**03b Questionnaire Development**

**1. Need for Instrument**

* Review of existing instruments, used established instrument if available and suitable
* Explain contribution and function of new instrument

**2a. Item Development - Content Validity**

* Define constructs, e.g. reading self-efficacy, so it is clear what will be measured (e.g. reading self-efficacy is…)
	+ Construct is a variable that is “constructed” from responses to multiple items or indicators. Indictors are questionnaire items used to form a construct; indicators provide an indication—a measure—of respondents’ positions on that which is measured.
	+ Example indicators of reading self-efficacy:
		- In general, how confident are you in your abilities in reading?
		- How confident are you that you will do well in reading this year?
		- How confident are you that you can learn to be a good reader?
* Describe theory, if available, of construct – this overlaps with construct dimensions below
* Specify need for non-construct variables, i.e., observable, single-item, and demographic variables
* Identify and define construct dimensions, provide indicators of dimensions, e.g., dissertation process anxiety
	+ Physiological over-arousal (or emotionality): somatic (body, not mind) signs of anxiety and may include headaches, stomach aches, nausea, diarrhea, excessive sweating, shortness of breath, light-headedness or fainting, rapid heartbeat, and dry mouth.
	+ Psychological - Worry: maladaptive cognitions, dread, negative thoughts. Include here catastrophic expectations of gloom and doom, fear of failure, random negative thoughts, feelings of inadequacy, self-condemnation, negative self-talk, frustration, comparing oneself unfavorably to others.
	+ Psychological – Impairment: poor concentration, 'going blank' or 'freezing,' confusion, poor organization. The inability to concentrate leads to impaired performance on tests.
* Develop item pool for construct (items will form a summated rating scale or index) and for non-construct variables
	+ Items should be appropriate for intended population (e.g., use pictures for poor readers like ☺ 😐 ☹ )
	+ Sources of items:
		- Theory, deduction or brainstorming
		- Research examples
		- Questionnaires
		- Expert feedback
		- Population feedback
		- Researcher experience
	+ Each dimension of construct should have separate item pool, enough to measure dimension adequately
	+ Index vs Scale (Instructor note: principal components vs factor analysis)
		- Scale: Items should demonstrate internal consistency, be correlated
		- Index: Items do not have to demonstrate internal consistency; sum of unrelated parts or items
			* SES – income, education, occupational prestige
			* Life Event Index – accumulation of milestones
	+ Response scale options
		- Likert (summated rating scale’ 1 = Very Dissatisfied to 7 = Very Satisfied)
		- Semantic differential (Weak \_ \_ \_ \_ \_ \_ \_ Strong)
		- Checklist (Which have you used to travel to work, check all that apply: [a] car, [b] walk, etc.)
		- Rankings (sort items from most to least important)
		- Multiple choice (which is your biological sex: [a] female, [b] male)
		- Open-ended items (e.g., How did you deal with the electronic harassment?)
		- Thurstone (complex process; many items rated by panel of 100 or more; ratings are from 1 to 11, from least to most positive or similar directions; mean [or median] for each item determined; low variability desired; items with equal distance means [or medians] selected to form 10 item scale)
		- Guttman (items are sorted so agreement with one means agreement with all preceding statements; each progressive statement represents a hardening or sharpening of opinion or knowledge; e.g., [a] 2+2=, [b] 2x2=, [c] 2x\_\_=6, [d] (2/6)^4= ; is deterministic, one can predict responses based upon total score)
	+ For each construct include one global, overall item; e.g., measure of dissertation process self-efficacy: “Overall I am confident I can complete the dissertation successfully” or measure of life satisfaction: “In general I am satisfied with my life.” This item can serve as construct validation for item analysis.
	+ Develop and explain scoring plan for construct formation (Take mean of items 2, 3, 6, reversed 9)
	+ Develop instructions for completing questionnaire
* Critical Item Analysis
	+ Read each item carefully and assess the following:
		- Wording clarity
		- Redundancy within and across dimensions
		- Fit with dimension and construct
		- Fit with targeted population
		- Fit with item scale, e.g., “Your level of satisfaction with current occupation” 1= Very dissatisfied, etc.
		- Remove or revise items as needed
* Draft Questionnaire Format
	+ Title
	+ Brief introduction with general description of questionnaire purpose; be very general, not specific, since this could sway responses
	+ Instructions for completing and submitting questionnaire
	+ If printed, best to use one side of paper, or be sure to include at bottom of page (“See Back” or “Over please”)
* Expert Review – knowledgeable individuals should
	+ Critically review items for the same issues noted above in “Critical Item Analysis”
		- review definitions and dimensions of constructs
		- assess relevance of each item to construct
		- appropriateness of items and questionnaire for target population
		- reading level adequacy of items and questionnaire for target population
		- wording clarity
		- questionnaire format/layout
		- likelihood items may be objectionable to respondents
	+ Edit items and questionnaire as needed
* Pilot Study (Field Test) of Instrument
	+ Use sample of respondents who match target population
	+ Use as large a pilot sample as possible
		- Small sample size – allows for critical feedback about questionnaire and items from respondents who match target participants; include open-ended item at end soliciting critical review and suggestions for revisions
		- Large sample size – allows for
			* Item analysis
			* Reliability assessment (test-retest, internal consistency, equivalent forms)
			* Rater Agreement
			* Validity assessment (predicted differences, correlations, etc.)
			* Critical feedback from participants; include open-ended item at end soliciting critical review and suggestions for revisions
	+ Make revisions based upon pilot study results, if revisions substantial, pilot test again

