Why Factor Analyses with Negatively Keyed Items Fail

Kimberly A. Barchard, University of Nevada, Las Vegas
James A. Russell, Boston College

Reference: Barchard, K. A., & Russell, J. A. (2020, October 28-31). *Why factor analyses with negatively keyed items fail.* Poster presented at the Western Psychological Association Annual Convention, San Francisco, CA.

Contact: Kimberly A. Barchard, kim.barchard@unlv.edu
Website: http://barchard.faculty.unlv.edu/examining-opposites/

Why Do Researchers Factor Analyze Positively and Negatively Keyed Items?

Researchers often want to know how many dimensions underlie a given set of items when they are either designing or evaluating a test. When designing a test, researchers might assemble a set of items covering the entire construct of interest and determine how many dimensions underlie that set of items. Then they might create one subscale for each dimension, so that the test can measure several related constructs simultaneously. Similarly, when evaluating a test, researchers often determine the number of dimensions underlying the existing set of items. If the dimensions correspond with the existing subscales, researchers interpret this as support for the use of those subscales.

The most common method of determining the number of dimensions is to conduct an exploratory or confirmatory factor analysis. If two items have a strong relationship, they are likely to load on the same factor. Depending upon the direction of that relationship, the items might have loadings that have the same sign or opposite signs. If their relationship is positive, they are likely to have loadings that are the same sign: both positive or both negative. If their relationship is negative, one item is likely to have a positive loading and the other a negative one.

Subscales generally consist of items that load together on a factor. Typically, items with positive loadings are positively keyed, and items with negative loadings are negatively keyed (e.g., the scores are reversed before adding them to the subscale total). Including both positively keyed and negatively keyed items on each subscale is beneficial for two reasons. The first (and well-known) benefit of including equal numbers of positively and negatively keyed items is that this reduces the effect of acquiescence response bias: If a person agrees with (acquiesces to) all items, they get a moderate score rather than an extreme score. The second (and typically overlooked) benefit of including both types of items is that this allows researchers to capture the full extent of a construct. For example, to measure the extraversion-introversion dimension, researchers should include items representing high extraversion (e.g., I am talkative) and items representing high introversion (e.g., I prefer to work by myself). Items that capture opposite ends of a dimension should have opposite scoring.

The Problem

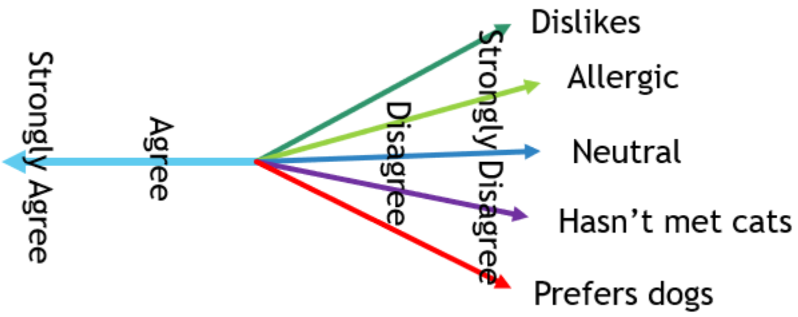
Although test designers typically intend negatively keyed items to measure the same dimensions as positively keyed items, negatively keyed items often load on separate factors or fail to load on any factor at all. Researchers typically conclude that the positively and negatively keyed items represent different constructs (despite this not making any theoretical sense) or that the negatively keyed items are poor measures of the desired construct (for inexplicable reasons). However, both these conclusions may be wrong.

Why Negatively Keyed Items Form Separate Factors

Positively and negatively keyed items form separate factors for two primary reasons. First, most item response scales are ambiguous. Many psychological scales use items with response scales that go from various degrees of agreement (or accuracy) to various degrees of disagreement (or inaccuracy). It is clear what different levels of agreement mean. Someone who *strongly agrees* with the statement "I like cats" in indicating greater affinity for felines than someone who *agrees* with the statement "I like cats." Greater agreement indicates greater liking. However, it is not clear what different levels of disagreement mean, because this item does not specify what the opposite of cat liking is. See Figure 1. Perhaps one person disagrees with this item because they dislike cats. A second disagrees because they are allergic to cats. A third disagrees because they feel neutral. People who disagree may not have much in common, and people who *strongly disagree* might not be more of any particular thing than people who simply *disagree*. Greater disagreement does not indicate greater amounts of any one particular thing. If a response scale assigns positive scores to agreement and negative scores to disagreement, then variation among positive scores may be related across items, but variation among negative scores may not be.

