item scales and indices

BACKGROUND

The following chapter introduces definitions and concepts that are important to understand before you proceed with scale development. Explanations are provided on what multi-items scales are and why PSI uses them. Reliability and validity of multi-item scales as well as methods for assessing them are introduced. Finally differences between multi-item scales and indices are discussed.

MULTI-ITEM SCALES AND TRaC SURVEYS

WHAT ARE MULTI ITEM SCALES?

Multi-item scales consist of a number of statements or items that can be used to measure variables. The variables they measure are not directly observable (they can not be seen) and can change over time. So while it is possible to directly observe some variables like behavior or condom availability in a store or hotspot, a theoretical variable such as attitudes towards condom use can not be directly observed. Such variables are often referred to as ‘latent variables’ (DeVellis, 2003). In TraC surveys, these theoretical or latent variables are determinants of behavior. Examples of determinants that can be measured by multi item scales include attitudes towards condom use with trusted partners, social norms towards using mosquito nets in a community, self-efficacy for condom use, or social support for family planning.
WHAT CONSTITUTES A MULTI-ITEM SCALE?

Multi-item scales consist of a related set of statements which measure a single variable or determinant. There are at least three defining characteristics of multi-item scales (Spector, 1992).

1. The scale is comprised of a minimum of three items. Less than three items would be analyzed as individual statements.
2. The scale measures a variable/determinant that is *continuous* and responses can range from very positive to very negative.
3. The scale provides response options designed to capture a broad spectrum of opinions, attitudes, and/or beliefs.

EXAMPLE OF MULTI-ITEM SCALES

Condom Use Attitudes Scale (Otto-Salaj, et al. 2001)

1) Using condoms can make sex more stimulating.
2) Using condoms is uncomfortable.
3) Using condoms is pleasurable.
4) Using condoms ruin the "mood."
5) Using condoms interrupt foreplay.
6) Using condoms is unreliable.
7) Using condoms is an effective method of preventing sexual diseases.

* Modified scales from the PSI Scales bank

WHAT IS A LIKERT RESPONSE SCALE?

In order to measure a continuous variable/determinant with a range of options, respondents are provided with a set of statements which they are asked to rate on a ‘*Likert scale*’. A Likert scale includes a set number of options which correspond to a range of positive to negative responses to a single statement or item. Responses can be scored across a three, four, or five point Likert scale. Some studies have used a 10 point Likert scale.
EXAMPLE OF LIKERT SCALES

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Using condoms can make sex more stimulating.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

The advantage of using a Likert scale is that it enables researchers to measure the strength or magnitude of variables that may differ across populations. For example, sex workers may agree with the statements related to effectiveness and reliability of using condoms but not items related to comfort or pleasure. Likert scale responses to all items in a single variable are combined or summed into a composite score. A respondents’ score on a multi-item scale is caused by or is a reflection of the underlying latent variable. For example, a high score on a condom use attitude scale will reflect that person’s positive assessment or evaluation of using condoms.

EXAMPLE OF MULTI-ITEM SCALE RESPONSES AND COMPOSITESCORES

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Using condoms can make sex more stimulating.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2) Using condoms is uncomfortable. R</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3) Using condoms is pleasurable.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4) Using condoms ruin the &quot;mood.&quot; R</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5) Using condoms interrupt foreplay. R</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6) Using condoms is unreliable. R</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7) Using condoms is an effective method of preventing sexual diseases.</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

R=items that need to be reverse coded so that a high score equals a positive attitude about using condoms

composite score=20/7=2.86


*This scale has been modified for use by PSI.*
**WHY SHOULD WE USE MULTI-ITEM SCALES?**

Using several items allows for the measurement of complex concepts such as the determinants found in PSI’s Behavior Change Framework. These determinants cannot be easily and accurately measured by a single statement or question. Likert scales enable a level of agreement or disagreement with a statement, which allows for differing opinions to be expressed by respondents as compared to two-option responses, e.g., ‘agree’ or ‘disagree’. Determinants as outlined in the behavior change framework are by and large better addressed by scaled items that enable variability between responses. The characteristics of multi-item scales allow for improved *scope* and *precision*.

