Chapter 4

What to Assess: Specifying the Construct Domains for Human Abilities, Dispositions and Other Attributes

4.1 Chapter Overview

A critical requirement for sound assessment design or selection deals with having clarity about "What" we would like to assess. How well we are able to specify construct domains directly affects the quality of items and tasks that we design or select thereafter, ultimately affecting the overall construct validity of the information produced by assessment tools. Specifying the constructs and construct domains with defensible and observable indicators should, in fact, be the very first responsibility of assessment designers. It should also be a guiding concern for researchers and assessment users who wish to select the best-matched items, tasks or from existing instruments for meeting their information needs in particular contexts of assessment use. Chapter 4 is concerned with this important topic.

Figure 4.1 shows how Chapter 4 connects with the rest of the book and the Process Model. The specific topics covered here are represented in the "What to assess" portal, leading into the first bullet under the "How to assess" portal in Phase II on specifying domains, sub-domains, and indicators. Details of the complete process are given in Figure 1.7, Chapter 1.

Figure 4.1 about here

[Process Model graphic reinserted in minimized form, with What to Assess under Phase I and bullet in Domains, sub-domains and indicators in Phase II in relief/contrast. *]*

Figure 4.1 *Connecting Chapter 4 to the Process Model and the rest of the book*

4.1.1 Chapter 4 Objectives

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

1. Explain the main tenets of domain sampling theory as applicable to specifying construct domains.

2. Locate relevant knowledge bases and data sources (literature-based, theory-based, expert-based, or other), to specify indicators for given constructs.

3. Distinguish among constructs with simple, stratified, ordered or unordered domain structures.

4. Apply suitable taxonomies to clarify the substantive nature, levels, or types of indicators and sub-indicators that define a construct domain.

5. Given a construct to assess, write and organize indicator statements for specifying the domain and sub-domains using appropriate guidelines, techniques and tools.

6. Evaluate the quality of indicator statements and overall domain specifications (e.g.,

poorly-specified or under-specified domains) using appropriate guidelines and conventions.

7. Explain the relationship between well-specified domains, validity and reliability of the information obtained from items, instrument and construct measures produced.

8. Define key concepts and technical terms. [See terms in bold font and glossary].

"Generosity is ascribed to a person who gives more frequently or regularly than others, or who gives larger amounts or amounts which are larger in relation to his income or other obligations… The actions in each case define the attribute".

E.E. Cureton (1951, p. 152)

4.1. Foundational Concepts: Specifying Construct Domains

4.1.1 Domain Sampling Theory

A first step in operationally defining a construct deals with specifying the domain with "indicators". Indicators are observable signs or descriptors of behaviors that serve as direct or indirect signals that the underlying construct that we are trying to tap, exists. Where would one begin and end the process of identifying and writing descriptive statements that represent hypothetical or still ambiguously defined constructs? Figure 4.2 illustrates a useful approach drawing on a school of thought called **domain sampling theory** (Tryon, 1957; Ghiselli, Campbell & Zedeck, 1981; Nunnally & Bernstein, 1994).

To appreciate the notion of domain sampling, consider the opening quote above carefully. We initiate the domain specification process by giving a label to the construct we would like to assess (e.g., *Generosity*), and by setting a boundary around a theoretical universe of all possible observable human behaviors and indicators that would denote the degree to which it may be present in members of the population. This hypothetical universe of indicators is referred to as the "**domain".** The assessment instrument is then viewed as a **sample** of all possible indicators and items that could, in theory, define that domain.

To begin with, the concept of *Generosity* may appear vague and abstract with unlimited possible conceptions and construct definitions. However, once we specify a domain that is based

on some established or acceptable knowledge bases with known limits, we could obtain a sense of the depth and breadth of human behaviors/indicators that would define it. The next steps in measuring the construct can then be approached more systematically, reasonably and manageably.

Domain sampling allows us to conceive of a construct as a collection of observable responses, behaviors, words, or actions--the "indicators"--all having some property in common. For example, the domain for *Generosity* mentioned above, is defined by a common predisposition of people wanting to give to others. The items could be deliberately crafted situations and exercises that help elicit those observable indicators of *Generosity,* while offering a credible sampling of the content covered by the theoretical domain.

Unlike probabilistic models of sampling, however, indicator/item sampling procedures that draw on the domain sampling model rely mainly on judgment-based, logical processes performed by assessment designers, experts and researchers involved. Hence, the more clearly and unambiguously we are able to specify the indicators—ensuring coverage of all relevant dimensions and components of a construct--the better we can design/select items or instruments that are a faithful reflection of the domain (AERA, APA, & NCME, 2014).