Figure 1

Ambiguous Response Scale



Consider the effect this has on correlations. If two items are designed to measure the same construct and are keyed in the same direction (see Figure 2), positive scores on the two items are likely to be related: More agreement on one item is associated with more agreement on the other item. However, negative scores on the two items are likely to have little relationship: The reasons people disagree with one item may be unrelated to the reasons they disagree with the other item, so that larger negative scores on one item are mostly unrelated to the magnitude of the negative scores on the other item. Therefore, the correlation will be reduced, but it will still positive because half of the response scales have a strong relationship.

On the other hand, if two items are designed to measure the same construct and are keyed in opposite directions (see Figure 3), positive scores on one item should ideally be associated with negative scores on the other. However, people may obtain negative scores for a wide range of reasons. Thus, no part of the response scale on one item has a strong relationship with any part of the response scale on the other item, and the correlation is devastated.

Figure 2

Items Keyed in the Same Direction

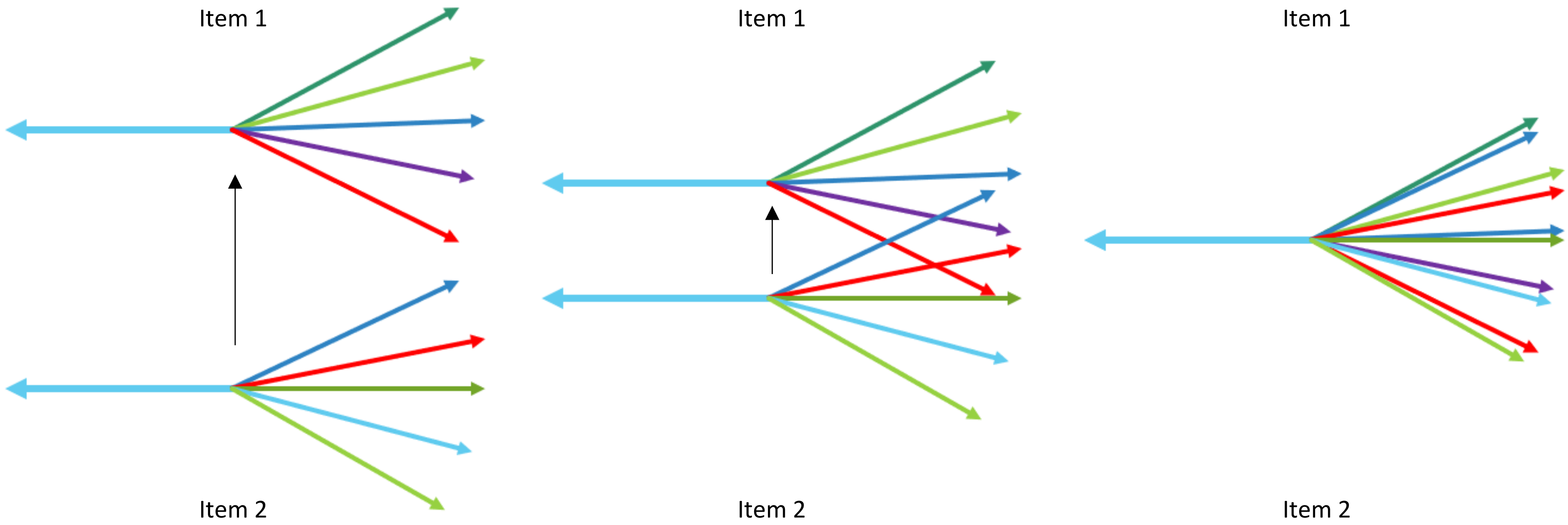
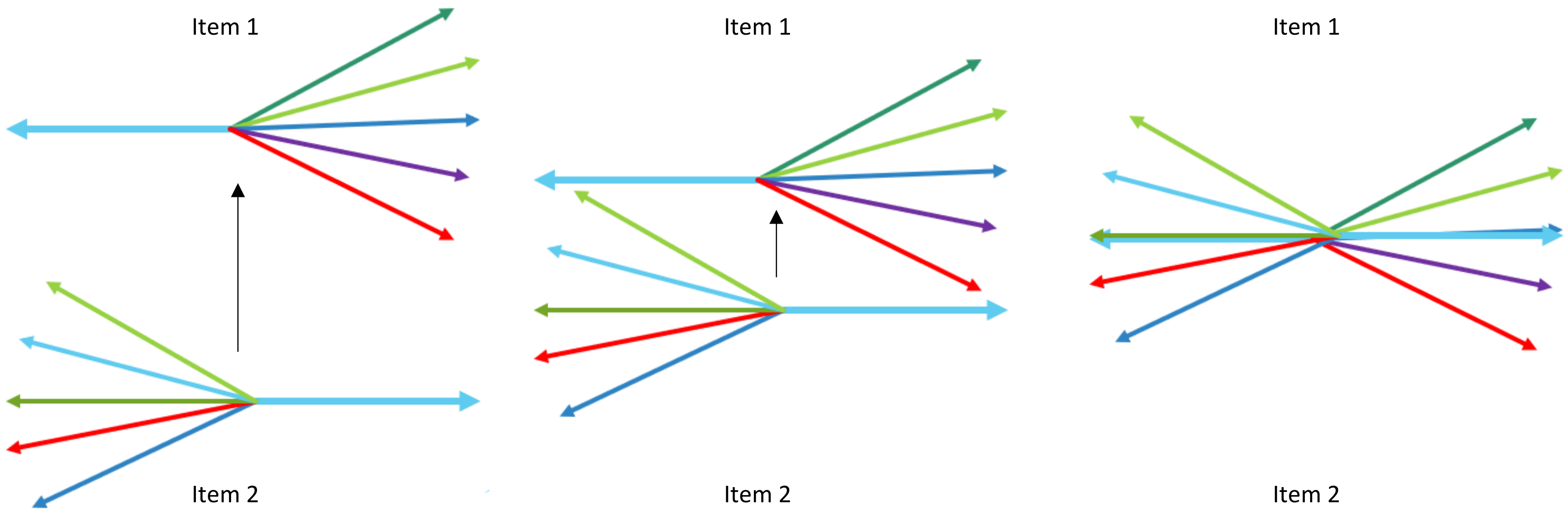