The variety of statements used to get at different aspects of a concept is referred to as the *scope* (Spector, 1993). For example, attitudes about using condoms can include assessments or evaluations in several areas such as comfort, pleasure and effectiveness. If a large number of slightly variant items was used to capture all aspects of the latent variable “condom use attitudes”, the scope of the scale would be considered wide.

**EXAMPLE OF DIFFERENT ASPECTS OF CONDOM ATTITUDES (SCOPE)**

1) Using condoms can make sex more stimulating.
2) Using condoms is uncomfortable.
3) Using condoms can be pleasurable.
4) Using condoms can ruin the "mood."
5) Using condoms interrupts foreplay.
6) Using condoms is unreliable.
7) Using condoms is an effective method of preventing STIs

The fact that scales have different response options and use several items to measure a concept allows for a larger variation in terms of the choices that respondents have and thus more *precision* in responses. For example, imagine you are asked to agree with the statement “It is easy to use condoms with my regular partner.” If you are given two answers “agree” or “disagree” neither of these responses may truly reflect your thinking. Perhaps you strongly agree with this statement or only slightly agree. Having more response options allows you to answer more precisely. Precision is also improved by having multiple items in a scale.

Finally, multi-item scales with better scope and precision produce responses that are more reliable or consistent over time (Spector, 1992). Those scales with good precision and scope do
not allow respondents a lot of room to be unsure or undecided about their answer, or have opinions that are not addressed by the available items.

**WHY ARE WE USING SCALES TO MEASURE DETERMINANTS IN THE PSI BEHAVIOR CHANGE FRAMEWORK?**

We use multi-item scales to measure determinants because it allows us to examine and improve the **reliability** and **validity** of our measurement tools. This increases our chances of detecting potentially powerful but subtle determinants of behavior. Using multi-item scales also improves the usefulness of results from segmentation tables (see ****). TRaC studies conducted using multi-tem scales allow us to identify precise and in depth information about determinants of behavior. This in turn allows programmers to address determinants of behavior about which they have detailed information. The use of scales also allows for the testing or confirmation of relationships proposed by the PSI Behavior Change Framework.

**BASIC PRINCIPLES OF SCALE RELIABILITY AND VALIDITY**

**WHAT IS A RELIABLE MEASURE?**

A reliable scale captures the latent variable with relatively little error. Reliability is the amount of variation (the spread of a distribution of scale score) in a set of responses due to respondents’ “true” scores on a latent variable. (Devellis, 2003). The “true” score represents respondents’ scores on a latent variable without any random measurement error related to problems with the items, the scale development procedures, or personal factors such as mood. (Viswanathan, 2005). Scales with low random measurement error have a higher amount of variance that can be attributed to the true score and thus are more reliable.

**WHAT IS A VALID MEASURE?**

Validity is the extent to which a scale measures the variable that it is meant to measure. When we ask if something is valid, we want to know if our scaled variable is really measuring what we think it measures. For example, how do we know that the scale in the box below is really measuring attitudes about using condoms, and not something else, such as beliefs or outcome expectation about using condoms. If a scale is valid then we can be confident that any changes in respondents’ scores are caused by changes in the underlying variable being measured.
EXAMPLES OF RELIABILITY AND VALIDITY OUTSIDE THE RESEARCH WORLD

Let’s assume I want to measure the reliability and validity of my new bathroom scale. To measure reliability, I will weigh myself two or more times in the span of a short period of time. At the first weighing, I am 64.8 kilos. Two minutes later, I weigh myself again, and the scale still reads 64.8 kilos. This means that the scale is reliable, and I can be pretty sure that if I spend one week eating fried bananas and doing no exercise that any increase in my weight will reflect an actual increase in my body weight and is not the result of a faulty scale.

