A Crash Course in Factor Analysis

Michael J. Culbertson

September 6, 2013

ISTEM Professional Development Series
Outline

- What is factor analysis (FA) all about?
- A little theory
- What kind of data do I need?
- How many factors do I have?
- What do I do with the results?
- Summary
What’s it all about?

- Constructing measurement scales
- Psychological characteristics: One item isn’t enough!
- Two main functions of FA:
  - Refining measurement scales
  - Validity argument

FA is all about constructing measurement scales.
That is, trying to get at subtle, complex, not-easily-observed psychological characteristics.

Individual items only get at a piece of the characteristics, plus they aren’t very reliable by themselves.

e.g., suppose we’re trying to measure “extroversion.”
We can’t see extroversion directly (though we might be able to see its effects).
We could ask something like “How much do you like being with other people?”
But, isn’t extroversion more complex than that?

So, we want to group a number of items together into a single scale.
Each item reflects a different piece of the characteristic (extroversion), and together they give a better picture.

But, we have to ask: Do these items really tap into the same characteristic? Are we justified in using them together?
Wishful thinking, here, isn’t enough!

FA has two main functions:
- Refining measurement scales (that is, choosing good items that tap into the desired characteristic well)
- Making a validity argument for grouping the items together (FA gives you evidence to support the claim that you can use the scale to measure the characteristic)
We’ll revisit these two functions at the end.
Choosing Factor Analysis

- When to use:
  - New scales
  - Modifying scales
  - Substantially different population
  - Combining scales into one instrument

So, when should you use FA?
Whenever you need to make an argument about using a group of items together to measure a psychological characteristic.
- Creating a new scale
- Significantly modifying an existing scale
  - e.g., rewording, adding new items, changing instructions, reducing length (this is a big one)
- Using a scale with a substantially new population (e.g., new culture/language, age group)
  Items may not work the same way for them as for others
- Combining two scales into one instrument (examining convergent/divergent validity)
Choosing Factor Analysis

- **When to NOT use:**
  - Existing validity work
  - Not combining items into a scale
  - Very little data

When isn’t it necessary to use FA?
- If someone else has already made a validity argument for your scale, no need to reinvent the wheel (unless good reasons, as above)
- If you aren’t going to combine your items into a single measurement of a psychological characteristics, makes no sense to run an analysis that does that
- If you have very little data, e.g., only 10 respondents—not much to work with there.
  (More on this in a moment.)
No need to get scared by theory: We won’t get into intensive math here, just the basic concepts necessary for conducting and understanding FA.

A model is a simplification of reality. Think of a model airplane: It’s much smaller and simpler than the real thing.
We’ll look at what kind of simplification of reality is used in FA.
Factor Model

- Latent characteristics
- Variable in degree
- Could be multidimensional
- Dimension = “factor”
- Items = Weighted combination of factors
- Weights = “loadings”

So, what kind of statement about reality do we think we’re making with FA?
Well, start with some latent (hidden) psychological characteristics.
Something we think is there, but that we can’t observe directly.
E.g., extroversion

Assume that the characteristic can vary (continuously) in degree
A little extroverted (like Gabi), moderately extroverted (like Emily), very extroverted (like Ayesha)
Note, if your characteristic is discrete, you want a different technique (such as clustering or latent class analysis)

Your characteristic could be a single thing or multidimensional
E.g., perhaps extroversion is really made up of three related things: outgoingness, crowd affinity, tolerance for being alone
Dimensions may be related, or may be quite distinct (either is okay, and this is something to be determined empirically)
In FA, each dimension is called a “factor.”

