Chapter 7

Designing Self-Report Instruments:

Survey-based and Interview-based Assessments

7.1 Chapter Overview

Self-report instruments belong in a broad category that includes survey-based scales, opinion polls, questionnaires and interview-based tools. Of these, surveys enjoy the most widespread popularity, with numerous online “survey builder” tools available today for interested users (See for example: *https://zapier.com/learn/forms-surveys/best-survey-apps/)*.

Their long history, coupled with widespread use, would suggest that the “how to” techniques for designing this type of assessment are easily learned and applied. But, a vast number of existing self-report items and tools have serious flaws, and even when well-designed, self-reported responses are inherently vulnerable to multiple kinds of error. To build and use self-report tools effectively, therefore, both assessment designers and users must be aware of the essential guidelines and literature on this particular assessment modality. Chapter 7 is devoted to this topic.

Readers will recall that as a part of discussing optional assessment operations, Chapter 3 outlined the general characteristics, advantages and limitations of survey-based and interview-based tools in Table 3.5. Building on that, Chapter 7 will now treat their distinctive features, advantages and disadvantages, guidelines for constructing items, and how to assemble self-report tools for different purposes.

To connect this chapter with the rest of the book and the *Process Model for Assessment Design, Validation and Use*©, see Figure 7.1. As with Chapters 5-6, the content of Chapter 7 speaks specifically to Phases II-III, which would follow from Phase I of the Process Model.

[Insert Figure 7.1 about here]

7.1.2 Chapter Objectives

After reading this chapter and completing the accompanying exercises, the reader should be able to:

1. Identify the origins and distinctive features of self-report assessments, specifically, survey-based and interview-based assessments.
2. Describe the utility of, and applications with, self-report tools in different disciplines.
3. Apply established guidelines/criteria for designing survey-based assessments.
4. Apply established guidelines/criteria for designing interview-based assessments.
5. Design, evaluate and select appropriate rating scales to accompany various types of self-report items.
6. Summarize the major sources of, and methods for ameliorating, error when designing self-report tools.
7. Critically evaluate the quality of self-report assessments for given domains, populations, inferential needs and assessment purposes.

7.2 Self-Report Instruments:
Examples, Definitions, and Historical Applications

> "*Survey research is the best known and most widely used research method in the social sciences……..(E)veryone in the United States (has) been affected by surveys. Politicians launch or scuttle campaigns and dreams based on voter surveys. Manufacturers discontinue or mass produce products....Federal aid programs hinge on the results of population surveys...* (Babbie, 1990, p.xvi).
>
> Interviewer: "*Would you say that Eden is a good, average, or poor place to live?*
>
> Interviewee: "*Oh, I'd say it is a good-average place to live."* (Anonymous)

7.2.1 What's in a Question? In Search of Eden

Often, the very best (or only) way to learn more about a phenomenon, condition, or an underlying construct is to directly put the question to people who might know about it. But, a good self-report question is more difficult to formulate than one would expect. What's in a question, or, the responses it engenders? Consider once again the excerpted exchange from the telephone survey cited above, and think about how *you* might have responded.

To begin, the key word in the question --"Eden"-- could carry multiple meanings for different respondents. If we interpreted it with reference to the story of Adam and Eve from the chapter on *Genesis* in the Bible, it would prompt one kind of answer. In addition, as the Garden of Eden in a Biblical sense lies outside the realm of direct experience of people in the modern world, some may respond with an ambiguous "*I don't know", or "What do you mean?"* Others who view religion as a private and personal matter, may choose to *not* give a response at all.

Note that already, the range of responses poses an interpretive dilemma. The item's structure seems to require respondents to pick just one of three possible choices the interviewer presented: a) good, b) average, or c) poor. Yet, all the obtained responses thus far fall outside those parameters. On top of that, we still need to deal with the meaning of the initially obtained response, "good-average". How should that be coded, summarized, and interpreted?

In other cases, people might interpret the word, "Eden" with reference to a real place, such as, a town in the state of Arizona that currently bears that name. Clarified and situated in a specific context, the same question is now answerable with one of the three choices provided. This is probably what the assessment designer intended in the first place, but without that particular qualification, the item is ambiguous.

In the same vein, a number of reasons could have led to the outside-the-box response--"good-average". Consider a few.

- The respondent may not have understood parts of, or the entire, *question*. The interpretive issues we encountered with the word, "Eden" are illustrative of this problem. Similar barriers with content-related or structural aspects of self-report items relate to the choice of wording, grammatical construction, directions, or mode and format of item presentation and the response choices.
- Respondents may find that there were *too few response options* to reflect their personal position on the topic; hence, the answer fell somewhere in-between.
- Respondents may find that the *response options* were illogical or otherwise unclear. For instance, a respondent might wonder: Does "good" stand for "above average" or "well above average"? Is "average" the same as "fair"? Why is the wording not in a logical progression, such as, from "below average"→ "average" → "above average"? Is there some hidden meaning in the answer choices that I'm missing?
- Other respondents may simply be "playing around", not taking the telephone survey exercise seriously. They just provided a "clever" answer to unsettle the interviewer and the interview process.
- Or, the response could have simply been a random variation, highly unlikely to be repeated even if the same person were to be interviewed again.

This is not an exhaustive list, but given the many possibilities, what is the best way to treat the survey response? The challenge is to address these difficulties in a disciplined manner that allows us to extract valid and reliable information on the construct that the item was originally meant to tap.

In self-report tools, demographic, cultural, linguistic, age-related or other population-specific differences could lead to various respondent difficulties in comprehending either the question, or the finer differences in the response options. The burden then falls on the researchers or practitioners who design the survey, or wish to interpret the information the tool produces meaningfully, to resolve the issues for the assessment purposes at hand. Self-report items that instigate too many speculations are a warning signal from a measurement standpoint.

Two pointers should be immediately clear from this first example. First, when different individuals interpret the same words with their own personal and individualistic frames of reference, results of self-report items become highly variable and suspect. Thus, an assessment designer's goal should be to craft questions that engender *consistent interpretations* of the item's content and language, regardless of who takes the survey. Second, when structured response items cannot "catch" the full range of answers that members of the targeted population could provide, unknown sources of error can complicate both the assessment process and data produced. Better designed items and instruments would minimize such issues.

To sum up, while anyone could ask a question or count the answers to a series of survey questions, there is a formal logic and skill set to designing high quality self-report items. There are scientific standards for performing studies that rely on self-reported data. The challenge is to determine the sources and seriousness of the errors, and to extract the highest quality of information from the tools on the targeted construct(s) and populations.

7.2.2 Definitions

A **self-report assessment** is a general type of measuring tool where individuals respond directly to a set of questions presented to them in oral or written form. The term, *self-report*, applies to any method which involves asking a participant directly about their own perceptions, recollections of experiences, feelings, attitudes, beliefs or behaviors related to some object or phenomenon with which they are expected to be familiar. Respondents should answer by themselves without interference from the researcher, assessment practitioner, or the medium through which the questions are presented.

Self-report questions could have **closed-ended** item formats, offering pre-set answer options from which respondents must select only one. Alternatively, instruments could include **open-ended** item formats with free responses allowed, but within limits. Other tools could combine both item formats.

Typically, the items are presented in written, "paper and pencil" mode, orally--whether in person or by telephone, or via computer-mediated formats on the internet or similar, alternative platforms. Tools may be administered in individualized, small group or large group formats. Two main sub-classes of self-report tools are illustrated in Figures 7.2 and 7.3 and defined next.

Insert Figures 7.2 and 7.3 about here

Survey-based assessments: The most popular category of self-reported instruments is the **survey,** also called a **questionnaire**. An opinion poll is a common example of a survey-based tool.

In written surveys, respondents read the question and either select a response from the options provided, or write in a response to an open-ended item by themselves. Individual items may elicit specific types of descriptive information on individuals. A cluster of items may often be organized in the form of a **survey-based scale** that measures a latent construct.

See Figure 7.2 for an excerpt of a client satisfaction survey on psychological counselling, meant to be administered in service provider settings. The key components of closed-ended survey items are marked, and include: (1) **survey items,** presented either as a statement or direct question, (2) **multi-point rating scales** tied to each item from which respondents must select *one* answer, and (3) **anchors,** or verbal descriptors tied to each point on the rating scale to clarify distinctions at each level of response.

Interview-based assessments: Interview-based assessments are the second main category of self-report tools. Here, items are presented by an interviewer either in a one-to-one, or small group setting. Again, respondents could answer items in a highly structured format by selecting a response from optional choices given, or by providing an open-ended answer within the parameters provided by the assessment designer.