Now, when I look at the bathroom scale, I realize that 64.8 is actually a measure of my body weight and not my brain size or my level of intelligence. To test this, I go off for a week and eat nothing but deep fried bananas and do no exercise. I come back a week later, step on the scale and see that I weight 65.8 kilos! This tells me that my scale is valid, and I am pretty sure it is measuring what I think it is measuring - body weight - and it is not a measure of intelligence.

HOW DO WE ASSESS RELIABILITY?

There is only one type of reliability; however reliability can be assessed through different methods. Reliability can be determined relatively early in the scale development process. Reliability is particularly important in connection with multi-item scales. It raises the question of whether each scale is measuring a single idea and hence whether the items that make up the scale are related to one another.

Methods to test the reliability of multi-item scales include:

- **test re-test** - Respondents complete the same multi-item scales twice over a short period of time (e.g., one week), and resulting composite scores for each scale are correlated with one another. Note that unlike other forms of reliability, test-retest requires data to be collected at two time points (see Chapter***).

- **multiple form** – Two versions of the same multi-item scales are created, administered at the same time to two groups of respondents, and resulting composite scores for each scale variant are correlated with one another.

- **split-half** – Multi-item scale statements are assigned to two subsets and each respondent completes both scale subsets. The respondents’ composite scores for each scale subset are correlated with one another. Scale items can be assigned randomly, selected, e.g., odd/even numbered items, or balanced based on characteristics of the statements (e.g., length of the statement).

- **internal consistency** – Respondents complete the same multi-item scales and item analysis is conducted to measure the association between the scale items. If items are
strongly related to the variable/determinant they are supposed to measure the items will also be strongly related to each other.

Internal consistency between items is the most common method for assessing reliability. This is also the method that will be used to assess the reliability of scales used in the PSI’s TRaC studies. The result of an internal consistency reliability analysis is presented in terms of a Cronbach’s Alpha score. Cronbach’s Alpha is defined as the “proportion of a scale’s total variance that is attributable to a common source” (Devellis, 2003). Alpha values will tell you how much the items in your scale ask about the same thing. The technical term for this is “common variance”.

Common variance is the variance in respondents’ scores that has a common or single source. This is examined by looking at the correlation between items. The assumption is that if items are strongly related to each other they are also strongly related to the construct they are supposed to measure. When Cronbach’s Alpha is calculated, we look at how much common variance there is between items and how much is specific variance or error variance. Cronbach’s Alpha scores range from 0 to 1. A Cronbachs Alpha of “1” indicates that there is no measurement error and that all of the items have common variance (see Chapter **** for more information on Cronbach’s Alpha). A good to excellent Cronbachs Alpha is a score .80 and above.

**WHAT ARE DIFFERENT TYPES OF VALIDITY?**

Unlike reliability, validity is never proven and the validity of a scale is determined over time. There are also several types of validity: face validity, content, construct and criterion-related (Devellis, 2003, Spector, 1992 and Netemeyer et al., 2003).

- **Face validity** is the extent to which an instrument measures what it appears to measure

- **Content validity** is the extent to which a multi-item scale covers all domains of a variable/determinant based on a review of literature, theory and other measures. In theory, a scale has content validity when its items are a randomly subset of the universe of appropriate items.

- **Criterion-related validity** describes the extent to which an instrument correlates with an established scale or gold standard that it is hypothesized to be related. Whether or not the hypothesized relationship is based on theory is irrelevant. Within this type of validity, there are five major subtypes including:

  - **Concurrent validity** is defined as the correlation between composite score data from a multi-item scale of interest (e.g., social support for abstinence) collected at the same time as a criterion variable (e.g. sexual debut) that is hypothesized to relate to that multi-item scale.

  - **Predictive validity** refers to the correlation between composite score data from a multi-item scale of interest collected before data is collected on the criterion
variable. If found to be correlated the scale is said to have validity for the particular variable it predicts.

- **Convergent validity** is defined as a strong correlation between the composite scores of a multi-item scale of interest and another scale that is supposed to measure the same latent variable/determinant.

- **Discriminant validity** is a modest correlation between the composite scores of the scale of interest and other scales measuring different latent variables/determinants.