Figure 3

Items Keyed in Opposite Directions



Positively keyed and negatively keyed items often fail to load on the same factors for a second reason. If two variables, X and Y, divide a dimension in half, they cannot correlate -1. For convenience, let scores on that dimension have a standard normal distribution with a mean of 0. In the top half of that dimension, X is positive and Y is 0. In the bottom half of that dimension, Y is positive and X is 0. If there is no measurement error, the scatterplot is L-shaped (see Figure 4), and Carroll (2000) showed that these the two variables correlate -.467. In contrast, if two identical items measure the same half of a standard normal distribution (without any measurement error), they have a straight-line scatterplot and correlate +1 (see Figure 5). Because of this, factors will tend to be composed of items that are all keyed in the same direction.

Figure 4
Two items that measure opposite halves of a bipolar dimension

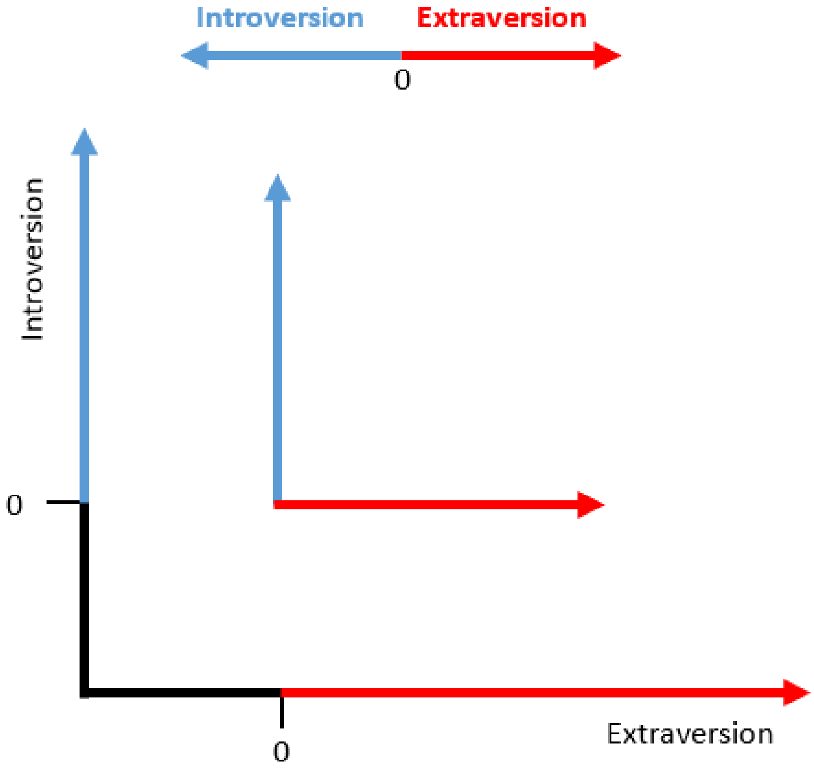
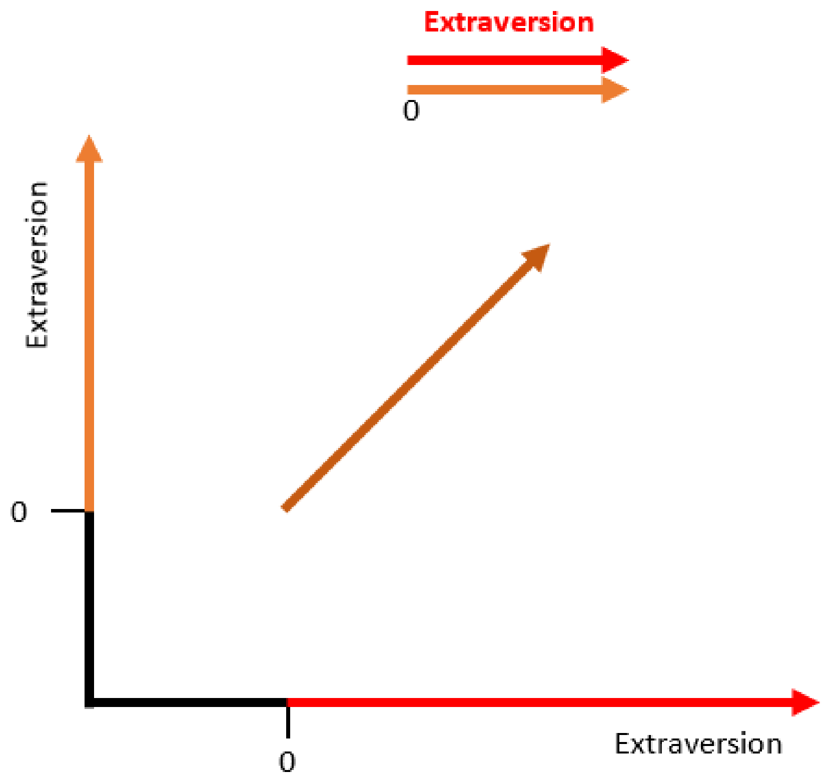


Figure 5

Two items that measure the same half of a bipolar dimension



The Solution

To solve these two problems, we must avoid ambiguity at the disagreement end of response scales, and then we must analyze the data appropriately.

To eliminate ambiguity at the disagreement end of response scales, researchers must either eliminate gradations of disagreement or else specify the meaning of the disagreement end explicitly. There are at least three ways to do this. First, for some variables, test designers may be able to use interval-level variables such as count data (e.g., How many cats do you own?). Such items distinguish among people at only the high end of the variable. Someone who owns four cats may like cats more than someone who owns one cat; however, there are many reasons for people to own zero cats, and so people who own zero cats may vary in their cat-liking. As such, this type of scale is referred to as *unipolar* to indicate that scores measure only a single characteristic: High scores indicate a large amount of that one characteristic and a score of zero indicates a lack of that characteristic.

Second, test designers can use response scales that provide gradations of agreement/accuracy without providing gradations of disagreement. For example,

Do you like cats at all? Yes or no? If yes, how much?
1 = a little, 2 = moderately, 3 = a lot, 4 = extremely

If someone reported that they did not like cats at all, their response would be scored 0. Thus, these items similarly distinguish between people at the high end of the scale without distinguishing between people at the low end of the scale and are thus unipolar. These items have the advantage that they can be implemented in a wide variety of self-report questionnaires, allowing researchers to continue using existing item stems.