We can’t observe the (latent) factors, so instead, we observe people’s responses to different items.
Each item response is modeled as a weighted (linear) combination of the factors

The weight given to each factor for a particular item is called a “loading.”
The primary task in FA is to find the loadings for all the items.
These loadings are then used to accomplish the two main goals (refining scale, validity argument).
Example

“I like to meet new people at social gatherings.” =
(Mostly) Outgoingness +
(Some) Crowd affinity +
(Very little) Being alone +
Uniqueness

For example, suppose we had an item “I like to meet new people at social gatherings,” and subjects responded with a number from 1 to 7 indicating how much they feel the statement applies to them.
In reality, there are a lot of things that go into the choice of what number the subjects will choose, but the factor model simplifies all of that to say that the number is based on a weighted combination primarily of three things: how outgoing they are, how much they like crowds, and how much they tolerate being alone.
For this item, perhaps the number has mostly to do with outgoingness, some to do with their crowd affinity, and very little to do with how much they tolerate being alone.
Each item also has some “unique” part, specific just to that item. This unique part accounts for all the other things that go into the choice of how to respond to this item.
Other Key Terms

- Factor structure
- Explaining variability
- Commonality
- Extraction
  - Principal axis factoring
  - Principal components

A few other terms that pop up in FA.
First, “factor structure” refers to the pattern of which items depend on (load on) which factors.

The idea of explaining variability.
Some people will endorse an item strongly (e.g., “strongly agree,” 7) and others weakly (e.g., “somewhat disagree,” 2).
We suppose that a large part of why some will endorse strongly and others weakly has to do with the fact that some have a high value for the latent factors and others a low value.
On the other hand, some of the variability in the item response will have nothing to do with the factors, just the item itself (uniqueness).

A key concept is the proportion of the variability of the item that we can attribute to the underlying factors (not uniqueness), referred to as “commonality.”
Commonality (proportion of variability explained by the factor model) is related to the item loadings
(Big item loadings -> lots of variability is explained by the factors;
small loadings -> not much variability can be explained by the factors—it comes from something else, unique to the item)

Since items are related to the same underlying factors, they will be correlated.
The patterns of these correlations depend on the loadings.
So, FA is about analyzing correlation matrices.
We call this “extracting” factors from the correlation matrix.
Popular method: Principal axis factoring (first removes an estimate of the items’ uniqueness)
Alternative: principal components (in my experience, works a bit better when number of subjects is small)
Often will have similar results, but some purists insist on PAF.
Types of Factor Analysis

- Confirmatory Factor Analysis (CFA)
  - Specify which items depend on which factors
  - Estimate loadings, Structure is assumed
  - Check fit or compare candidate structures

- Exploratory Factor Analysis (EFA)
  - Every item loads on every factor
  - Determine structure empirically
  - Generally, start with EFA

There are two types or modes of factor analysis: confirmatory and exploratory.

In CFA, you specify which items depend on which factors—that is, you provide a particular factor structure, and then simply estimate the loadings assuming that the structure is correct.

Good for checking fit if you have a strong reason to believe that your structure is correct, or compare the model fit of several candidate structures.

In EFA, you assume nothing about the structure and let every item load on every factor (even if in a miniscule amount).

Then, you determine the structure empirically by looking to see which items load onto which factors.

Generally, it’s best to start with EFA, even if you have a strong idea about the structure, to provide a check of your assumptions.

Most of the following is about EFA (the part about determining the structure); the rest applies to CFA as well.
Now for the last bit of theory.  If EFA, if we let every item load onto every factor, we run into a problem: How do we know which factor is which?  Remember, the factor model says that the item responses are weighted combinations of the factors, but the computer doesn’t know anything about the actual factors, so it might come up with any linear combination and can’t tell the difference between them.  For example...

But, instead of using the factors themselves, you could use linear combinations (linear combinations of linear combinations are still linear combinations...).

This is called “rotational indeterminacy,” if you want to use big words.  The solution is: First, the computer starts with the factors that capture the most variability.  Then, you “rotate” the factors to achieve “simple structure,” which basically means that items tend to have large loadings on one (or two) factors and small (near zero) loadings on all others.  This makes it easier to interpret the factors, since you only have to look at the small set of items that load on the factor (and not the others).

