Building items

Nils Myszkowski, PhD

Outline

Response formats
Content validity
Response biases
Scoring biases
Review

Aims of this session

- At the end of this session, you should be able to:
  - Understand the importance of building valid items
  - Know how to ensure content validity
  - Choose and understand the importance of response formats
  - Understand the importance of bias reduction in psychological testing

Terminology

- The dimension that is measured is called the **construct**.
- The whole measurement tool is called the **scale**.
- To increase accuracy, dimensions are measured through responses on multiple indicators, called **items**.
- If the measurement tool is intended to measure multiple **facets** of the construct, it can be called an **inventory**, and the measure of one of the facets is called a **subscale**.

Outline

Response formats

Content validity

Response biases

Scoring biases

Review

Response formats

Response formats

- Items have to be agglomerated for the test to output an interval variable (remember, the aim is to quantify an amount of something) called a scale **score**.
- Most scoring systems are based on summing the item scores.
- Some are more computationally demanding (notably IRT models), and require a more complicated estimation of the latent construct. The will be briefly covered in the "advanced techniques part".

Response formats

- Summing scores requirements:
  - For scores to be computed, an interval level of measurement (at least) is necessary, and **all the items should have the same scoring format** (because the item scores have to be added).
  - In other terms, when summing all the item scores up, you have to assume that "one point always means the exact same thing for every item".
  - When building a test, you will need to choose one response format and use it for all of your items.
  - **In other words, don't mix different responses formats!**

Response formats

Types of response formats

- Examples of response formats:
  - Multiple Choice Questionnaire
  - True/false
  - Open-ended questions
  - Self-report Likert Scales
  - Forced-choice
  - Time to solve a problem

Outline

Response formats
Content validity
Response biases
Scoring biases
Review

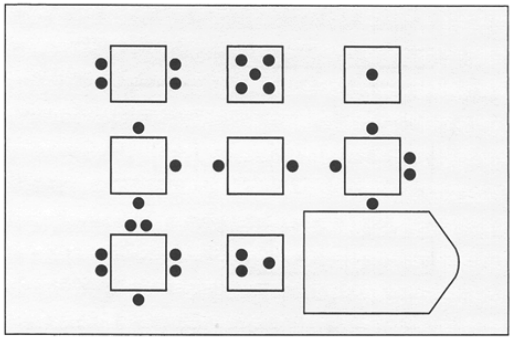
Content validity

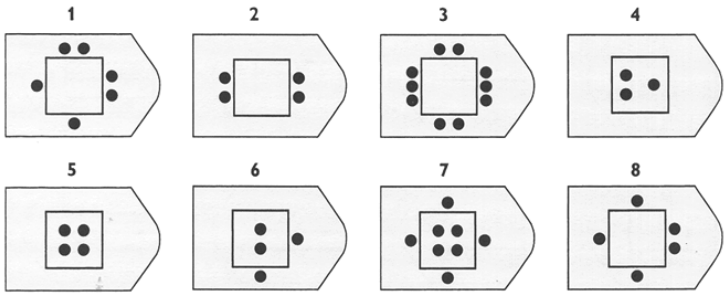
Observable behaviors

- When constructing items, researchers focus on finding observable behaviors that reveal individual differences related to the construct.
  - For example, to build a measure of chronic depression, researchers try to find observable symptoms of chronic depression (loss of appetite, loss of enjoyment in things that were once pleasurable, depressed mood most of the day, suicide thoughts, etc.)

Content validity

- The degree to which the content of the items truly represents the facets of the construct is called content validity.
  - For example, a verbal intelligence test that includes non-verbal matrices as items has bad content validity:





Content validity

Two approaches

- There are two basic approaches to item generation :
  - The **deductive** approach (used the vast majority of the time)
  - The **inductive** approach
- Alternatively, a **combination** of both may be used.