Once the domain is specified with indicators, alignment of the indicators to items is typically verified through the process of **content validation,** mentioned when presenting the Process Model in Chapter 1. Content validation is typically carried out through objective reviews, interviews or surveys of qualified experts and informed individuals who are external to the assessment design processes (Guion, 1977; Schmeiser & Welch, 2006).

From a practical standpoint, it would be impossible to include items in an instrument that represent every single indicator of a domain. Regardless, thinking of the tasks or items as a smaller, but random sample of the content defining some larger universe does make available several useful ideas on statistical inference and probability. We can then think of making extrapolations from a person's construct measure (e.g., raw score or a derived score) to the larger theoretical domain. Several mathematical formulations of score reliability and generalizability start with this basic premise, drawing on principles of domain sampling theory (Crocker & Algina, 2006; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). More on quantitative methods of reliability estimation will follow in Chapter 10.

To summarize, the domain sampling model of assessment design treats the term "domain" as analogous to the term "population" in the social and behavioral sciences, where the population units are actually the observable "indicators" representing the underlying construct. Eventually, the items on a test/assessment that are meant to tap into those indicators, serve as the "sample" of the universe of indicators. To be credible, as pointed out, the content of the indicators must have a scientific basis, or connections to some established theory, formal literature, expertise-based or agreed-upon knowledge base. Further, indicators must be stated in adequately clear terms, so that the best-suited items and task formats can be created or selected for the assessment.

Insert Figure 4.2 about here

4.1.2 Different Kinds of Domain Structures

Most construct domains are conceived to have either a **simple** (non-stratified) or **stratified** structure. For example, a domain tapping into only addition skills in math with two digit numbers would have a simple structure. Within simple domain structures, all tasks would deal with one common theme, with tasks viewed as interchangeable in terms of content and difficulty levels. Such a domain is said to be **homogeneous**.

In contrast, a domain tapping into a wider range of arithmetic skills might be conceptualized with a **stratified** structure. For example, a stratified math domain could be composed of four qualitatively different **sub-domains** in terms of content and skills**:** addition, subtraction, multiplication and division.

Figure 4.2 illustrates a stratified domain structure for *Generosity,* with five hypothetical strata or sub-domains. Each sub-domain in this instance would be defined by observable indicators of generosity that were substantively different, with each potentially yielding a sub-scale or dimension score. For example, if we categorize generous acts demonstrated by people in family, workplace, larger community, religious, and in national/global contexts separately, we would have constructed a stratified domain framework for generating five potential sub-scales. While each sub-domain would be homogeneous within itself, there should also be sufficient content overlap to tie these together under a common, theoretical framework representing *Generosity*.

Construct domains can also be conceived of as **ordered domains**. Whereas unordered domains are homogeneous**,** with the indicators (and items generated thereof) viewed typically as equivalent in terms of value or weight, ordered domains are deliberately hierarchical or

developmental in structure. Ordered domains allow designers to make distinctions between indicators and items that may be progressively more difficult or higher in intensity levels.

For example, if the universe of content and skills falling under a broad domain of *Mathematics Proficiency* is graduated by difficulty at primary, intermediate, middle and high school levels and these levels correspond with changes in subject matter complexity, we would have specified an ordered domain for the construct. Cognitive assessments based on ordered domains are useful in mapping individual development and growth over time in given competency areas or subject area domains.

See Boxes 4.1 and 4.2 for two examples of ordered and unordered domains, respectively with corresponding indicator statements for two constructs. More discussion of these boxes will follow.

Reflection Break

Recall the Likert versus Thurstone approaches to attitude measurement in Chapter 2. What domain structures did each have in mind—ordered or unordered? Discuss.

Insert Boxes 4.1 and 4.2 about here

4.1.3 Advantages of the Domain Sampling Approach

Although it not the only approach one could adopt for designing assessments, domain sampling techniques are useful in conceptualizing and measuring constructs in any theoretical or

applied field, and is the recommended approach in this book. The approach is evident in a majority of large scale, educational assessment projects, both old and new.

By providing *a-priori* construct domain structures, it is not only possible to design or select tasks/items that match the critical content dimensions of a construct, but also validate the instruments and response data obtained from tests and assessments using that theoretical domain as a framework. A routine question during validation deals with the extent to which the structural patterns in respondent data obtained from instrument/scales, once analyzed, fit the domain structures given by some theory base. Having a domain in hand prevents blind, and often inconclusive, "trial and error" explorations about the underlying structural properties of assessments with unanswered questions remaining as to how best to interpret the construct measures they yield.