Interviews are unique in that they permit the use of **item probes.** Probes are follow-up questions that help clarify or confirm the response that interviewees provided to preceding item(s).

Figure 7.3 provides an excerpt of an interview-based assessment to assess the competence of healthcare professionals who are required to interview their patients or clients skillfully using a technique called Motivational Interviewing (MI). MI is applied today in healthcare settings in an effort to motivate patients to care for themselves (after Rollneck, Miller, & Butler, 2008). The interview-based tool in this example is meant to measure the performance-based competency domain of MI, in which healthcare staff, counsellors, dentists, or doctors are expected to be adept (adapted from Alqirq, 2017).

In the interview application in Figure 7.3, the professionals are assessed formatively in training contexts with a combination of closed-ended and open-ended questions. Item probes are employed for deeper assessment of counsellor competence on specific performance indicators.

Because of their dependence on live exchanges, interview-based assessments are more costly to administer than other assessment modalities. Often, they add further costs by incorporating a training component for interviewers. This may be necessary for standardization of interviewing procedures for higher stakes decisions, such as, diagnostic interviews. A benefit, however, is that interviews provide a means for the respondent to clarify, explain or defend their responses. Thus, the format is advantageous from a validity perspective, especially to ensure that a recorded response is truly reflective of the interviewee’s status on the variable or construct being measured.

Examples of interview-based assessments are *viva voce* examinations, which are oral examinations that typically supplement written or laboratory examinations for students seeking degrees in various scientific disciplines. Assessments in early childhood educational contexts are often interview-based. Many tests of language or conversation skills employ interview-based methods, such as the *Test of Spoken English.* Interviews are also often applied in many healthcare fields, for example, as "intake assessments" of patients leading up to further diagnostic testing or prescription of treatment.

Reflection Break

A. **Identify examples of widely-used or published self-report tools from the following fields or disciplines.**
   - **Education**
   - **Psychology**
   - **Health**
   - **Other Social Science Field**

B. **What constructs are measured in each case above? For which populations are the tools designed? What are the declared purposes—score-based inferences and uses—for each? Discuss whether self-report modality is the best one for each application.**

7.2.3 Historical Applications

We saw in the first section that self-reported tools require careful thought during design. What were the early drivers of this mode of assessment, and in what contexts were the instruments used in the earliest applications? What types of item formats were useful, and why? In this section , we take a look back at the history of these tools.

Public Censuses and Surveys: An early use of questionnaires arose in the context of public censuses and surveys. The purpose of a **census** is to gather descriptive data from every individual in a given population. The information is collected to serve some larger public or social purpose, and is typically used by government-level decision-makers. In contrast, a **sample survey research** gathers data from a smaller subset of a defined population with the intent of extrapolating the results from the sample to the larger population. In technical and historic terms, the concept of a survey originated within a broader context of conducting this form of large scale research for informing governmental actions and policies, although the term “survey” now commonly refers to the data-gathering instrument only (Babbie, 1990; Rossi et al, 2013; Dillman, 2011; Dillman et al, 2014).

Reportedly, early census questionnaires were invented by the Statistical Society of London around 1838 (Wilcox, 1934; Hilts, 1978). The purpose then was to gather descriptive information on given characteristics of regionally-defined populations, such as, their age, occupation, income, taxpayer status, housing, and so forth. The U.S. government’s Census Bureau administers similar questionnaires to residents and tax-payers routinely today, guided by similar purposes. The earliest U.S census dates back to 1790. (See: https://www.census.gov/history/.html).

Typically, items on descriptive surveys and censuses elicit facts on a person’s demographic characteristics (e.g., gender), group membership (e.g., affiliation to a political party), preferences (e.g., religion), or specific behaviours (e.g., food items individuals buy at the grocery store). In policy contexts, descriptive item-level data are statistically compiled for analysing how variables may be distributed in defined social groups. Explanatory and predictive research using descriptive survey items is common in the many social science and health fields, where researchers look for the relationships among different variables (Babbie, 1990).

Examples of two descriptive items adapted from a census-based questionnaire are shown in Box 7.1. The first item deals with the size of households. The second item utilizes parenthetic clarifications for each response category. With the highly structured response options in the item, we see that detailed directions are an essential accompaniment to item prompts, so as to obtain accurate and unambiguous responses on the facts sought.

Insert Box 7.1 about here

Origin of Attitude Surveys and Psychological "Scales": The study of psychology as an experimental science began in early 19th century Germany, and planted the early seeds in applying self-report techniques for psychological scale development. Scientists in so-called "perceptual" laboratories led the way towards "scaling" of human perceptions, moving the field away from philosophy, where it was originally situated (Wundt, 1873). The influence of this early tradition of psychological scaling research is evident in the psychoacoustics studies of S.S. Stevens we discussed in Chapter 1 (Stevens, 1946; 1968).

Two distinct traditions evolved in the scaling of psychological constructs, respectively. These involved scale construction with ordered versus unordered items. To examine these traditions in detail, see the illustrations in Figure 7.4 and review Boxes 7.1-7.2.

Figure 7.4 and Boxes 7.1-7.2 about here

Some scaling studies attempted to locate items on a scale continuum by their difficulty or intensity levels as determined by human judges (Thurstone & Chave, 1927; 1929). Thurstone and Chave (1927), for example, created a scale on the perceived seriousness of crimes with items, such as, assault, theft, and homicide, applying an integrated set of methodological principles called the Law of Comparative Judgments. In this method, human judges compared the content of each item with that of the others by making paired comparisons. The items were then ordered by their "intensity" levels based on those judgments and placed on a scale continuum, as illustrated in Figure 7.4.

To measure intensity of an attitude item, for example, the judges were asked: How strong is this item on attitude X as compared to the others? To measure difficulty of an item in an ability domain like mathematics, they would be asked instead: How difficult is this item compared to others? In the final version of the instrument, the item ordering was random rather than from low to high intensity. Thurstone scales did not use titles or subtitles in the layout that might suggest item intensities to respondents, thereby possibly swaying their responses. Thurstone's item-development process is considered the foundational method for scaling itemsusing human discriminate on processes.

The **Thurstone scale** in Box 7.2 was also developed following the Law of Comparative Judgments. Note the interesting differences in the numeric values under “weights” that reflect how Thurstone’s judges rated item intensity levels with respect to the specific item’s content on the attitude.

Other scaling methods of the time attempted to place both items and persons conjointly on a single continuum by their perceived levels of difficulty or ability (Guttman, 1941). Still others were interested in locating individuals by their empirically-derived strength on a trait using a “total score” on the scale. This last approach ignored the difficulty or strength of individual items (Likert, 1932). Review the overlapping and contrasting features of the three approaches to attitude scale construction in Figure 7.4.

Also a psychologist, Rensis Likert (1932) was a contemporary of Thurstone and Chave (1927). But, as clear from Figure 7.4 and Box 7.3, he applied the self-report technique in an entirely different manner. Likert assumed that all items tapping a construct were essentially replications of each other and therefore, interchangeable in intensity. Any differences in item intensity, even if these existed, did not really matter in the scale construction process.

Likert posited instead that a ”total” scale score could be created as the simple sum of an individual’s responses to a series of attitudinal statements (the items), each accompanied with a five-point response scale that is now famous: Strongly Disagree, Disagree, Uncertain, Agree, Strongly Agree (SA-A-U-D-SD; see also, Figure 7.2). Because of the positive and negative response points on either side of a neutral point on the rating scale, a **Likert scaled item** is described as **bipolar.**

The statements served as the items on a Likert-type instrument. Individuals selected and endorsed only one SA-A-U-D-SD response to each item. As shown in Box 7.3, the total score resulted from summing a person's numerically weighted responses. This score reflected the person's overall attitudinal strength on the total score scale.

Boxes 7.2 and7.3 together illustrate the differences in scoring and scaling methods of Thurstone and Likert, respectively, with a common set of items tapping into Attitude towards the Movies. In contrast to the Thurstone scale, it is the *individuals*--not items—that are "scaled" on a continuum on their total scores in Likert's approach.

Another historic technique, Osgood's (1964) **semantic differential** scale, was developed to answer the question: What kinds of *meaning* do people attribute to different concepts in life, like government, schooling, or corporations? To answer this question, he developed thematic items accompanied with a seven-point, bipolar rating scale. The rating scales were anchored with opposing adjectives at each end. Items could be ranked from good to bad (the evaluation dimension); from strong to weak (the potency dimension); and from fast to slow (the activity dimension). Semantic differential scales are scored in ways similar to Likert scales, and offer an alternate technique for scaling people on a construct continuum.