Content validity

Deductive approach

- "This approach requires an understanding of the phenomenon to be investigated and a thorough review of the literature to develop the theoretical definition of the construct under examination. The definition is then used as a guide for the development of items." (Hinkin, 1995)
- In short, it is a **theory-based** approach.

Content validity

Inductive approach

- This approach requires "asking a sample of respondents to provide descriptions of their feelings about their organization or to describe some aspect of behavior."
- These individuals, called **Subject Matter Experts** (SMEs ; Kline, 2005) are supposed to have a valuable expertise in the construct, either because:
  - They belong to the aimed population of the measure (i.e., Students for a measure of the level of stress of students)
  - They are in close contact with the population (i.e., Clinicians that are in contact with patients for a clinical measure)
  - They have academic knowledge in the field (i.e., Personality Psychology researchers for a personality measure)
- Interview content is analyzed and classified —until saturation of the themes —to determine facets and build items.

Content validity

What to be concerned with...

- The primary concern in generating the items is **content validity**.
- Content validity is the extent to which the measure adequately captures the specific domain of interest, and no other content.
- In short, the content of the items must help measure the contruct, and the construct only.

Content validity

What to be concerned with...

- Ensuring content validity can be challenging, as content validity in ensured by ruling out potential measurement **biases** (errors) that can contaminate responses.
- Through this course, we will review the main sources of bias in psychological testing, as well as ways to avoid or reduce them.

Content validity

Content validity analysis

Content validity

- As we explained earlier, content validity has to be maximized when building items.
- In that process, we notably try to reduce biases.
- But we can also assess the content validity of the items produced

Methodology

- One way to do, introduced by Lawshe (1975) so is to list the items, and have them evaluated by Subject Matter Experts (SMEs).
- The judges will read each item and respond for each to the question : "Is the skill or knowledge measured by this item 'essential,' 'useful, but not essential,' or 'not necessary' to the performance of the construct?"

Methodology

- Another way to proceed, possible only for multifaceted constructs, is to ask SMEs to correctly replace each item in their facets.
- We then select items with the highest rates for each facet.

Analysis

- We then analyze the results produced.
- If we attribute 3 to 'essential,' 2 to 'useful, but not essential' and 1 to 'not necessary'...
- A first way to analyse the results is to analyse the location indices (mode, median, mean) for each of the items. (see the next course, which will focus on univariate distribution analyses)
- We would then keep the items that have the highest values for these location indices (highest means, medians, modes).

Analysis

- Another way is to keep items that are judged essential by most judges (which generally produces consistent results with using location indices).
- Typically, the Content Validity Ratio (CVR) is used.

Content Validity Ratio

- Content Validity Ratio (CVR)

$$CVR = \frac{n_{essential} - \frac{N}{2}}{\frac{N}{2}}$$

- $n_{essential}$ is the number of judges who answered "essential for an item
- N is the total number of judges
- $-1 \leq CVR \leq 1 : -1$ means perfect disagreement, $+1$ means perfect agreement
- Positive CVR indicates more than half the judges judging the item as essential.

Example

- Imagine that, for Item 1, 23 out of 30 judges judged the item as essential.

$$CVR = \frac{n_{essential} - \frac{N}{2}}{\frac{N}{2}}$$

$$CVR_1 = \frac{23 - \frac{30}{2}}{\frac{30}{2}}$$

$$CVR_1 = \frac{23 - 15}{15}$$

$$CVR_1 \approx .53$$

Thresholds

- No available threshold value here is really meaningful.
- A higher CVR is obviously desirable.
- As a rule of thumb, we generally want more than half the judges judging an item as 'essential' to include it ($CVR > 0$).

Proposed thresholds

- Lawshe originally proposed critical values, which were later corrected successively by Wilson et. al. (2012), then Ayre & Scally (2014, reported on the next slide).
- They correspond to "the lowest level of CVR such that the level of agreement exceeds that of chance for a given item, for a given alpha (Type I error probability, suggested to be $.05$ using a one-tailed test)
- In other terms, if $CVR > Threshold$, then the CVR is significantly higher than 0 (meaning "significantly more than half of the SMEs judged the item as 'essential'").