**2b. Writing Items**

Below are a few suggestions for writing clear items (Crocker & Algina, 1986)

* Use as few items as possible to obtain valid scores; longer questionnaires reduce response rate
* Items should have one interpretation
	+ Example of recommendation letter sentence:
		- Poor: “You would be lucky to get him to work for you.”
		- Better: “His work performance is inconsistent and incomplete; rarely does he finish a task.”
* Avoid items to which most respondents agree or disagree because reduces item discrimination (the ability for an item to distinguish respondents on the concept measured)
	+ Example
		- Most agree: “When people need help after facing devastating natural disasters, someone should be there to help them.”
		- Differences emerge: “When people need help after facing devastating natural disasters, a national government insurance policy supported by an income tax rate increase of 1.25% should be required.”
* Have a few items that are reversed to help prevent response set (marking items without carefully thinking about each item)
	+ Example
		- I can learn the most challenging statistical analysis procedure taught in this class.
		- I believe I will perform well on statistical-related test items in this class.
		- The more difficult statistics become in this class, the less certain I am in learning those statistics. (Reversed response likely)
* Items should be as short as possible
* Avoid complex sentences with multiple segments, “if” or “because” links, etc.
	+ Example
		- If p ≤ α reject Ho, otherwise fail to reject Ho
* Items should be correct grammatically unless specific idiom or vernacular is intended.
* Items with absolute or indefinite qualifiers can create ambiguity or uncertainty of meaning
	+ All, Always, None, Never
	+ Only, Just, Merely, Many, Few, or Seldom
	+ Example
		- Poor: I am always washing my hands
		- Better: If possible, I wash my hands before eating
* Use vocabulary that can be understood easily by respondents
	+ Reading level checker can be helpful, e.g.
		- <https://www.webpagefx.com/tools/read-able/>
		- <https://readable.io/>
	+ Example
		- Grade Level = 19:
			* “Some people have confidence in mathematics and some do not; statistics is based upon mathematics but also relies on logic and some folks have trouble with logic as well; how confident are you in your ability to learn complex statistics in this class?”
		- Grade Level = 10:
			* “How confident are you in your ability to learn complex statistics in this class?”
		- Grade Level = 6:
			* “Do you think you can learn statistics in this class?”
* Avoid use of negative (e.g., not, none, never)
	+ Double negatives really create the problem, although negatives can as well. Ok to use negatives with some items, but always check for clarity.
	+ Examples
		- Confusing: “I am not confident that I cannot learn statistics in this class.”
			* Response scale: Not true of me ---- Very true of me
		- Confusing: “I am not confident that I learn statistics in this class.”
			* Response scale: Not true of me ---- Very true of me
		- Clearer: “I am confident that I can learn statistics in this class.”
			* Response Scale: Not true of me ---- Very true of me
		- Clearer: “Learning statistics in this class is difficult for me.”
			* Response Scale: Not true of me ---- Very true of me
* Items should focus on one construct (i.e., unidimensional); do not use double-barreled items
	+ Examples of poor items
		- “Schools that perform poorly several years in a row should be closed and their teachers fired”
		- “I have the competence to work effectively and can influence the way work is done in my department”