- **Known group validity** refers to situations in which certain groups (e.g., populations) have a significantly higher mean composite score on the scale of interest than other groups. In these cases, the criterion variable is categorical rather than continuous.

- **Construct validity** describes the extent to which a multi-item scale’s composite score correlates with other measures’ scores in predicted ways based on a particular theory. Methods used to assess concurrent, predictive and known group validity described above can also be used to assess construct validity when the researcher’s intent is to measure relationships as predicted by a theory.

### How Is Validity of Multi-Item Scales Assessed by PSI

Everyone agrees that validity is important, but how do you assess validity? At PSI, researchers can address some, but not all aspects of validity.

**Face validity** is not useful for all instruments, especially multi-item scales that try to measure a latent variable/determinant without asking about it directly. Items in a scale may not always be self-evident in terms of the relationship between the items and the underlying variable. The purpose of a scale may also be more or less evident to different groups. For example, a researcher who developed a scale may understand the relationships between items and the latent variable/determinant, but research participants or a “cultural” expert may not recognize these relationships. A scale may also be designed to measure one variable but end up measuring a different variable (Deveillis, 2003). For these reason, PSI assessments do not place a high priority on face validity.

**Content validity** can be assessed through a review of items by a panel of experts in the early stages of a study. In addition, results from **exploratory factor analysis** provide information about the dimensionality of scales and provides data about the content validity of scales. Content validity is therefore one type of assessment used by PSI during scale development.

Factor analysis can serve as preliminary evidence of **construct validity.** Data collected using PSI’s TRaC surveys can used to assess construct validity and **concurrent, predictive, and known group validity.**

While the reliability of scales will be assessed for all studies, certain types of validity such as criterion and construct validity will not be conducted for every study. However, face and content
validity can be assessed through a review of items by a panel of experts in the early stages of a study. The assessment of criterion-related validity (concurrent, predictive, and known group) requires the inclusion of additional measures in TraC studies and assessing construct validity requires the use of advanced statistical methods to conduct confirmatory factor analysis. Results from confirmatory factor analysis will provide strong evidence about the validity of a multi-item scale.

**UNDERSTAND HOW INDICES DIFFER FROM MULTI-ITEM SCALES**

**WHAT ARE INDICES?**

Measurement tools that use statement and response options and result in a composite score can be classified as multi-items measures. Such measurement tools are further categorized as scales or indices (Devellis, 2003).

Multi-item scales measure latent variable/determinants that are believed to have a common source or cause. Alternatively, indices are made up of a set of items that can be used to measure latent or observable variables that do not have a common source but can be combined to form a construct (Netemeyer, 2003). For example, scores on an SES index are not believed to be caused or reflect a latent SES variable that exists within individuals. Instead SES scores are made up of different characteristics or categories of information such as income, education, employment status that do not share a common cause but when combined make up a person’s SES. Similarly a set of questions with Likert scale response options for brand personality can be considered an index if it is assumed that the scores on this variable are not caused by one underlying variable but rather result from multiple factors like ease of use, label, and product availability. Because indices do not have an underlying cause or source, they are not expected to exhibit internal consistency. For this reason, reliability is not assessed for indices. In TraC studies, both knowledge and SES are indices.
QUALITY IMPROVEMENT CHECKLIST

The following section highlights the most important information you will need to know to ensure that your scale development is insightful and accurate.

CHECKLIST OF IMPORTANT CONCEPTS, DEFINITIONS AND RATIONALE

☐ MULTI-ITEM SCALES AND TRAC SURVEYS
  o What are three important characteristics of scales?
  o Why does PSI use scales to measure determinants of behavior?

☐ BASIC PRINCIPALS OF SCALE RELIABILITY AND VALIDITY
  o What is a reliable scale?
  o What is a valid scale?
  o How are the reliability and validity of a scale assessed?

☐ INDICES AND SCALES
  o What is the difference between scales and indices?
  o Why is reliability assessed for scales but not indices?
REFERENCES


