The items you need and the validation processes you will choose all depend upon “what kind of scale you are writing” – you have to decide:

- measuring, predicting, or measuring to predict?
- construct / content?
- quantifying or classifying or multiple classifications?
- target population?
- single scale or multiple subscales?
- want face validity?
- relative/rank or value reliability?
- alternate forms?

Components of a single item…

Item = target construct + systematic error + random error

Systematic error sources
- other constructs
- social desirability / impression management
- asymmetric response scales (e.g., average, good, great, awesome)

Random error sources
- non-singular (double-barreled) items
- response patterns (e.g., answer all 5s)
- inattention / disinterest
Item-writing

Kind of item?
Judgement vs. Sentiment – what they know or what they think?
Absolute vs. Comparative – what want them thinking about?
Preference vs. Similarity – want ranking/selection or values?

Things to consider …
• don’t mix item types too much – confusing to respondents
• consider what you are trying to measure
• consider what you will do with the response values
  • are there “correct answers” or “indicative responses”
  • e.g., ratings are easier to work with than ranks
• consider the cognitive abilities of the target population

Item Writing, cont. – focusing on sentiments…

How many response options?
• binary – useful if respondents have “limited cognitive capacity”
• 5 – the Likert standard (perhaps a historical accident?)
• 7 +/- 2 – based on working memory capacity

Important Issue #1 – middle item ???
• some don’t like allowing respondents to “hug the middle”
• research tells us that, if given an even # responses, they will “hug the middle” and increase error variance

Important Issue #2 – verbal anchoring ???
• some like to have a verbal anchor for all items & other like to anchor the ends (semantic differential) & some also like to anchor the middle
• some research has shown that labels are “less interval” than #s – i.e., the anchors “hurt” getting interval-like data

Important Issue #3 – Item Sensitivity

Item sensitivity relates to how much precision we get from a single item
• Consider a binary item with the responses “good” & “bad” – big difference between a “1” and a “2”
• Consider a 3-response item with “dislike” “neutral” “like” – huge “steps” among 1,2 & 3 – can lead to “position hugging”
• Consider a 5 – response item “strongly disagree” “disagree” “neutral” “agree” “strongly agree” – smaller “steps” among 1,2,3,4 & 5 – should get less “hugging” and more sensitivity
• Greater numbers of response options can increase item sensitivity – beware overdoing it (see next page)
Important Issue #4 – Scale sensitivity

Scale sensitivity is the “functional range” of the scale which is tied to the variability of data values

• Consider a 5-item true-false test – available scores are 0% 20% 40% 60% 80% & 100% -- not much sensitivity & lots of ties

How to increase scale sensitivity?

• Increase #responses/item sensitivity – can only push this so far
• Increase # items – known to help with internal consistency
• both – seems to be the best approach

Consider:
• 1 item with 100 response options (98 +/- 2???) (VAS?)
• 100 binary items (not getting much from each of many items)
• 50 items with 3 options (50-150)
• 20 items with 6 options (20-120)
• 12 items with 9 options (9-108)

Working the #item - #responses trade-off

Important Issue #5 – Item & Scale difficulty / response probability

“What” you are trying to measure from whom will impact “how hard” the items should be…

• obvious for judgment items -- less obvious for sentiment
  • consider measuring “depression” from college students vs. psychiatric inpatients – measuring very different “levels” of depression

“Where” you are measuring, “Why” you are measuring & from whom will impact “how hard” the items should be …

• equidiscriminating math test for 3rd graders vs. college math majors
• identifying “remedial students” math test for 3rd graders vs. college math majors

Validation Process

Over the years several very different suggestions have been made about how to validate a scale – both in terms of the kinds of evidence that should be offered and the order in which they should be sought. Couple of things to notice…

Many of the different suggests aren’t “competing” – they were suggested by folks working in different content areas with different measurement goals – know how scales are constructed and validated in your research area!

Urgency must be carefully balanced with process – if you are trying to gather all the forms of evidence you’ve decided you need in a single or couple of studies you can be badly delayed if one or more don’t pan out…
Desirable Properties of Psychological Measures

Interpretability of Individual’s and Group’s Scores
- Population Norms (Typical Scores)
- Validity (Consistent Accuracy)
- Reliability (Consistency)
- Standardization (Administration & Scoring)

Let’s start with 2 that are very different…

“Cattell/Likert Approach”
- focus on criterion-related validity
- requires a “gold standard” criterion – not great for “1st measures”
- emphasizes the predictive nature of scales & validation
- a valid measure is a combination of valid items
- a scale is constructed of items that are each related to criterion
- criterion-related validity coefficient is the major evidence
- construct validation sometimes follows
- tend to have “limited” internal consistency and “complex” factor structures – these are not selection or validation goals

“Nunnally Approach”
- focus on content validity
- does not require a “gold-standard” criterion
- is the most common approach for “1st measures”
- emphasizes the measurement nature of scales & validation
- a valid measure is made up of items from the target content domain
- internal consistency, test-retest reliability & content validity (using content experts) are the major evidence
- construct validation usually follows

Process & Evidence Approaches
Obviously there are all sorts of permutations & combinations of these two…

One-shot approach … “a good scale is made of good items” so select items that…
- related to the criterion construct (convergent validity)
- don’t relate to non-criterion constructs (divergent validity)
- interrelate with each other
- show good temporal stability
- represent the required range of difficulty/response probability

You might imagine that individual items that meet all these criteria are infrequent. Remember -- the reason that we build scales is that individual items don’t have these attributes, while composites are more likely to have these attributes.

<table>
<thead>
<tr>
<th>Interesting variation – content validation of predictive validity…</th>
</tr>
</thead>
</table>

Range restriction is a real problem when substituting concurrent validation with incumbent populations for predictive validity with applicant populations. Validities of .60-.70 often shrink to .15-.20

Alternative approach is to base predictive validation on “content experts” (SMEs – incumbents & supervisors) – three steps…
1. Get frequency & importance data from SMEs supporting that the content of the performance criterion reflects the job performance domain for that population (O*NET)
2. Get data from SMEs that content of the predictive scale reflects the predictive domain for that population
3. Get data from SMEs that the predictive domain and criterion domains are related

<table>
<thead>
<tr>
<th>Example … Want to use “job focus” to select among applicants, so have to show that it is predictive of job performance.</th>
</tr>
</thead>
</table>

Job performance criterion is validated by having SMEs rate each of several job performance specifics (e.g., “notice whether or not all the bolts are tightened” is frequent and/or important to perform the job successfully.

Job performance predictor is validated by having SMEs rate whether each of items is related to “job focus” (e.g., Does your mind sometimes wander so that you miss details?).

Predictive validity is established by showing that each of the job performance predictor items is related to at least one of the job performance specifics