The domain sampling approach also helps us begin the design process with a bounded construct universe. Recall it was that Tyler (1949) who first demonstrated how to define the boundaries of a scholastic achievement domain in a Table of Test Specifications (see Chapter 2). From the cells of such a table, he showed how items tied to different curricular objectives could be sampled systematically for building classroom examinations (see also, Schmeiser and Welch, 2006, p. 317). From a pragmatic standpoint, setting a border in measuring a construct helps contain the design process within those limits.

When specifying achievement domains, "indicators" are typically the curricular objectives or statements of expected **learning outcomes** or **competencies**. For a recent application of domain sampling, you might check out the mathematics domain and task/item specifications of Common Core State Standards assessment programs that were released in 2015

to the public (https://www.smarterbalanced.org/wp-content/uploads/2015/08/Mathematics-Content-Specifications.pdf; retrieved on 10-15-17*).*

4.1.4 Alternatives to Domain-sampling: Empirical Methods of Scale Development

There are other approaches to instrument development, however. A widely used approach originated in psychology and is sometimes referred to as the "empirical method of scale development" (Comrey, 1988, p. 754). In this method, the design process does not start with specifying a domain with observable indicators. It begins, instead, by compiling pre-existing items or fashioning new tasks based on common sense and judgments of researchers or assessment designers in particular fields. In the empirical method, superficial characteristics of items are relied upon to devise an instrument. The emphasis is placed instead on the observed properties of the scores or the item response data that result, rather than a conceptual structure underlying the construct domains. For designing instruments intended to predict individual performance on a future criterion, for example, what would matter more is the empirical correlation of the scores with some chosen criterion measure, regardless of the configuration of the domain.

Chapter 2 mentioned the origins of the factor analysis tools as yet another "empirical method" in psychological measurement. Effective use of factor-analytic methods to select item collections from available instruments to build new or different scales is an empirically-driven method of scale development. In the field of personality psychology, for example, factor analytic approaches were dominant empirical methods relied upon for instrument design purposes, used frequently to establish conceptual distinctions among similar and dissimilar sets of items (see Briggs & Cheek, 1986; Comrey, 1988).

Reflection Break: Application Exercise 1

What do you see as the pros and cons of the domain sampling versus the empirical methods of assessment design? As an assessment designer, which one would you prefer to adopt, and why?

List two useful ideas from domain sampling theory.

Give examples of tests or assessments that you know of, with the following types of domain structures: stratified, nonstratified, ordered, or some combinations of these.

Recall Joseph Mayer Rice's "common sense" approach to making up a spelling test and Edward L. Thorndike's analysis of his spelling words (Chapter 2). If you were designing that particular test for 5th-8th graders attending school, how would you approach conceptualizing the "spelling ability" domain?

Are you familiar with any other approaches to assessment design in other fields, outside the two discussed? Share and give an example.

4.2 Domain Sampling and Measurement Quality

As pointed out, how well we are able to specify the domain has a direct influence on both the validity and reliability of the resulting construct measures. This section will highlight this relationship.

4.2.1 Domain-sampling and Content-related Validity

If we are to claim overall construct validity of measures, evidence of **content-related validity** is absolutely necessary. Indeed, it is the first kind of evidence that we should seek when designing instruments. Such evidence is obtained by evaluating the domain indicators, items and the overall assessment instrument for levels of (a) **content relevance,** and (b) **content representativeness** vis-à-vis any existing theory or knowledge about the construct (AERA, APA & NCME, 1999; 2014). Together, such evidence helps determine the degree to which measures produced, such as, the raw scores, derived scores or aggregated scores when people are grouped, will have content-related validity—or **content validity,** hereafter**.**

By following a domain sampling approach, both these properties--content relevance and content representativeness--can be built into the assessment design process in a more methodical and deliberate fashion. To understand these twin indicators, let us revisit the *Generosity* domain in Figure 4.2.

Content relevance refers to the extent to which the indicators and sampled items/tasks are substantively rooted in an agreed-upon body of knowledge about the construct(s) of interest. When attempting to assess *Generosity*, for example, we should ensure that the domain is based on a strong theoretical base relevant to human behaviors that fit widely accepted definitions of the concept of "generous" in psychology or other appropriate literature sources. Similarly, when trying to measure other constructs, say, *Job Competence* in a professional area like nursing, we should be able to claim that the specified domain is a faithful reflection of established knowledge about what expert professionals in that field typically know and are able to do on the job.