Review an example next. To determine what the word "democracy" means to people today, the following survey exercise could be administered. Item 4 is reverse-oriented on purpose. What might be the value of this strategy?

Democracy

1. Good 1---2---3---4---5---6---7 Bad
2. Dynamic 1---2---3---4---5---6---7 Static
3. Strong 1---2---3---4---5---6---7 Weak
4. Ineffective 1---2---3---4---5---6---7 Effective

Reflection Break

A. *"Likert assumed that all items tapping a construct were essentially replications of each other"*. **This statement suggests that Likert conceptualized the attitudinal domains he was measuring as (choose one, and refer back to Chapter 4, if needed):**

   a. Stratified
   b. Non-stratified
   c. Hierarchical
   d. Non-hierarchical
   e. More than one could apply (Explain):

B. Which scaling method makes assumptions consistent with domain sampling theory? Explain.

C. Would you prefer the scaling method of Thurstone, Likert, or Osgood, if you wished to design a scale for a psychological construct? Explain.

D. Directions: Select a self-report instrument of your choice to answer the following questions. Skip if the item does not apply to the tool.

   a. **Indicate if the tool is a *survey-based* or an *interview-based* tool.**
   b. Identify: 1) a *descriptive item*, 2) an *open-ended* and/or 3) a *closed–ended* item in the tool.
   c. Identify a group of items that suggests a *"scale"*. What underlying construct is the designer trying to measure?
   d. Do you see signs of Likert-scaling, Thurstone scaling, Osgood's or another scaling method in the instrument? Explain.
   e. Evaluate whether the above item formats and scales are suitable from a validity, reliability and utility perspective. Think of the construct(s) measured, the population, and the instrument's purposes. Explain your answers.

7.3 Distinctive Properties of Self-Report Instruments

7.3.1 Constructs Best Measured

Ideal Constructs for Survey-based Tools: Not all constructs are well-measured with self-report techniques, but several can be. Outside descriptive information on fact-based characteristics of individuals, items in questionnaires and surveys typically target information on non-cognitive, social, or health-related constructs. Broadly, non-cognitive constructs include interests, social-emotional mind-sets, dispositions, and personality characteristics. While interests have to do with a person's likes and dislikes (feelings), attitudes and dispositions could deal with feelings, values, or behaviors that people may espouse with respect to a particular attitudinal object.

In education, parent, teacher and student attitudes towards schools in general, or various aspects of schooling are routinely assessed through surveys (see Phi Delta Kappa Poll, 2018 at: http://pdkpoll.org/results). Health-related constructs could seek experience-based perceptions of well-being, pain or recovery status from a health setback. Note that the characteristics best-tapped by survey-based tools are all *preexisting* or *naturally occurring* variables and constructs (Box 7.1-7.2).

Ideal Constructs for Interview-based Tools. In contrast, interview-based assessments can be designed to measure cognitive domains, as well as, non-cognitive and health-related constructs. Figure 7.2 illustrated an example of an interview-based assessment for assessing a competency-based domain related to MI. Many personality, attitudinal, and psychological constructs have also been measured successfully with interviews, in particular, for applications in the mental health field (Shea, 2016).

An illustrative set of interview-based items on a non-cognitive construct is shown in Box 7.4. these items intend to measure "locus of control", a personality characteristic dealing with factors to which individuals attribute their successes in life (Lefcourt, 2014). The literature distinguishes between an *internal* versus *external locus of control* on which individuals could vary. We see a closed-ended interview question in Box 7.4, followed by three open-ended probes. Interviewee responses would be noted down by the assessor in the latter, and coded systematically to identify the "look-for" indicators tied to the psychological literature on locus of control (after Author, 2003).

7.3.2 Typical Distributions versus Maximum Performances

Another distinctive feature of self-report tools follows from the first, and applies to the nature of constructs we are attempting to measure. Compare the three items in Table 7.1. With naturally occurring characteristics or non-cognitive constructs, we cannot categorize responses as "correct" or "incorrect", nor attempt to examine acceptability levels of an answer with the help of scoring keys or rubrics. Rather, the goal is to devise the assessment with items that depict the *typical and natural distributions* of the targeted construct or variable as accurately as possible.

Insert Table 7.1 about here

The first two questions in Table 7.1 depict personal opinions and values that are neither wrong nor right. Accordingly, assessment conditions must be structured to capture the actual and typical distributions of individuals on these indicators.

In contrast, the third item taps into a cognitive dimension—recall of concept knowledge--where there *is* a correct response. In the latter, we would prefer to obtain information on what a person's best recollection of the factual information is—that is, the item must be able to tap into their *maximum performance* potential. This important distinction on measuring cognitive versus non-cognitive characteristics must be clear to item designers relying on the self-report modality from the start.

7.3.3 Populations Best-suited for Self-report Tools

Different self-report tools make particular assumptions about populations they target. They expect a certain level of verbal literacy in respondents in the language of the instrument.

Surveys are typically designed for *readers*. The reading level should be adjusted to fit population profiles in terms of language literacy levels, but most surveys expect respondents to know how to read and write in the language of the survey. Further, survey respondents are expected to have the capacity to write in or enter responses accurately by themselves using computer-based, paper and pencil or technology-mediated platforms through which surveys are administered.

Interview-based formats make fewer assumptions about language proficiency. They are particularly well-suited for special needs populations, such as, very young children, nonreaders, the ill, and the elderly, where variations in reading ability or literacy levels can be accommodated. Appropriate kinds of interpretive or translational support may be needed to enhance measurement quality with special populations.

Expert probing and clarification are one way to obtain valid information from these groups, making *interviewer training protocols* an essential part of assessment design requirements. Interview-based modes are also best applied when highly sensitive information is sought from individuals (such as, violation of parole guidelines by prisoners; or levels of abuse in the home environment). So, when surveys are vulnerable to unreliable or false information, interview-based assessment modes can play a compensatory role.

7.3.4 Types of Data from Self-Reported Tools

There are four main types of self-report items, each yielding different kinds of data.

1. **Items with highly structured response formats that help "scale" a latent construct**: Highly structured response items on self-report tools often collectively serve as a **scale,** as shown earlier in Boxes 7.2-7.3, where the score denotes a person's level on an underlying construct continuum.
2. **Items with highly structured response items that yield a social "index"**. In the same vein, responses to items on self-reported *background characteristics*, such as, one's education level, job type, and income level, could also be grouped to produce a quantitative **index** of a social construct, such as, Social Economic Status (Babbie, 1990). For example, *Home Educational Resources* is an index variable constructed by International Association for the Evaluation of Educational Achievement (IEA) based on students' answers to six survey items on the TIMSS: (1) Number of books at home, (2) Having a study desk for own use, (3) Having a computer, (4) Having a dictionary, (5) Father's education, and (6) Mother's education. For further details, see the

TIMSS student and teacher questionnaires at the IEA website at: *https://timssandpirls.bc.edu/*

3. **Items that yield fact-based, descriptive information and stand alone**. Other structured response items could individually serve as “stand-alone”, descriptive items as in Box 7.1. Each descriptive item provides specific, fact-based information about the persons studied, allowing statistical analysis on variables like gender, nationality, and other demographic characteristics.
4. **Open-ended, self-report items**. Finally, self-reported data from open ended-responses yield text or narrative data that allows deeper insights into the persons, such as, their reasons for particular responses they provided. The information gathered from such items, however, must be systematically coded and analysed to extract predominant themes and sub-themes, following accepted standards of qualitative research (Wisdom & Creswell, 2013).

7. 3.6 Measurement Issues and Threats

Self-report measures are treated rather dubiously by many users, despite their popularity. They, therefore, demand support with necessary evidence of validity and reliability. The ultimate question is: To what extent can we trust the data? This section briefly considers the most common measurement problems with self-reported data.

Item-level Ambiguities and Semantic Issues: As the first section to this chapter demonstrated, subjective frames of reference, idiosyncratic responses and semantic issues could cloud the quality of self-reported data. These issues could stem from the language, wording or content of items. Items that offer either *too few* or *too many* response options on multi-point scales, or with *unanchored answers,* could also add error variance.