Rotation
- EFA: Which factor is which?
  - 1) Outgoing, 2) Crowds, 3) Alone
  - 1) Crowds, 2) Alone, 3) Outgoing
  - 1) 50% outgoing/50% crowds
    2) 50% crowds/50% alone
    3) 50% outgoing/50% alone
- Etc...
There are two types of rotations: orthogonal and oblique. Orthogonal assumes that factors have nothing to do with one another (0 correlations). Oblique allows factors to be correlated. Psychological characteristics are almost always correlated, so you should generally use oblique rotations. I typically use “oblimin,” but others probably work equally well.
Outline

- What is factor analysis (FA) all about?
- A little theory
- What kind of data do I need?
- How many factors do I have?
- What do I do with the results?
- Summary
So, what kind of data do you need?
Well, you have to have a number of items related to one or more latent characteristics. These items need to have variability in their responses: If everyone gives the same answer, the item is useless.
It’s best if the items are normally distributed, though this is not strictly necessary. Likert items are, by definition, not normally distributed, but you can generally get a fairly good approximation with 7+ categories.
If very small number of categories (e.g., 2 or 3) or severely non-normal, you can average together small “parcels” (2-4 items averaged together, but make sure the parcels have very good face validity—that is, content obviously very similar).
Alternatively, since FA is about analyzing correlations, you can start from a correlations matrix, instead of the raw data, which means you can compute tetrachoric/polychoric correlations.
If you have missing data, try starting from a correlation matrix computed using pair-wise deletion.
But, if there are too many missing responses, you may run into the problem of the correlation matrix not being positive definite, which kills the calculations.
But, the big question of how much data do you need?
There are rules of thumb out there, but I think they mostly scare people away from doing factor analysis.

Guadagnoli and Velicer (1988) published a simulation study to investigate how well you could recover a factor pattern under various conditions.
Let’s look at some of the results for their weak loading condition (all the items had loadings of 0.4).
Here, N is the sample size (number of subjects),
g² is a measure of overall pattern recovery (they suggest, arbitrarily, 0.01 as the largest acceptable value),
Type I is the proportion of items mistakenly identified as belonging to a factor (when they shouldn’t have),
Type II is the proportion of items mistakenly identified as not belonging to a factor (when they should have).

For a small configuration (3 factors, 36 items), acceptable results with upwards of 200 respondents.
For more items (72 items, still with 3 factors), you can get by with fewer subjects—closer to 150.
For more factors, you need more subjects: over 150 for 72 items, and over 200 for 36 items.
If the loadings are stronger, you can get by with fewer subjects (see paper for details). I generally take 100 subjects as a bare minimum, preferably at least 150.

If loadings are strong and there aren’t too many factors, this should be enough to get something meaningful.

If your sample size is small (and you can’t collect more data), give EFA a try to see if you can make sense of things.

If you can, great!

If not, at least you gave it a try. You’ll have to fall back to other techniques/arguments.
Now, on to the juicy stuff: How do you decide how many factors you have?
First start with the theory behind your instrument: How was it written?
This is only a starting point, and must be verified empirically.
Then, take a look at the eigenvalues of the correlation matrix (provided with EFA output).
Eigenvalues have to do with combinations of items (eigenvectors) that capture the most variability.
The larger the eigenvalue, the more variability the combination captures.
They’re always organized in decreasing order.
Look for a precipitous drop (e.g., 8, 6, 1)—but, you probably won’t find one.
(A drop says that adding a factor won’t capture much more variability.)
Generally don’t take any factors with eigenvalues less than 1 (but this is just an upper limit).
Based on the above, choose a range of number of factors to investigate.
Perhaps one more and one less than your “target” number of factors (or more, if necessary).
Then, extract a factor model solution for each number of factors (e.g., 3, 4, 5).
Let’s take a look at an example of some items about nutrition from a health education needs assessment conducted with university students.