Content validity

Content validity analysis

Table 2. Simplified Table of $CVR_{critical}$ Including the Number of Experts Required to Agree an Item Essential.

Panel Size	$N_{critical}$ (Minimum Number of Experts Required to Agree an Item Essential for Inclusion)	Proportion Agreeing Essential	$CVR_{critical}$
5	5	1	1.00
6	6	1	1.00
7	7	1	1.00
8	7	.875	.750
9	8	.889	.778
10	9	.900	.800
11	9	.818	.636
12	10	.833	.667
13	10	.769	.538
14	11	.786	.571
15	12	.800	.600
16	12	.750	.500
17	13	.765	.529
18	13	.722	.444
19	14	.737	.474
20	15	.750	.500
21	15	.714	.429
22	16	.727	.455
23	16	.696	.391
24	17	.708	.417
25	18	.720	.440
26	18	.692	.385
27	19	.704	.407
28	19	.679	.357
29	20	.690	.379
30	20	.667	.333
31	21	.677	.355
32	22	.688	.375
33	22	.667	.333
34	23	.676	.353
35	23	.657	.314
36	24	.667	.333
37	24	.649	.297
38	25	.658	.316
39	26	.667	.333
40	26	.650	.300

Figure: Thresholds of significance of CVR (Ayre & Scally, 2014)

However

- Significance here can give us more certainty that the CVR is positive, and but the really important statistic here is the CVR itself, and N.
  - A CVR being significantly higher than 0 with 40 panelists means only 65% judges decided it is essential, which shouldn't satisfy us. In that case it is advisable to build new items!
  - A CVR significant with 5 panelists means that the CVR is perfect $(+1)$, but at the same time, it was judged by a very limited amount of judges !
- The bottom line is that we want **more panelists** (N), and a **higher Content Validity Ratio (CVR)**, altogether.

Other considerations

- We can also add at this stage that there could be other considerations:
  - Desired length of the instrument
  - Expected psychometric qualities
    - e.g. More items generally lead to stronger internal consistency
  - Quality and severity of the SMEs
    - Biases towards "fun" items, "original" items, items that relate to SME's life, etc.
  - Other items submitted to the SMEs
    - Order effects (contrast effects notably : A "bad" item will make the next one seem better) → Randomized presentation is advisable

Outline

29

Response formats

Content validity

Response biases

Scoring biases

Review

Response biases

30

Response biases

- Social desirability bias
- Comprehension biases
- Response format biases
- Emotional biases
- Applicability bias
- Cultural biases
- Stereotype threat bias
- Non-response bias

Social desirability

- The **social desirability** bias is the tendency of respondents to respond in order to present themselves more or less favorably.
- It is notably a problem of self-report questionnaires.
- Respondents may respond to present themselves favorably either consciously...
  - It's called **other-deception**, or impression management, or **faking**
- ...or unconsciously.
  - It's called **self-deception**.

Context specificity

- Social desirability is very **context-specific**.
- Social desirability can notably be a dramatic issue when responses are public and something is at stake.
  - For example, in recruitment situations or for a legal psychology expertise.
  - But it does not mean that social desirability is not a problem when nothing is (apparently) at stake (in research, for example...).

Construct specificity

- Social desirability is also **construct-specific**.
- The measure of some constructs is notably more problematic than others, because it can be more or less socially desirable to have some individual characteristics.
  - For example, it's highly socially desirable to have high intelligence, high creativity, specific political orientations, etc. Social desirability is less of a problem if you want to know about somebody's locus of control, or how much people like milk shakes.

Solutions

- There is no perfect solution against social desirability biases.
- Typical solutions include:
  - Building **neutral items** with less socially desirable responses
  - Using **forced-choice response formats** with equally socially desirable responses
  - Using **lie scales** (also called social desirability scales) to detect social desirability response tendency (some responses can be invalidated depending on lie scale scores)

Response biases
Social desirability bias

Appendix: The Social Desirability Scale-17 (SDS-17)

Instruction

Below you will find a list of statements. Please read each statement carefully and decide if that statement describes you or not. If it describes you, check the word "true"; if not, check the word "false".