For example, when respondents are inclined towards selecting both options in a Yes-No or True-False item format, but wish to do so conditionally, the forced choice format precipitates validity and reliability problems. An example of such an instrument with persistent measurement issues is the dichotomously scored, *Myers-Briggs Type Indicator*, a personality measure that is often applied is occupational contexts (Gardner & Martenko, 1996; reprinted in 2016; Pittenger, 2005).

Response Rates and Missing Data: Survey-based instruments tend to produce very low return rates, especially mail-out and online questionnaires. Typically, individuals who do return the surveys are those who want their opinion heard, and either have a very positive or a very negative viewpoint. When studies or practical data gathering efforts are guided by research questions that examine group-based differences, the lack of responses in one or other group may introduce **selection biases** that produce misleading conclusions.

Missing data from self-report tools must be treated carefully and thoughtfully. *Systematic factors* could yield missing data for particular groups. For instance: some respondents may have run out of time due to reading difficulties; others may have been fatigued as the instrument was too long; others may have chosen to avoid particular items due to a religious or cultural factor; still others may have been excluded from the services on which they were being questioned, and felt ill-equipped to offer responses. Often, these issues arise when closed-ended items do not incorporate *Not Applicable* or *Unable to Respond* options. Beliefs of data analysts that the gaps in survey data are random are often wrong, and must be verified before statistical algorithms based on such assumptions are employed to fill the information.

Faking, Exaggeration, Acquiescence Bias and Social-Desirability Biases of Respondents: Self-report measures can be falsified in either the positive or the negative direction by respondents. For example, one person might choose to say nice things about their home environment even when family members are abusive towards them; another might choose responses that make their home environment look more abusive than it actually is.

Exaggerated responses and “faked” responses are common in self-reported data in cases where respondents are too embarrassed to reveal private details of their lives, or wish to present themselves in best possible light to the external world.

There are two kinds of respondent-specific biases that may affect self-reported results. The **social desirability bias** reflects the unconscious tendency of people to say things that will meet with social approval. For example, if asked as to whether we like being in a school or the workplace with people from different ethnic backgrounds, we might say, "Of course!" without really thinking about where we personally stand on the issue. Socially desirable responses are forms of culturally conditioning that influence individuals to say what they believe others want to hear (Krumpal, 2013). To detect levels of social desirability bias in survey response patterns of individuals, tools like the Marlowe-Crowne Social Desirability Scale (Reynolds, 1982) can be embedded in questionnaires.

The **acquiescence bias** is a human tendency to readily agree, rather than disagree, with statements. “Response sets" are similar and refer to our tendency, conscious or unconscious, to mark a particular response choice repeatedly without giving due consideration to the item content. Thus, a person might demonstrate an "uncertain" **response set** when asked to fill out a questionnaire. Both response patterns mask the true standing of a person on the construct, leading to unknown invalidity levels in the data (Richman et al, 1999).

Forgetfulness of Respondents: Lastly, subjects may also forget pertinent details on the topics on which they are being questioned, such as, details of happy childhood experiences, or symptoms of pain or trauma. Self-report studies are inherently biased by the person's feelings *at the time they fill out* the questionnaire or participate in an interview. More negative responses could result when the person feels poorly for reasons other than the issue of focus; the reverse pattern emerges when the individual feels better generally.

Tackling Measurement Errors applicable to Self-reports: In sum, a key concern with self-reported instruments, particularly questionnaires, is that there may contain quite large measurement errors. Some of these errors are random, and caused by unintended mistakes by respondents, interviewers and/or coders, affecting reliability levels. Others are systematic, such as, a repeated response to particular questions due to a given language barrier.

The design work on items and scales is crucial to curtailing the amount of measurement error, as are the validation efforts undertaken thereafter. Different analytic tools are available for estimating the quality of items and scales from self-report tools. One example deals with using Multi-Trait Multi-Method (MTMM) experiments. These studies examine validity levels through correlation matrices of self-reported results with other measures, removing effects of assessment methods (Campbell & Fiske, 1959; Saris et al, 2010). Other tools help in predicting item and scale response patterns using algorithms, like the Survey Quality Predictor software (SQP). The specific methods above rely on specific assumptions and definitions of validity and relaibility that we will discuss in further detail Chapters 9-10.

So, compared to other assessment formats treated in earlier chapters—such as, performance assessments or even structured response tests--the relative ease of item compilation, administration and lower cost of self-report techniques are definite advantages. But, issues uniquely applicable to these tools must be confronted during item design, scale development, administration, scoring, coding, and interpretation of the data. As the next few sections suggest, some of the issues can be contained, or avoided altogether, during the design phases of the Process Model. The rest must be confronted during validation.

Reflection Break

A. Collect some self-report tools used in your field, organization or institution. Identify ONE example of a self-report item that falls in each of these categories, using your best judgment:

a. Items with too many responses on a multi-point rating scale
b. Items with unanchored or poorly anchored rating scales
c. Items that have semantic ambiguities
d. Items that will possibly lead to socially desirable responses
e. Items that will possibly lead to response sets or acquiescence biases
f. Items with other kinds of errors (identify and explain these).

B. In each case (if you found the issue), a) define the type of error and why it is a measurement threat in terms of validity, reliability and utility, and b) explain when and how you would fix the issue.

7.4. Applying the Process Model to Design or Select Self-report Assessments

7.4.1 Background- The Applied Setting

We next review an applied instrument design case where two complementary self-report tools are designed to screen individuals on the construct, *Subjective Perceptions of Unhealthy Intimate Relationships* (after Bogart, 2003) starting with one set of assessment design specifications. Violence and abuse in intimate relationships is an unfortunate social problem today affecting many young and older adults, both on college campuses and society at large (Wiersma et al, 2010). A growing body of research shows that intimate partner relationships is associated with various levels and kinds of abuse in adult dating, cohabitating, or marital partnership situations. For victims, initial screening is the first step towards providing treatment, as may be necessary.

Screening instruments, by definition, are short and quick assessments. They yield information relevant for specific types of formative decisions, such as, the need for more in-depth, diagnostic testing in clinical settings, or for selection testing and placement in educational or workplace settings (see Chapter 2; also AERA, APA & NCME, 2014).

The original version of the instruments presented was developed following the Process Model by a counselling psychologist who worked at the Student Services office of a large, urban university in the U.S. (Bogart, 2003). The original design is modified and adapted for the present illustrations. See Boxes 7.5-7.6 in unison with Tables 7.2 and 7.3.

New psychometric concepts or techniques referenced in this section are defined, but more will follow in Chapters 9-11.

7.4.2 Phase I. Specifying the Assessment Context: The Construct, Population, and Purposes

In the top of Box 7.5, we see details of the *assessment context* specifications. The setting is practical rather than academic. The *construct* is identified, along with the domain detailed in Box 7.6, using suitable literature sources (e.g., Burge,1998).

The targeted *population* comprises young adults/students attending universities, inclusive of other adults involved in intimate relationships with partners. Intake assessment with an initial screening component is standard practice in clinical contexts.

The assessment purposes are specified in two parts--*inferences* to be made on individuals assessed (the units of analysis), and specific *uses* to be made in the clinical counseling and screening contexts. This up-front declaration is necessary to ensure that the final items and tool have the properties to support both the proposed score-based inferences and uses of a tool in applied settings.

Boxes 7.5-7.6 about here

7.4.3 Phase II. Specifying the Assessment Operations

Optional Assessment Modes: The instrument *design specifications* deal with screening for signs of exposure to abuse in university students. To provide optional ways to make students comfortable in sharing details about their relationships with partners or spouses, both the survey- or interview-based assessment modality may need to be used. Combining both modes would yield higher levels of validity (better content domain coverage) and reliability (internal consistency reliability estimates improve with more items). The design specifications allow either one of, or both the options.

Multiple "Parallel Forms" Design: Two sets of self-report items were created and organized as two "parallel forms" assessments, each in a different assessment modality in this application of the Process Model. The items were systematically sampled from the same domain in Box 7.6. Other "parallel" instruments may be similarly designed aligned to the domain and item specifications. The detailed domain in Box 7.6 together with Box 7.5, comprise the assessment design specifications in full.

Strictly speaking, **parallel forms** of a test/instrument must be equivalent on every facet of the tool (AERA, APA & NCME, 2014). That is, alternate but parallel forms must be built guided by the same set of design *and* statistical specifications. Specifically, when both the tools are administered to individuals, the statistical distributions of scores must meet particular criteria. To claim we have "parallel forms", the means, variances and error variances must be equal, or close to equal (Bandalos, 2018; Price, 2018).