There were 11 items that covered both nutrition attitudes and nutrition behavior, so based on how the scale was designed, we might expect two factors.

Since there are 11 items, we have 11 eigenvalues. The first 3 are > 1, but the 4th is kind of close.

There’s a pretty big drop between the first and second eigenvalue, and somewhat larger drops to get to the second and third, but then the differences get fairly small.

The first 3 eigenvalues cover 52% of the variability in these items.

So, based on this, we might extract 1, 2, and 3 factors—maybe we’ll throw in a 4-factor model, just to see what happens.
How many factors?

- Name the factors
- How do factors change?
- Avoid junk factors
- Which model makes most sense?

For each of the models you extract, look at which items load together onto a common factor:
  Ask “what do these items have in common?” and give them a name.
  It’s nice if this maps onto how the instrument was written, but be prepared for surprises.

Look to see how the item groupings change when you increase or decrease the number of factors.
  What items get smushed together? What items break apart?

When you have “too many” factors, the last factor (or two) tends to be junk (one or two random items).

Based on all the evidence, decide which model (which number of factors) makes most sense.
Okay, let’s look at the nutrition example again.
The item names are on the left. “PNA” refers to the attitude items, and “PNB” refers to the behavior items.
Here are four models—1 factor through 4 factor.
I’ve color coded the loadings to help quickly identify which items load on which factors:
green are > 0.6, blue > 0.4, orange > 0.3.

Since the scale was written as a group of attitude items and behavior items, let’s start with the two-factor model.
Except for the item about impediments to eating balanced meals, all of the attitude items load together on the first factor, as would be expected by the instrument theory.
but, it looks like a couple of the behavior items are mixed in
(being aware of the nutritional content of what you eat and refraining from eating something because of nutritional content)
So, the instrument theory doesn’t quite work—we’ll have to be more flexible in our interpretation.

The eigenvalues suggested a pretty strong first factor (big drop between first and second eigenvalues).
What happens if we go from a two-factor to a one-factor model?
Well, the one-factor model looks pretty much the same as the first factor of the two-factor model.
While this factor may be fairly stable, using a one-factor model would probably throw out
useful information.

So, let’s bump up to a three-factor model.
(Recall that there were three eigenvalues > 1.)
Here we see that the second factor of the two-factor model remains fairly intact (as the third factor).
In the three-factor model, what was the first factor has broken into two:
The new first factor seems to be about diet:
   We have three attitude items about saying it’s important to eat plenty of fruits and vegetables, fiber, and food low in fat.
   Plus an item about how often you eat high-fat foods (reverse coded).
   This item is about behavior, but it loaded with the attitude items because (apparently) the fact that they were all about diet was more important than the distinction between attitude and behavior.
   But, notice how the behavior item has a smaller loading than the attitude items—this is because even though the items are all about diet, the attitude items all ask about “how important is it to you to eat” such-and-such, and the behavior item asks “how often do you eat.”
   They’re related, but not quite the same, so the behavior item doesn’t fit as well with the other attitude items.

The new second factor seems to be about thinking explicitly about what’s in your food:
   We have an item that asks how familiar you are with reading food labels,
   and the two items about being aware of and refraining from eating because of nutritional content.
   The item about impediments to eating balanced meals is thrown in here, too.
   Impediments to eating balanced meals might have something to do with being aware of what you eat, but not a ton.

What’s the third factor about?
   We have a couple of items about coffee and energy drinks together with an item about how often you eat out.
   The coffee and energy drink items might go together, but the fit with eating out is a bit tenuous.
   Perhaps we’d call it something like “excessive eating habits of questionable value”?
   Seems like a stretch.

What about the four-factor model?
The fourth factor is rather suspect from the beginning, since the eigenvalue was less than 1 (though it was close).
The four-factor model looks a lot like the three-factor model, except that the impediment item is split off on its own.
This, combined with the low loadings in the previous models, suggests that the impediment item is simply about something different than all the other items.
I would call the fourth factor a “junk factor”—it just absorbs a single item that behaves differently than all the other items. This doesn’t mean that we won’t ask the impediment item—it might be important to know about what discourages students from eating balanced meals for its own sake. But, we probably won’t include it in our measurement scale, since it doesn’t seem to be tapping into the same thing as the other items.