Items

1. I sometimes litter.
2. I always admit my mistakes openly and face the potential negative consequences.
3. In traffic I am always polite and considerate of others.
4. I have tried illegal drugs (for example, marijuana, cocaine, etc.).
5. I always accept others' opinions, even when they don't agree with my own.
6. I take out my bad moods on others now and then.
7. There has been an occasion when I took advantage of someone else.
8. In conversations I always listen attentively and let others finish their sentences.
9. I never hesitate to help someone in case of emergency.
10. When I have made a promise, I keep it--no ifs, ands or buts.
11. I occasionally speak badly of others behind their back.
12. I would never live off other people.
13. I always stay friendly and courteous with other people, even when I am stressed out.
14. During arguments I always stay objective and matter-of-fact.
15. There has been at least one occasion when I failed to return an item that I borrowed.
16. I always eat a healthy diet.
17. Sometimes I only help because I expect something in return.

Note

Answer categories are "true" (1) and "false" (0). Items 1, 4, 6, 7, 11, 15, and 17 are reverse keyed. Item 4 was deleted from the final version of the SDS-17.

Figure: An example of a lie scale. Usually the items are disseminated among the items of another scale.

Response biases

Response biases

- Social desirability bias
- Comprehension biases
- Response format biases
- Emotional biases
- Applicability bias
- Cultural biases
- Stereotype threat bias
- Non-response bias

Comprehension biases

- Items or instructions that are not 100 % correctly understood are more frequent that you think.
- In particular, through the construction of a measure, researchers should watch for specific sources of comprehension biases:
  - Not all respondents understand items or instructions in the same way. (ambiguity)
  - Respondents do not understand them in the way the researcher intended (clear for you but not for the respondents - Watch out notably for figues of speech and double negatives...)
  - Questions are complex (too long, include technical terms)
  - Items are double-barreled. “"How satisfied are you with your pay and job conditions?"
  - Context changes the interpretation.

Response biases

Response biases

- Social desirability bias
- Comprehension biases
- Response format biases
- Emotional biases
- Applicability bias
- Cultural biases
- Stereotype threat bias
- Non-response bias

Response format biases

- Response format may also induce biases.
- In particular, individual response tendencies have an influence on the way individuals respond:
  - **Central tendency bias**
    - Individuals respond using the "middle" responses
  - **Extreme response bias**
    - Individuals respond using extreme responses
  - **Acquiscence bias**
    - Individuals acquiesce or approve anything
    - Usually dealt with by using a balanced number of positively and negatively worded questions (named reversed items)

Response biases

Response biases

- Social desirability bias
- Comprehension biases
- Response format biases
- Emotional biases
- Applicability bias
- Cultural biases
- Stereotype threat bias
- Non-response bias

Emotional biases

- Respondents may be affected by their emotions during the testing process, leading to biases.
- Their responses can be biased by **positive**...
  - Enthusiasm
  - Pride
  - Amusement
- ...or **negative** emotions.
  - Anger
  - Fear of negative evaluation (inhibition)
  - Stress

Emotional biases

- Generally, administration instructions are done so that respondents are in an emotional state that is neutral or positive when being administrated a test.
- Notably, inhibiting situations (for example, taking an intelligence test) are counterbalanced by some form of encouragement in the instructions.

Response biases

Response biases

Social desirability bias
Comprehension biases
Response format biases
Emotional biases
Applicability bias
Cultural biases
Stereotype threat bias
Non-response bias

Applicability

- Items should be **equally applicable and appropriate to any individual** of the aimed population. If not, individuals may respond randomly or not respond, leading to biases.
- Watch out notably for **loaded questions**, which presume something, which not be true for some respondents.
  - For example "I have stopped going at parties." for an introversion scale. Are you sure all respondents went to parties before?