A first step in creating parallel forms assessments, however, is the design of equivalent, random samples of items tied to a common set of indicators that helps operationalize the construct. This step was followed in this case, and is reflected in the parallel *item structures* (closed-ended, Likert type items), as well as, the samples of items in both instruments covering the same *content* and *taxonomic levels* specified in the domain (Tables 7.2-7.3). But, as the administration conditions and accompanying materials clearly differ in survey-based and interview-based modes, we have a variant of strictly parallel forms design here.

Domain of Indicators: The general indicator of the domain in Box 7.6 is broken down in terms of *three sub-domains* dealing with emotional, physical and verbal forms of abuse. The sub-indicators for emotional abuse only are excerpted here.

Note that the domain also includes an indicator with "reverse-oriented" wording representing affirmative behavioral indicators of a healthy intimate relationship. Such indicators allow the design of items with similarly reverse-oriented wording. When randomly mixed with other items, such items keep respondent-rooted errors in check, such as, proneness towards response sets or acquiescence biases, or a lack of engagement in the assessment process.

Although not shown in the boxes, the assessment designer drew on psychological literature sources and guidelines relevant for clinical counselors and therapists who serve the targeted population, to justify the original domain. Such justification is compliant with a key guideline for specifying domains defensibly (see Chapter 4). Some sources were the following.

Burge S.K. (1998). How do you define abuse? *Family Medicine*, 7(1); pp. 31-32.

Eisenstat (1999). Examples of Abusive Behavior *New England Journal of Medicine*, Volume 341(12), pp. 886-892. Table 1.

National Coalition Against Domestic Violence (2000). Predictors of Domestic Violence, retrieved from web site, http://www.ncadv.org/problem/predictors.htm, May 6, 2002.

Wolfe D.A., Scott K., Reitzel-Jaffe D., Wekerle C., Grasely C., Straatman A-L (2001). Development and Validation of the Conflict in Adolescent Dating Relationships Inventory. *Psychological Assessment*, 13(2); pp. 277-293.

Taxonomic Levels: To set item-writing parameters, the designer applied a suitable *taxonomy* for assessing *non-cognitive constructs*, as appropriate. Four potential levels of non-cognitive processing could have been targeted for measurement (Table 4.2, Chapter 4), as follows.

- “Belief-specific” and “value-expressive” component of dispositions: What a person *perceives to be true* about something; their opinions, beliefs, values, or perceived knowledge about some present or past experiences, events, places, persons or objects. A Likert-scaled item to tap beliefs about abuse in relationships is:

Any form of abuse in a marital partnership is wrong. SA-A-U-D-SD

- "Emotional" component of dispositions: What a person *feels about* some present or past experiences, events, places, persons or objects. A Likert-scaled item at this level is:

  I feel oppressed in my marriage. SA-A-U-D-SD

- **"**Behavioral" component of dispositions: What *a person would do*, in terms of social behaviors, responses, actions, or practices in relation to some experience, event, place, person or object. A modified-Likert type item at this level is:

  I keep secrets from my partner. This statement about you is:

  a) True

  b) Mostly True

  c) Mostly False

  d) False

- "Metacognitive Component" of dispositions: The habits and skills a person would demonstrate to self-evaluate mind-sets, feelings, and behaviors; to self-correct course; or to make improvements to, and alter, unproductive attitudes.

  A modified-Likert type item at this level:

  I think aloud about why I feel bad about my relationship with my partner. This statement about you is:

  a) True

  b) Mostly True

  c) Mostly False

d) False

Note that the assessment designer in this case focused on two of the above only, as shown in Box 7.6 and Tables 7.2-7.3.

Specifications of assessment conditions, materials and scoring: Some conditions are similar for both tools, but other requirements vary depending on the mode of assessment, as shown in Box 7.5. For instance, interviews would require standardization regarding the content of item probes, the number questions that is ideal, and when to apply the probes. These design specifications are provided only for the interview-based assessment segment.

In contrast, the scoring specifications apply to both modalities. A Likert *scaling method* is applied for designing both tools and a total score is obtained by summing item ratings. Even though screening decisions carry lower stakes than diagnoses, some guiding specifications are necessary for consistent decisions. The decision-making guidelines for the score ranges shown should ideally be derived as a part of the design process (Phase II-III), and validated with expert opinions and through empirical studies (Phase IV).

7.4.4 Phase III. Design or Select the Instrument

Designing Survey-based Assessments: The survey instrument comprises 10 items, a mixed sample taken proportionally from the three sub-domains. The short length of the survey is apt for the screening purposes. The number of items falls within the 10-20 range in the design specifications.

To build the survey-based assessment, traditional Likert scaled items were incorporated with a forced choice, four-point response scale, one of which that persons would endorse (Table 7.2). An endorsement scale is well-suited for measurement of *perceptions/beliefs* and *feelings*.

The lack of an “Uncertain” response category in the items prevents response-avoidance by respondents. Note also that, per the specifications, the survey embeds a small number items that are worded affirmatively (see items marked with asterisks). With respect to the construct continuum, this form of item construction is "reversed"—that is, a higher rating for persons would denote a lower standing. The "reverse-oriented" items could also help prevent faking and various respondent-specific biases.

Complete assembly of the tool requires that the items are formatted in a respondent-friendly manner. The entire survey must be readable, inviting, printed in sufficiently large font size, and with clear directions. The purposes must also be declared. Assurance of confidentiality and anonymity, as appropriate, are critical for securing the trust of the respondent, as well as, consistent with ethical research and professional practices in evaluation (see AEA Guidelines at www.eval.org; Yarborough et al, 2011).

Designing Interview-based Assessments: The interview-based assessment taps into *experience-based perceptions* of respondents relating to partner behaviors. The frequency-based, four point rating scale is therefore, more appropriate for gauging the perceived abuse in terms of how often the subject experienced the same (Table 7.3). Per the specifications, this instrument is between 10-20 items, again, and sampled across the three sub-domains of abuse. The short length of the tool is again, is consistent with the specified assessment purposes (Phase I).

While item-writing and compilation rules similar to survey items are applied here, interview-based items are presented in *question format*, even when closed-ended. Further, along with directions for interviewees, interviews are accompanied with detailed *scripts and protocols* to guide interviewers, as shown.

Insert Tables 7.2 and 7.3 about here

7.4.5. Phase IV Validate the Items, Instrument and Construct Measures

The initial items and overall instruments with 15 items each were subjected to three preliminary validation exercises before the final versions shown were compiled. Details on all the psychometric methods will follow in Chapters 9-11. The validation techniques included:

(1**) Content validation of items and assembled tools, by three external, expert judges**. A Content Validity Index (CVI; after Polit et al (2007) was calculated for each item and scale, to examine levels of agreement with regard to content relevance and representativeness of items vis-à-vis the domain by pairs of judges;

(2) **Cognitive interviewing of respondents and reliability of total scores as a part of a small scale pilot test of both instruments** (N=20 for the survey; N-10 for the interview). Reliability was estimated with the **Cronbach's alpha** technique for total scale scores. This method ensures that all items that yield the total scale score are consistently yielding the same information for all persons. The **cognitive interviews** (Corbone et al, 2002), performed with **focus group** interview methods (Rabiee, 2004), asked a sample of 5-8 respondents from each modality to speak about their response processes, item interpretations, and reactions after they took the survey or interview. The aim was to examine item clarity, item comprehensibility/readability, and comfort levels of respondents during the process (see Corbone et al, 2002) so that revisions could be made to tools.

(3) **Setting criteria for decisions.** The pilot test results from both instruments were reviewed a second time by two expert counselors. They set "cut scores" for separating respondents in High, Low and very Low/Nil categories through consultations and review of scores of individuals with "known" histories of the condition versus those without (Cizek, 2013). This process helped arrive at the best decision-making points on each scale.

7.4.5 Phase V. Evaluate Readiness of Tool for Use

Revisions and Decisions following First Iteration: To arrive at the final forms from the versions in Tables 7.2 and 7.3, the length of the survey and interview was increased to 15, and revisions made to obtain balanced sets of items from each of the three sub-domains of abuse. These lengths provided sufficiently reliable scores of .80 and above for the formative assessment purposes outlined (Phase I). Item diagnostic statistics and expert comments from the pilot-tests, such as the CVI, helped remove poorly functioning items. Several initial items were revised or removed due to ambiguity or a lack of adequate match with the content or taxonomic level of indicators (10-20%). The directions and the interview script were also edited and refined for clarity and efficiency.