Okay, what have we seen in this example?
Instrument theory only partially meshes with empirical results.
Instead of attitude and behavior, the major breakdown seems to be “diet” and “explicit awareness of nutritional content.”
We saw how a factor was conserved when we decreased the number of factors, and we saw how a factor broke apart when we increased the number of factors.
We saw what it looks like when an item doesn’t seem to fit, and how such items form junk factors when the number of factors is too large.
We saw a factor with items that didn’t seem to be very cohesive.

So, at the end, we’re supposed to decide which model makes most sense:
What do you think?
Perhaps the three-factor; or perhaps a two-factor with only the relevant diet/awareness items.
A few tips.

It can be easier to see which items are grouped together if you sort the rows in the loadings matrix in descending order by all the items on the first factor and then all the items on the second factor, etc. (Use a rough cut-off of 0.3 for the initial go-round.)

Many times if you have items that are positively worded and items that are negatively worded, these will separate out into two factors. This may have more to do with the wording than with the underlying psychological constructs.

(Also look for other wording and content effects.)

When you have an oblique rotation, you’ll get several different loadings matrices.

Look primarily at the factor pattern matrix.

Finally, note that loadings generally aren’t stable from sample to sample!

(This is one reason to use your best judgment, not strict cut-offs.)

But, the structure should stay pretty similar, if items are good and the sample is big enough.
Okay, so we have selected a model that seems to make sense. Now what?
Well, remember what I said were the two main functions of FA:
- refining measurement scales (choosing good items)
- making validity arguments (justifying item groupings)
First, refining measurement scales. Once you have the number of factors, it’s time to clean them up a bit. Here, we’re trying to choose good items that hang together well (they’re all about the same thing). Often, junk will creep in (e.g., items that don’t load well onto any factor).

Three things to look for:
- Good loadings (generally > 0.4, but don’t be too strict about it)
- Items load onto only one factor (it’s hard to say the sub-scale is about X if some of the items are also about Y)
- Items on one factor should have a cohesive substantive interpretation (ask: which one is not like the others?)

Throw out suspicious items.
Remember: Substantive interpretation always trumps statistics!

One reason I mentioned at the beginning to use FA is to reduce the number of items on a scale.
Here, consider pitching lowest loading items first.
But, always keep an eye on substantive concerns!
Balance domain coverage, avoid items that are too similar in content, Balance positively/negatively worded items, similar wording patterns, etc.
Finally, we can make our validity argument. You want to group items together to say “even though I can’t measure X [extroversion] directly, if I ask all these items, they collectively get at it.” You need an argument (evidence) that the items really do get at X. Well, you can’t prove that the items really get at X (since you can’t observe it). But, you can at least show that the items get at the same thing as each other (whatever that thing is). Then, you use an argument about the face validity (based on item content) to say “that single thing that these items are getting at is probably X, the thing I’m interested in” So, you need: items that have strong loadings on one factor each (they’re getting at the same thing), don’t load onto other factors (no contamination), and are similar in substantive content (cohesive interpretation). Put those together, and you’ve got a validity argument for using your scale to measure X. Now you can use the average of the items as a proxy/surrogate for a measure of X.
1) General sense of how many factors, and reasonable upper limit
2) Use an oblique rotation.
3) Which model makes most sense?

5) But don’t be heavy handed...

If necessary, bolster sub-scale validity argument with Cronbach’s alpha.
How do you report your results?
Basically, give a summary of your choices with a brief justification (e.g., eigenvalues, models compared, etc.)
Mention items removed (and why)
Explain your names for factors
Give CFA model comparison statistics (if applicable)
Show table of loadings for final model