**2c. Increasing Response Rate**

* Short questionnaires – the number of items should be as few as possible because shorter questionnaires tend to have higher response and completion rates (Rolstad, Adler, & Rydén, 2011)
* Wording below are quotations from Edwards et al (2002) who performed a meta-analysis to identify factors related to response rates for postal surveys (and results apply for other surveys too); table on page 3 is particularly helpful.
	+ “The odds of response were more than doubled when a monetary incentive was used (odds ratio 2.02; 95% confidence interval 1.79 to 2.27) and almost doubled when incentives were not conditional on response (1.71; 1.29 to 2.26).
	+ Response was more likely when short questionnaires were used (1.86; 1.55 to 2.24).
	+ Personalised questionnaires and letters increased response (1.16; 1.06 to 1.28),
	+ as did the use of coloured ink (1.39; 1.16 to 1.67).
	+ The odds of response were more than doubled when the questionnaires were sent by recorded delivery (2.21; 1.51 to 3.25)
	+ and increased when stamped return envelopes were used (1.26; 1.13 to 1.41)
	+ and questionnaires were sent by first class post (1.12; 1.02 to 1.23).
	+ Contacting participants before sending questionnaires increased response (1.54; 1.24 to 1.92),
	+ as did follow up contact (1.44; 1.22 to 1.70)
	+ and providing non-respondents with a second copy of the questionnaire (1.41; 1.02 to 1.94).
	+ Questionnaires designed to be of more interest to participants were more likely to be returned (2.44; 1.99 to 3.01),
	+ but questionnaires containing questions of a sensitive nature were less likely to be returned (0.92; 0.87 to 0.98).
	+ Questionnaires originating from universities were more likely to be returned than were questionnaires from other sources, such as commercial organisations (1.31; 1.11 to 1.54).”

**2d. Questionnaire Format**

* See Fanning (2005) for many practical examples for formatting and layout of questionnaires.
	+ Fanning, E. (2005). Formatting a Paper-based Survey Questionnaire: Best Practices. Practical Assessment Research & Evaluation, 10.
	+ <http://pareonline.net/pdf/v10n12.pdf>
* Toepoel et al. (2009) conducted an experimental study of item layout and found horizontal presentations seem to work better than vertical presentations, and linear better than non-linear.
	+ Linear: Poor, Fair, Good, Very Good, Excellent
	+ Non-linear (not on sample line):
		- Poor, Fair, Good
		- Very Good, Excellent
* Studies like this work well as methodological dissertation topics.



vs.



* Items on left and responses on right also seems to work very well, and makes manual data entry easier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Strongly Disagree | Disagree | Somewhat Agree | Agree | Strongly Agree |
|  |  |  |  |  |  |
| Overall this course is well designed. | 1 | 2 | 3 | 4 | 5 |
| Materials are available when needed. | 1 | 2 | 3 | 4 | 5 |
| Instructor responds to questions in timely manner.  | 1 | 2 | 3 | 4 | 5 |

* Likert responses, positive to negative or negative to positive?
	+ Negative to Positive: Poor, Fair, Good, Very Good, Excellent
	+ Positive to Negative: Excellent, Very Good, Good, Fair, Poor
	+ Friedman et al (1994) argue that direction can present a biasing effect in responses, however, their data in table 1 shows little evidence for this since only 3 of 10 items were significantly different.
	+ Chan (1991) reached a similar conclusion to Friedman et al. but his findings were also mixed but generally more supportive of the bias.
	+ Weng and Cheng (2000) found no evidence of bias from either order.
	+ Overall research seems to be inconclusive about order bias; where there are differences, the differences do not seem large.

**3. Data Entry (to be added)**

**Sections below explained in detail through remainder of course.**

**4. Reliability Assessment**

* Test-retest
* Internal Consistency
* Parallel-forms
* Rater Agreement

**5. Item Analysis**

* Difficulty
* Discrimination
* Correlation with total score
* Contribution to reliability

**6. Validity – Structural Assessment**

* Correlation Matrix
* Exploratory Factor analysis
* Confirmatory Factor Analysis

**7. Validity – Construct Assessment**

* Construct
	+ Correlated with related constructs
	+ Mean differences with known groups
	+ Correlated with similar measures
* Convergent – related as expected
* Divergent – unrelated as expected

**References**

Chan, J.C. (1991) Response-order Effects in Likert-type scales. Educational and Psychological Measurement, 51, 531-540.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R., & Kwan, I. (2002). Increasing response rates to postal questionnaires: systematic review. *BMJ : British Medical Journal*, *324*(7347), 1183.

Friedman, H.H., Paul J. Herskovitz and Simcha Pollack (1994), "Biasing Effects of Scale-Checking Style in Response to a Likert Scale." Proceedings of the American Statistical Association Annual Conference: Survey Research Methods, 792-795.

Rolstad, Adler, & Rydén (2011). Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis, Value in Health, 14, 1101-1108.

Toepoel V, Das M, van Soest A. (2009). Design of Web Questionnaires: The Effect of Layout in Rating Scales.

Journal of Official Statistics, 25, 509–528.

Weng, L. & Cheng, CP (2000). Effects of Response Order on Likert-type Scales. Educational and Psychological Measurement, 60, 908-924.