Developing and Evaluating Self-Report Scales

Sally Planalp, The University of Utah

1. **To develop**: Define your concept clearly, preferably both conceptually and practically. Consider, above all, what you want to use the survey for. Include subcategories if the concept can be subdivided into components or aspects.

   To evaluate: Decide what concept most closely fits the conceptual framework you are operating in and the uses to which it will be put. Determine if all subcategories are relevant to your purposes.

2. **To develop**: First, generate a broad pool of items that represent your concept exhaustively. Be sure to capture all subcategories and be open to adding new ones if you think of more without going beyond the basic concept. You may want to use experts, novices, or both to generate items. This is ***face validity***. Play around with different wordings. Second, eliminate items that are difficult or impossible to answer, that are obviously redundant (though you can be liberal here), that include words that are not widely understood, or that have any other obvious fatal flaws. Third, consider what general instructions you will give when responding to the set and how they might affect your results. Unless there is strong reason not to, use continuous scales that are symmetric and anchored at both ends, with gradations as fine as you think are reasonable (e.g., strongly agree to strong disagree on a 7-pt. scale). Other types of answers (yes/no, rank orders) are inappropriate for later analyses.

   To evaluate: Look at the items on the questionnaire to see if they tap what you intend to measure. Make sure all the subcategories still work in specific versions. Make sure the instructions are consistent with your use.

3. **To develop**: Randomize the order of items and administer the mega-questionnaire to a sample population that is as similar as possible to the target population, but without going to great effort. Run reliability analysis (especially alpha with each item excluded) and factor analysis (generally, orthogonal with rotation, including getting the correlation matrix).

   To evaluate: Refine items using several statistics:

- Overall alpha, or **inter-item reliability**: The ideal is probably in the .80 range. Lower alphas mean that the items do not cohere in measuring a single construct; higher means that they are too redundant.
- Alphas with each item omitted: Exclude an item if alpha increases substantially when it is dropped. It means that the item does not correlate well with the other items.
- Correlation matrix: Look for items that correlate .95 or higher. They are probably too redundant. You are wasting respondents' time by asking the same thing twice.
- Factor analysis: Factors indicate subdivisions of items that group together. In general, use the Eigenvalue >1 criterion to decide how many factors are sensible. For each item, look at the factor loading for each item in the scale. The loadings indicate how strongly that item represents the factor. Factors are usually labeled based on the pattern of loadings (usually the highest and lowest). You hope for "pure" loadings, i.e., items that load high on only one factor. This is complicated. Maybe just consult or become an expert in statistics.

- Refine your scale by excluding items that hurt the alpha, that have impure factor loadings, or that load on factors that do not represent what you want in your scale.

4. **To develop**: Do #3 again with the refined scale. Report alphas and factor results.

   To evaluate: Look for a final alpha of at least .70 and factors that are interpretable and make sense in light of how the concept was originally formulated. Make sure all the components of the scale are represented. That's ***content validity***.

5. **To develop:** Check test-retest reliability by giving the scale to the same group of people twice at a time interval that optimizes the trade-off between remembering what they said earlier (bad) and changing in the time between filling out the scales (also bad). Run correlations between scales, and even between items on the scale. Hope for an overall correlation of at least .70, preferably higher.

   To evaluate: Look for test-retest reliability of at least .70 under appropriate conditions.

6. **To develop:** Check to see if your scale is similar to other scales it ought to resemble and different from other scales it ought to be different from (even nearly the opposite of). Pick several pre-existing scales, about half of which should be similar (positively correlated) or different (negatively correlated). Report the reasons why you thought they should be similar or different. Administer all scales to the same subjects, and look at the correlations. This is assessing ***convergent*** (similar) and ***divergent*** (different) validity, two versions of ***construct*** or ***discriminant*** validity.

   To evaluate: Look for correlations that are in the direction predicted (positive or negative) and perhaps also of the strength expected (strongly or weakly correlated). If any correlations are in the .90's, consider revising your scale or give up and use the other scale with which it correlates on the grounds that the existing scale is giving you the same information.

   To develop: Find out if your scale corroborates claims that are obviously true, so obvious in fact that if they didn't you would question your scale. This is called ***criterion-related*** or ***predictive validity***. Two common methods are to see if the self-report scale corresponds to the expected behaviors (e.g., extraverts talking more than introverts) or corresponds to known group differences (talk show hosts reporting more extraversion than accountants).

   To evaluate: Make sure the criteria they use to evaluate validity are obvious and sensible and that the results turn out as expected. If not, they may be measuring something different from what the researchers intend.