Response biases

Response biases

- Social desirability bias
- Comprehension biases
- Response format biases
- Emotional biases
- Applicability bias
- Cultural biases
- Stereotype threat bias
- Non-response bias

Cultural bias

- Cultural differences may also induce biases, because tests are often made for individuals of a specific culture, and therefore rely on knowledge or skills that are specific to it.
- For example, a typical issue is the **language bias** : Language proficiency may bias performance scores at specific tests.
- As a consequence, when testing influences decisions (and that's usually the case), cultural biases may lead to discriminatory practices.
- A proposed solution to this is the use of "culture-fair tests", which are usually non-verbal and do not include verbal instructions. But it's often difficult to be certain that a test is really culture-fair, and they are never perfectly fair.

Response biases

Cultural biases

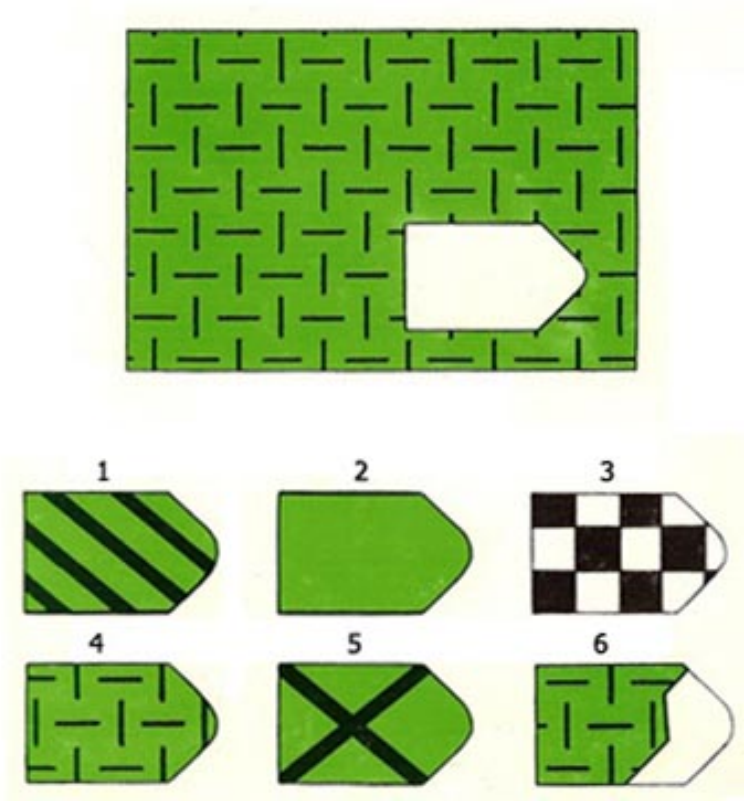


Figure: Non-verbal matrices are sometimes discussed as culture-fair test, as they are totally non verbal and include no instruction.

Response biases

Response biases

- Social desirability bias
- Comprehension biases
- Response format biases
- Emotional biases
- Applicability bias
- Cultural biases
- Stereotype threat bias
- Non-response bias

Response biases
Stereotype threat bias

Discriminatory biases

- When an individual is put in a testing situation that may involve negative stereotypes about their performance in similar tests, they tend to underperform.
- Such reactions depend on the content of the test, the situation, what is at stake, the test taker's personality, etc.
- Thus it is important how the test is presented.
- In addition, it is important to check measurement invariance.

Response biases

Response biases

- Social desirability bias
- Comprehension biases
- Response format biases
- Emotional biases
- Applicability bias
- Cultural biases
- Stereotype threat bias
- Non-response bias

Non-response bias

- Non-response bias occurs when the individuals that do not respond a specific item or a test have different characteristics that individuals who respond.
- For example, individuals with a specific political orientation may avoid responding political questions. The responses are therefore biased, because only part of the aimed population responded, and their responses did not represent the responses of the population.