Third, test designers can specify the meaning of both ends of the response scales. For example,

I like cats a lot	I like cats a little	I neither like nor dislike cats	I dislike cats a little	I dislike cats a lot

or

I like cats a lot more than dogs	I like cats more than dogs	I like cats and dogs equally	I like dogs more than cats	I like dogs a lot more than cats

Because this type of item explicitly specifies the meaning of the two ends of the dimension, it is referred to as *bipolar*.

After researchers have eliminated the ambiguity of the disagreement end of the response scales, the next step is to analyze the data appropriately. If researchers use bipolar items like the ones above, then the items cover the full length of the dimensions, and the data that results from these items can be analyzed using exploratory and confirmatory factor analysis methods that assume linearity. However, these bipolar items assume the two ends of the response scale are opposites, which makes it impossible to test whether the two ends load on the same factor. Thus, this type of item cannot be used to determine the number of dimensions that underlie a construct.

If researchers use unipolar items like the ones above, then the items only cover the high ends of the dimensions and so oppositely keyed items have non-linear relationships. Therefore, alternative statistical methods are needed to model the relationships between the full underlying dimensions. An effective method of modeling these relationships is to use censored data analysis. When data points are *censored*, researchers have partial information about the values (Fox, 2016): They know that the value is at least as large as some value (e.g., the person's age is at least 55) or no larger than some value (e.g., the concentration is no more than the lower limit of detection of .001).

Many censored data analysis methods exist (e.g., Allignol & Laouche, 2020; Josse et al., 2020). However, most of these methods allow censoring on only one of the variables. Therefore, these methods cannot estimate correlations when both variables are censored and cannot be the foundation for exploratory or confirmatory factor analysis.

Fortunately, Holst and Budtz-Jørgensen (2013) developed a maximum likelihood method that allows censoring on both the X and Y variables. This method has now been implemented in the R package *lava* (Holst, 2020), making it freely available to all researchers. Recent research (Barchard & Russell, 2020b; Holst et al., 2015) shows that the *lava* package provides accurate estimates of the relationships between uncensored variables as long as censoring is only moderate and the assumption of multivariate normality is reasonable. By using strictly unipolar scales and analyzing the resulting data using censored data methods like the ones in the *lava* package, researchers will be better able to determine the relationships between positively and negatively keyed items, allowing them to measure the full breadth of psychological constructs.

References

Allignol, A., & Latouche, A. (2020, June 3). *CRAN Task View: Survival analysis* [List of R packages]. https://cran.r-project.org/web/views/Survival.html

Barchard, K. A., & Russell, J. A. (2020a, June 1-September 1). *I'm less and less happy until finally I'm sad: Estimating correlations when variables divide a construct into parts.* Poster presented at the Association for Psychological Science poster showcase, Chicago, IL. http://barchard.faculty.unlv.edu/examining-opposites/

Barchard, K. A., & Russell, J. A. (2020b, February 9-13). Correlating positively and negatively keyed items: The problem and the solution. Poster accepted for presentation that the Society for Personality and Social Psychology Virtual Annual Convention.

Barchard, K. A. (2020, July). *CensorCorr: Estimating the effect of censoring on correlations (n = 500,000).* [Excel file]. http://barchard.faculty.unlv.edu/research/examining-opposites/

Carroll, J. M. (2000). *The psychometrics of a bipolar valence activation model of self-reported affect* [Unpublished doctoral dissertation]. University of British Columbia.

Fox, G. (2016, May 12). *Introduction to analysis of censored and truncated data* [Recorded workshop]. University of South Florida. https://www.youtube.com/watch?v=aPN10YYrC1M

Holst, K. K. (2020a). *lava: Latent Variable Models (Version 1.6.8)* [Computer software]. https://CRAN.R-project.org/package=lava

Holst, K. K., Budtz-Jørgensen, E., & Knudsen, G. M. (2015). *A latent variable model with mixed binary and continuous response variables.* https://www.researchgate.net/publication/279864661_A_latent_variable_model_with_mixed_binary_and_continuous_response_variables

Josse, J., Tierney, N., & Vialaneix, N. (2020, June 9). *CRAN Task View: Missing data* [List of R packages]. https://cran.r-project.org/web/views/MissingData.html