Given the evidence and revisions (Phase V), the modified versions of the tools were considered ready for use by practitioners and researchers involved, but with only formative, screening purposes in mind (Phase I). Ongoing psychometric evaluations were planned for the future (Phase IV) to gather supporting evidence on the quality of parallel forms, cut scores, and other aspects of the instruments and measures.

Reflection Break

A. What are three benefits to designing tools *after* specifying the domain in detail?
B. List two benefits of preparing detailed assessment specifications, as shown in the illustrative case in Boxes 7.5-7.6.
C. Create a new "parallel form" of (1) a *survey-based assessment* and (2) an *interview-based assessment* based on the specifications provided in Boxes 7.5-7.6.
D. Critique the quality of any 5 self-report items from each tool in Tables 7.2 and 7.3. Which ones would you revise, and why?

7.5 Detailed Guidelines for Writing Items and Assembling Self-Report Tools

Let us now take a more in-depth look at guidelines for item construction and compilation of self-report tools, following from the illustrative application of the Process Model. Some of these guidelines apply to both assessment modalities; others, however, are specific to either survey-based or interview-based assessments that the reader should note. The previous examples in Boxes 7.5-7.6 and Tables 7.2-7.3 will be used again to demonstrate how specific guidelines might be applied. Table 7.4 and 7.5 present new information introduced in this section.

7.6.1 General Guidelines

Guideline 1. *Specify the assessment context in terms of the construct, population and assessment purposes.*

The starting point in developing self-report measures, as with all assessment instruments, is specifying the assessment context with clarity, coupled with the domain of observable indicators for the construct(s), using appropriate theoretical and empirical support. See Boxes 7.5-7.6 again to see how this important guideline is applied for self-report tools.

The boundaries of the population and its characteristics must be clear to the designers and any researchers involved, as well as, to all potential assessment users. In the example—male and female college students and adults in intimate relationships—sets that periphery. Given this, a question to screen out only those who fall within this category can now be included in the section on the "subject's background" that is included in most self-report tools (discussed under assembly of self-report tools).

Likewise, both the inferences (meanings the scores will carry) along with the units on whom the inferences will be drawn, *and* the uses (type of decisions) to be made with scores, must be separately clarified. As the example shows, these are not the same thing.

Examine the sub-indicators of emotional abuse in Box 7.6—these could not have been formulated arbitrarily without knowledge on documented forms of abuse found in intimate relationships. To be able to measure the construct, the domain must be specified next in adequate detail. Even if we start with a vague, brainstormed indicator, such as --"Subject makes (or endorses) negative statements about partner ", this starting indicator must be rooted in some kind of formal knowledge. It also becomes much easier to specify the component indicators of a construct in concrete and observable terms when grounded in relevant data sources.

Guideline 2. *Write items to match the (a) content focus, and (2) level of cognitive, affective or other type of behavioral processing (taxonomic categorization) of the indicators.*

The methodological literature on domain sampling stresses the need for homogeneity of items in given construct domains and subdomains. This means that the items should be constructed to generate responses that "hang together" for particular individuals. In other words, items measuring "emotional abuse" (Indicator 1.3 in Box 7.6) should not generate responses that indicate one's experiences of, or feelings about, "verbal abuse" or “physical abuse” (Indicators 1.1-1.2). Even when the two subdomains are hypothesized to be related, the items must be written to ensure that responses pertinent to one dimension are not conflicted by the subject's responses on another. Domains/sub-domains must have an internally *homogeneous* item structure.

To achieve homogeneity during item construction, we could follow two strategies. First, we could attempt to keep the *content* of items tightly linked to the content specified in the indicators. A second strategy is to write items that are homogeneous with respect to the specific *taxonomic levels* of indicators. Some illustrations follow.

Suppose we wish to write items to measure the indicator, “Reports feelings about emotional abuse”. When parsed, “feelings” reflects the taxonomic category, and “emotional abuse” reflects the content focus of the indicator. Let us evaluate the extent to which following items would be a good or poor match with the underlined aspects of that indicator.

- Item A: *How often does your partner treat you roughly - grab, pinch, shove or hit you?* This item does not match the content focus of the item as it reflects a form of physical, but not emotional abuse. Further, the item taps into experience-based perceptions of partner behaviors (*how often does your partner do this?*), not the subject's feelings--the taxonomic category in the indicator is overlooked. Poor!

- Item B: *How often does your partner say things that make you feel incapable of making decisions on your own?* This is another item with a poor content match! It reflects verbal, not emotional, abuse—the content focus in the indicator. Again, the item taps into experience-based perceptions of partner behaviors, rather than the subject's feelings.
- Item C: *How often do you feel afraid that your partner will give you the "silent treatment"?* This item is a much better match with the content specified, as it reflects a form of *emotional abuse* listed in the literature-supported indicators of the domain, and taps into the subject's *feelings* of fear. Good!
- Item: *How often do you feel guilty because your partner blames you for making him/her angry?* This item also shows a good content match! The item reflects a form of *emotional abuse* per the list of indicators in the domain, and also taps into the subject's *feelings* of guilt.

Guideline 3 *Make sure that the items are written in direct, concrete, and clear language.*

Good writing skills and the ability to make appropriate word choices are essential qualifications for good item writers. As mentioned in a previous section, all respondents in the targeted population should read and interpret the same item in the same way (Jaeger, 1997) There are several strategies to apply this guideline, demonstrated next.

Make items concrete by providing a common reference framework for all respondents. A concretely stated item is clear because it provides a well-defined reference framework that a respondent can use to generate the response. Such a framework could provide a timeframe, location or a particular episode/experience. Consider the concreteness levels of three items next, meant to measure forms of “verbal abuse” by partners. Items have this response scale: All the time-Some of the time-Rarely-Very Rarely-Never.

- Item A: *My partner blames me.* Vague!
- Item B: *In the past month, my partner blamed me.* Better! Item is situated within a timeframe.
- Item C: *In the past month, my partner blamed for making him/her angry*. Good! Item is situated within a concrete timeframe and type of experience.

Use either complete sentence or question formats to construct items. This strategy facilitates communication of the essential idea in an item. For example, the structure of the first item (A) below with an incomplete sentence in the stem is less desirable than that of the second. Further, the second item utilizes a traditional 5-point scale that is likely to capture more variable responses than the 3-point item.

- Item A: *My relationship with my partner is:*
    a) *Satisfactory b) Neither Satisfactory nor Unsatisfactory c) Unsatisfactory*
- Item B: *I have a satisfactory relationship with my partner.*
    a) *Strongly Agree b) Agree c) Uncertain d) Disagree e) Strongly Disagree*

Avoid wordy (verbose) items, or the use of technical language and other jargon. Consider this item on prevalence levels of partner abuse among college students.

- Item A: *Year-specific, perceived norms of partner abuse in dating relationships of college freshman students, is a cause for concern*. Response scale: SA-A-U-D-SD

Technical and academic phrases like "year-specific perceived norms” may confuse and will carry little meaning for lay persons, leading to misinterpretations and invalid data. If used, the terms must be defined in user-friendly terms. A better item would use simpler and more direct language at a lower reading level, such as the following.

- Item: *I am concerned that college freshmen are physically assaulted by their dating partners.* Response scale: SA-A-U-D-SD

Concrete references to the exact class/level--*"freshmen"*—and type of abuse--*"physical assault"*-- add clarity to the item, as well. An eighth grade reading level in the language is should be assumed for most adult surveys to maximize communication clarity and readability.

Avoid abbreviations. An item using acronyms such as "The WHO condemns abuse," hampers readability and direct communication, as well. It makes the item-writer's assumption that all respondents are aware that the abbreviation, WHO, refers to the World Health Organization. As a rule, we should spell out such labels to prevent multiple interpretations leading to added error variance in responses, or missing data on the item.

Avoid double- or triple-barreled items. Items that contain more than one idea or question are poor for measurement purposes. For example, the item, "*My partner abuses me verbally and physically,*" is double-barreled. It will be ambiguous for individuals who vary in their stance or experience on one versus the other form of abuse. Such an item is better split into two separate items, and further concretized, as shown in previous examples.

Avoid leading questions. An item similar to the following: *"Given that I am still married, I have learned to cope with my partner's emotionally abusive behaviors*". The preceding clause to the item--"Given that I am still married" is leading. It strongly suggests a socially desirable answer to the respondent and should be avoided. The item wording should not suggest responses that lean in any direction. It is also better to operationalize "*emotionally abusive behaviors"* into specific acts listed in the domain (see sub-indicators).