Outline

Response formats

Content validity

Response biases

Scoring biases

Review

Scoring biases

Scoring biases

Order bias

Consistency bias

Expectancy bias

Order bias

- The order of scoring can also lead to measurement biases.
- In particular, test administrators should watch out for contrast biases
  - For example, a respondent's performance will seem higher if the previous respondent's performance was low.

Scoring biases

Scoring biases

Order bias

Consistency bias

Expectancy bias

Consistency bias

- Scorers tend to seek a consistent view of respondents, which leads to scoring biases.
- Therefore, scorers will usually heavily rely on the first information that they have (called the "anchor") to make their judgment. This is called the anchoring bias.

Scoring biases

Scoring biases

Order bias

Consistency bias

Expectancy bias

Expectancy bias

- The expectancy bias is due to the researcher or the test administrator allowing his or her expectations to - purposely or not - affect the outcome of a test.
- For example, the exact same responses to an intelligence test can lead to different scores, because psychologists administrating tests may expect or anticipate good or bad performance results for various reasons (because of previous test results, or because of stereotypes attached to the socio-economical background of the respondent for example).

Outline

Response formats

Content validity

Response biases

Scoring biases

Review

Review

What you should remember

- Efficiently constructing tests is based on building items wisely and on assessing pre-versions psychometrical qualities.
- Items are generated based on observable behaviors, defined either by theory (deductive approach), by empirical data (inductive approach), or by both.
- Because item scores are summed up, dimensions are measured using the same response format for all the items.

Review

What you should remember

- There is a wide variety of biases, and all of them may jeopardize the validity of your measure.
- There are ways to reduce biases. However, these ways are rarely perfect, and they may sometimes induces other biases (for example, an encouragement to counterbalance inhibition can be taken differently by different respondents - some of them may feel empowered, whereas some of them may be vexed because they think you did not trust their ability! - , leading to even more bias).
- Bias reduction is not only a concern when building measures: It is also very important to be aware of such biases in psychological practice, when administrating tests.

Standardization

- Standardization has much to do with reliability, but it's important to be aware of it from the very test construction, as it also can be a tool to reduce measurement bias (notably the emotional bias and the expectancy bias).
- Standardizing is developing and implementing standards (precise common rules). It means that the items, the instructions to the respondent, the instructions for scoring, and the conditions of administration, etc. should be **as equivalent as possible across the respondents**.
- Another way of saying this : You need to make sure that variations of the scores are only due to variations across participants, and are not due to different conditions, raters, or instructions.

Review

Exercises

Exercise

- Below is the counterproductive behavior checklist. Identify possible biases and imagine what you could do avoid them when measuring such a construct.

10-Item Short Version of the Counterproductive Work Behavior Checklist (CWB-C)

How often have you done each of the following things on your present job?	Never	Once or twice	Once or twice/month	Once or twice/week	Every day
1 . Purposely wasted your employer's materials/supplies	1	2	3	4	5
2. Complained about insignificant things at work	1	2	3	4	5
3. Told people outside the job what a lousy place you work for	1	2	3	4	5
4. Came to work late without permission	1	2	3	4	5
5. Stayed home from work and said you were sick when you weren't	1	2	3	4	5
6. Insulted someone about their job performance	1	2	3	4	5
7. Made fun of someone's personal life	1	2	3	4	5
8. Ignored someone at work	1	2	3	4	5
9. Started an argument with someone at work	1	2	3	4	5
10. Insulted or made fun of someone at work	1	2	3	4	5

Exercise

- Below are different items. Identify the possible measurement biases.
  1. Course objectives were clearly stated at the start of class. (Agree-Disagree)
  2. I believe in acting from the heart. (Agree-Disagree)
  3. I watch the news everyday while having breakfast. (Agree-Disagree)
  4. I think the POTUS should have more power, while the SCOTUS should have less power. (Agree-Disagree)
  5. My instructor is organized and prompt. (Agree-Disagree)