Avoid loaded questions. An item can be "loaded" with *pre-supposed assumptions* that may or may not be true for a given individual respondent. As such, there is a logical fallacy to asking a loaded question. For example, the next item example is a loaded one.

Item: *"How often do you blame your wife when you're angry about issues at home?*" This item assumes that: the respondent is (a) has a wife, (b) lives at home with his wife where there are "issues", and (c) is probably male. Neither of these assumptions may be true for certain members of the population. Such items should be avoided or appropriately contextualized in the instrument so that assumptions can be verified. Items that screen individuals with respect to specific assumptions may remove some problems with loaded questions.

Guideline 4: W*hen scaling or creating an index of a construct, include an adequate number of items from the domain and subdomains.*

A general rule of thumb is to build reliability into the assessment design process by developing 3-7 well-designed items per subdomain. A useful practice is to create a few more items per indicator than we need, provided that the instrument does not become too long. This will enable deletion of poorly functioning or redundant items after content validation and empirical item evaluations are completed.

The actual number of items we construct per subdomain and overall domain will influence the internal consistency reliability of results. The minimum number of items necessary for obtaining optimal reliability levels is, in the end, an empirical question (taken up in Phase IV of the Process Model). The principle of improving internal consistency reliability by adding more items, naturally, will not apply to descriptive items that stand alone.

Guideline 5: Choose open-ended and closed-ended item formats judiciously.

Not all items on self-report instruments can be structured with only one possible response. In some circumstances, open-ended items are very useful although less efficient. They are onerous for gathering data, as well as, for coding, analysis, and interpretation. Here are some factors to consider before you make the choice.

- When responses on the construct are likely to vary so greatly that using a highly structured response format will result in a loss of information
- When the measurement process is still exploratory, and not enough is known about either the construct or the population to design structured-response items effectively
- When we seek to probe deeply or clarify answers that respondents give to structured items (as we saw in the examples of interview-based items in this chapter)
- When the respondents are too young or have special needs that are barriers to using a structured-response tool (e.g., if they cannot use a paper and pencil)
- When time and resources available permit us to properly conduct and compile the results of an open-ended assessment

<u>Guideline 6</u>: Use thought and judgement to select the best item-response choices or scale points for closed ended items.

A key design-related decision with structured self-report items is the choice of the most suitable item-response scale. In making this decision, numerous modifications to the basic ordered Likert scale are possible options. Fink (1995) offered the following useful classification that may assist in selection of item scales. Some guiding criteria follow.

Endorsement:	Strongly Agree, Agree, Uncertain, Disagree, Strongly Disagree
	Definitely true, True, Don't Know, False, Definitely False
Frequency:	Always, Very Often, Sometimes, Rarely, Never
Intensity:	Absent, Very Mild, Mild, Moderate, Severe, Very Severe
	To a great extent, to a moderate extent, to some extent, Not at all
Comparison:	More than last year, About the same as last year, Less than last year

Select a response scale that best matches match the content and taxonomic foci of the item. We should evaluate the degree to which the *verbal anchors* (descriptors) for points on the response scale are consistent with the content and taxonomic foci of the item and the original indicator.

Endorsement scales are useful for measuring beliefs, attitudes, and perceptions. Frequency-based scales are common when measuring behavior- or practice-based indicators. Intensity scales are applied for measuring feelings, such as, the degree to which one is satisfied with a service, or the seriousness levels of a condition (e.g., symptoms of pain). Comparison scales are used when measuring experience-based perceptions.

For other variations of multi-point response scales, see Table 7.4. Although all the examples are of 5-point scales, one could have fewer or more scale points, depending on the application. Pictorial descriptors for scale points are often used for children’s surveys.

Scale points must represent an increasing gradient. The response choices in self-report items should represent a *gradient* (a slope), moving progressively from negative to positive or from low to high. While a central point often serves as the neutral anchor--as in the “Uncertain” response in the bipolar Likert scale--the numeric weights allocated to scale points in all self-report items, increase progressively and in order.

Word choice of verbal descriptors on multi-point scales must fit the wording in the item stem. The wording of verbal anchors on the scale should be *logically and semantically consistent* with the stimulus statement or question (the stem). Often, the lack of concreteness and clarity in the stem makes it difficult to find the right wording for response choices. A few appropriately and poorly worded item-response sets are illustrated next.

Poor stimulus-to-response wording match:

Stimulus:
How do you feel about your relationship with your partner?

Responses:

a. Unhappy

b. Satisfied
c. Sad

Better stimulus-to-response wording match:

Stimulus:
How satisfied are you about your communications with your partner?
Responses:

a. Very Satisfied
b. Satisfied
c. Not Satisfied

In the first item, the stem is vague: “How do you feel about your relationship?” A heterogeneous mix of descriptors dealing with different dimensions of feeling on the response points—Happiness, Satisfaction-- invites responses from different reference points, making the item vulnerable to errors or unreliability. The linguistic construction is also weak. Further, the responses are *not* on a gradient.

In the second, the stem hones in on the feeling of *satisfaction* alone, and one aspect in a relationship. The adjectives in *both* the stem and response options refer to *same* feeling of satisfaction. This improves the communication of the question as a whole.

Balance the number of positive and negative response points on item scales. An equally balanced set of options would not suggest a direction for the respondent to take. A heavier loading with more positive response options, for example, creates a built-in bias in the item.

A positively-biased response scale

Stimulus:
How happy do you feel about your marital relationship?
Responses:

a. Very happy
b. Happy
c. Indifferent
d. Unhappy

Use judgment in choosing the number of scale points in closed-ended items. Finally, the number of points in the response scale can vary from two to as many as ten. The most commonly used number is five.

On occasion, we might choose to drop the middle option, such as "Unsure," to create a four-point scale that forces respondents to take a position. The use of a "not applicable" or "no response" as a fifth option could yield more valid information in circumstances where segments of the population may not have a position on the topic. The "not applicable" responses would have to be separated from the others during scoring. For populations that do not have the attention span, time, or inclination to respond to demanding item structures, a three-point scale may be the best. Literature in social sciences and health areas suggest that we stay with five to seven options for self-report tools (Babbie, 1990; Fink, 1995).

In theory, more scale points would potentially add more variability to response distributions, a property that improves the quality of the scale from a classical test theory perspective. However, if the indicator is not properly mapped by the scale points, respondents are likely to use only a narrow segment of the scale to respond. Unanchored response points could add further issues. Consider the following item example.

Item: *Did the client consent to participate in marriage counseling? Circle one response.*

Response rating scale: 1-2-3-4-5-6-7-8-9

Based on the wording of the stimulus of this descriptive item, a binary response option (Yes/No) is a logical fit and would suffice. Too many, unanchored response points, as shown, increases the likelihood of whimsical responses, affecting reliability levels. Additionally, having a large number of scale points suggests there is an underlying construct that is being scaled. This false assumption may impede the respondents' abilities to discriminate among *true differences* on the variable' of interest—a validity issue.

Response category usage in individual items should be empirically checked out through pilot tests. The most useful number of scale points can thereby be set. In sum, we need to be alert to the common pitfalls in item and rating scale construction with self-report instruments.

Guideline 7. *Mix in a small number of items with negatively-oriented, or reverse-oriented, wording in a random manner, but use negative language carefully in items.*

A few negatively-oriented items in a tool are useful in preventing response sets or thoughtless, disengaged responses from those surveyed. Items with reverse-oriented wording serve the same function, as illustrated in Tables 7.2 and 7.3.

Research suggests that negatively worded items tend to load on a "semantic factor" when factor-analysed (DiStephano & Motl, 2006) That is, the items with negative words behave like a factor, eclipsing the information from factors on the construct of interest. Hence, the recommendation is to include only a few negatively-oriented items. Evaluating how such items perform before including them in the overall scoring system, should be a prerequisite.

If negative words are used in items, they should be italics or underlined so that there is no confusion in the minds of readers. Double negatives are confusing, and should be avoided. For example, the jingle, " Nobody doesn't like Sara Lee (cakes)" may be catchy, but could well pose measurement difficulties by confusing respondents as to whether they should agree or disagree.

Guideline 8. *Screen all items for the appearance of bias towards relevant social groups (e.g., age, ethnicity, gender, special needs and so on).*

This final and important guideline cannot be overlooked during the design phase or validation of any assessment, not just self-report tools. The overall assessment and items must *appear to be* free of biases during the initial screening and content validation. That is, objective reviewers must be able to state that all groups will be able to access and participate fully in the assessment, if they so choose.

Particular groups or individuals should not be excluded, or placed at a disadvantage, due to the manner in which the instrument is crafted or presented. Fairness and equity are high priorities in the current *Standards* (AERA, APA, & NCME, 2014). Four types of biases could apply.

- Readability bias- This form of bias is manifested when the reading level is too high or poses other barriers to easy interpretation of items by some groups.
- Inflammatory bias – This issue arises if specific language or examples incorporated in items are inflammatory to certain groups or sects, such as, portrayal of religious texts in a one-sided manner.
- Stereotypical bias- This bias makes false assumptions about certain groups that perpetuates stereotypes, such as, boys being better at math than girls.
- Biases due to assessment materials or conditions- This occurs when the conditions format, materials, or platforms through which assessments are administered systematically hamper or prevent participation by some.

Many of these issues arise inadvertently; however, screening, revising, and as needed, deleting items when biases are detected, is sound practice in assessment design endeavours

Insert Table 7.4 about here

Reflection Break

A. Create a "negatively-oriented" item using the domain in Box 7.6. What properties make it effective?
B. State the indicators measured by the items/item sets that follow were designed to tap.
C. Evaluate the quality of the self-report items that follow using the guidelines provided. Explain each issue you identify. In each case, redesign and improve the items, as needed.

1. *Rate the quality of* (a named product) *on safety, utility, and efficiency.*
1 =Poor,
2=Fair,
3=Good,
4=Very Good

2. *Did the teacher demonstrate knowledge of safety procedures during science class? Circle your response.*
   1-2-3-4-5-6-7, where 1=None and 7= Advanced

3. *Items 1-8.To what extent were you trained in the following topics as a part of your professional education? Enter your selected response in the blank to the left of each item.*

_____ *1. Classroom instruction*
_____ *2.Child development...* (6 more topics listed)
Choose one of the following responses.

a. Inadequate
b. Fairly adequate
c. Adequate
d. To a great extent but not adequate
e. More than adequate

7.6.2 Specific Guidelines for Interviews

The interview-based method should be a justifiable match for the targeted content, behaviors, and conditions in the domain, as well as the purposes and populations. Because interviews demand a lot of time from the designer, interviewer, and scorer (usually the same individual in local settings), we should select this method when it is the best fit for construct indicators, purposes, and populations.

Interviews in cognitive domains work best where examiners wish to check for deep understandings and reasoning processes, such as, in the defense of a doctoral thesis. Informal cognitive interviews are particularly well suited for *formative* assessment purposes. In such applications, teachers might wish to embed one-on-one questioning of students with their day-to-day instruction. Interviews generate valuable information on individual student misunderstandings that can be easily overlooked in large group settings.

Some factors to consider when designing more formal interview-based assessments are the following.

- What instructions should we give the respondent/examinee to help him or her understand what we expect?

- Once a question is asked, how much time should we allow before we move on to the next question?
- How many probes will be allowed after each prompt?
- What is an acceptable probe?
- When the interviewee rambles, how should we bring back the focus to the interview's purposes?
- What procedures will we use to code and score responses for a) cognitive constructs, and b) non-cognitive constructs?

Well-designed questions, probes, coding guidelines, and scoring rubrics must accompany interview-based assessments. Formal interview examinations demand that we prepare instructions for both the interviewers and the scorers.

7**.6.3 Assembly of Self report Tools**

The final steps in compiling an instrument concern its assembly. These require careful attention to several details. Refer to the two illustrations in Tables 7.2 and 7.3 as we review some basic guidelines.

Assemble after a preliminary tryout and content validation: It is always a good idea to ask content validators to review all aspects of an instrument during Phase IV of the Process Model, including an early version of an assembled tool. This approach allows all parts of the tool to be refined through an objective, external review. Pilot-testing is always recommended with self-report items and tools, even when empirical evaluations can only be conducted on a small scale. Assembly of the final versions should ideally occur after.

Ensure the Directions, Layout, and Presentation are clear: A key responsibility in this step involves writing clear directions for respondents, users and assessors. The layout of all instruments should be easy to follow for assessors and respondents. Finalizing the layout and presentation of the tool is as important for containing errors, as is the item-writing.

Create the "Background" section: Although not illustrated in the example, adding a short section to collect background information on respondents is often a necessity. Typically, this section contains mainly descriptive items, and is placed in the front or at the end of the instrument. Commonly, items seek information on demographic variables, such as, age, gender, nationality, race/ethnic roots, native language, occupation, income, education/ grade level and so on.

Relevant guidelines for item-writing rules, directions and layout would apply to this section, too. The purposes for the background section should be made clear to respondents. The section should be short, focusing on critical context variables only. It should not distract respondents from the purposes of the main instrument, nor take too much time.

Directions for Sections: For written questionnaires and interviews, general directions are a must. In addition, some questions might require specific instructions, as when the response scale suddenly changes for a series of items, or the stem is common for a series of items. Questionnaires with many parts require separate instructions for each part.

Titles and Sub-titles: The practice of using brief but descriptive titles or subtitles for sections of a questionnaire is also useful for orienting the respondent to what lies ahead. This strategy could enhance the validity of responses obtained.

In other instances, however, titles or sub-titles on surveys are best avoided. An example would be in the measurement of constructs dealing with a clinical condition (such as, a phobia), where the use of a title might encourage faking or unconscious denials from respondents.

Adding an Interview Script: An ideal interview script includes very detailed directions on every aspect of conducting the interview—such as, how to present questions and probes, pauses allowed in between questions, how to handle unexpected responses, and how to score the instrument. Complete scripts begin with a greeting and end with an expression of thanks to the interviewee for their cooperation. If a highly standardized interview delivery is desired by the designers/users, the entire script should be prepared in advance for interviewers to follow in a verbatim manner.

Assuring Anonymity and Confidentiality: This step is critical as it may help curtail response avoidance, faking, non-response, or other forms of respondent-generated biases. Salient concepts are now defined. *Confidentiality* is communicating to respondents that results will not be shared with anyone other than a select group of individuals whom the respondents know and can trust. Making sure that the confidentiality is maintained is equallys important. *Anonymity* deals with asking respondents to conceal their identities, thereby erasing fears that how an individual responds can be linked to names, or perhaps, be used against them.

Obtaining Informed Consent refers to a two-step, formal procedure. In the first step, the assessor communicates to respondents the purposes for assessment and how the results will be used. In the second step, respondents sign an agreement indicating that they understand and are participating voluntarily.

Insert Table 7.5 about here

7.6 Summary

This chapter dealt with design principles that apply for two closely-related types of self-report assessments, survey-based and interview-based tools. Ideally, these tools should be applied when direct questioning of defined groups of respondents is the best or only means for learning more about construct(s).

Self-report assessments were defined as a general type of measuring tool where individuals respond directly to a set of questions presented to them in either oral or written form. Examples of different types of items/item sets in self-report instruments were provided. These included open-ended items, closed ended items, descriptive items, or groups of items that help create a scale for an underlying construct, or operationally define an index. The chapter discussed several multidisciplinary examples, demonstrating when and where each item type is best applied.

Self-report instruments are popular because they seem easy to devise and use. But, the chapter illustrated the specific and many challenges to designing sound self-report items. Given the general measurement issues surrounding self-reported data, scientific standards and guidelines should be followed to extract information of the highest quality from the tools,.

Selected historical designs of self-report assessments and items were presented. While "fact-based" descriptive items were used in censuses from the earliest times, scales and scaling methods were introduced by early psychologists only in the early 1900s. Two historical scaling techniques had a lasting impact in measurement of constructs in the social, educational and health sciences: Thurstone scales and Likert scales. To a lesser extent, Semantic Differential Scales have also been used. Of these, Likert-scaled items continue to be widely employed across fields today.

The case application of the Process Model in this chapter dealt with the design of dual, but complementary assessments in the interview-based and survey-based modalities, starting with a common set of assessment design specifications. The applied setting dealt with screening in clinical counseling contexts for college-going adults. A domain sampling method was applied to create "parallel" sets of items in the two assessment modes.

The chapter presented detailed guidelines on how to design items for each instrument type, discussing common pitfalls that occur during item design, and techniques to recognize and correct the issues through thoughtful review and validation. Examples of well-written and poor items were provided to illustrate each point. The chapter ended with suggested guidelines for assembling self-report tools. Table 7.5 synthesizes the main guidelines in the form of a checklist that readers could use to design, select or critique self-report items and instruments.