In contrast, the notion of content representativeness has to do with ensuring that we have represented all the strata or levels of a domain adequately in terms of content. If a construct is best represented by a stratified domain, a sufficiently large item sample should be taken from each relevant content stratum; if it is a hierarchically organized domain, the number of items should be representative of all levels that define it. As suggested in Figure 4.2, the indicators and items should ideally be *proportionately* distributed or balanced in a reasonable manner across different strata or levels of a domain.

If existing psychological theory on *Generosity* suggests a domain structure with five equally balanced and homogenous strata, then a content- representative assessment would sample indicators/items proportionately from each stratum. Similarly, a content-representative

assessment of nursing competence would show coverage of all relevant sub-domains of nursing expertise, which may cover areas like knowledge of medical procedures, skills in patient care, and so on. There are quantitative indices of content validity that will help summarize results of a content validation study, called the **Content Validity Index** (CVI). More on that will follow in Chapter 9.

4.2.2 Reliability and Domain Sampling Theory

The domain sampling model of assessment design is also linked to reliability formulations based on Classical Test Theory (CTT), and extends to the broader framework of **generalizability theory** (Cronbach et al, 1972; Crocker & Algina, 2006; Shavelson & Webb, 1991). Once we accept that an assessment is a sample of items from a defined domain, we can think of the raw score or the average score for persons, as a quantitative index of that universe. Recall that the raw score for individual respondents/examinees (representing the construct measure) is simply the summed responses to the assessment items.

From a domain sampling perspective, the number of items generating the raw score matters a great deal because the size of this item sample determines reliability levels. If we have a smaller sample of items or behaviors--say, 2 or 3--the raw score is likely have more sampling error associated with it, leading to less reliable scores. Conversely, increasing the number of items to 5, 10, or 20 will tend to reduce the sampling errors. Long-standing psychometric research shows that quantitative estimates of reliability using CTT methods are related to “test length” based on item numbers until we reach a point of no returns (Allen & Yen, 2001 after Gullicksen, 1950; p. 96).

At the same time, could one take an unreliable instrument and increase the reliability estimates by simply increasing its length with a hodgepodge of items? Not necessarily. Only if the items are homogeneous in terms of content and in that sense, "parallel", will we be likely to find that halving the number of items on a test, will generally lower the corresponding reliability estimates. More on the relationship of test length and reliability will follow in the Chapter 10 (Crocker & Algina, 2006; Ghiselli, Campbell & Zedeck, 1981; Allen & Yen, 2001).

So, a general rule-of-thumb is: We should avoid building assessments with a thoughtless and heterogeneous mix of items. The aim, rather, should be towards establishing substantively similar items within domains/sub-domains that yield meaningful and reliable scores. Domain sampling ideas are helpful to achieve these ends.

Reflection Break: Application Exercise 2

Identify examples of instruments that tap into (a) simple (non-stratified) domains, (b) stratified domains, and (c) ordered domains. In your view, to what extent was the targeted construct appropriately conceptualized in each case?

Review the Process Model in Figure 4.1. Does it lean more towards the domain-sampling or empirical methods of scale design? Explain.

4.3 Techniques, Tools and Conventions for Specifying Construct Domains

Now, we turn to some "how to" steps, tools and techniques for specifying construct domains situated in the Process Model, excerpted in Figure 4.1. Simultaneously, examine Table 4.1, which outlines the three main steps that we should follow. Boxes 4.1-4.4, Figure 4.3, and Table 4.2, illustrate with examples how each step may be applied.

To start, for ease of communication, the section will classify constructs in one of four categories. This classification should be qualified in light of new knowledge evolving in the neurosciences showing the human brain to be the physiological/biological source of most human

responses and behaviors. Given that, it is offered here merely as a practical tool to facilitate instrument design processes which vary depending on the construct type. The categories are:

- **Cognitive constructs**, dealing with one's intellectual skills, mental capacities, cognitive processes, and knowledge levels in a domain;
- N**on-cognitive constructs**, dealing with one's social-emotional mind-sets and dispositions related to a domain;
- H**ealth-related constructs**, dealing with one's state of physical, physiological or psychological well-being, disability or disease/disorder; and
- **Social and demographic constructs**, dealing with societally-defined characteristics of individuals, groups or entities (e.g., ethnicity, nationality, socio-economic status, religion).

4.3.1 Step 1: Locate Appropriate Resources and Knowledge-Bases for Specifying Construct Indicators

Once a construct is identified and situated in its assessment context in Phase I of the Process Model (Figure 4.1), an assessment designer's first task is to locate credible sources from which to derive the construct's indicators. In many cases, a quick google search might suffice to get us started. In other cases, added research and alternative data sources may be necessary. This section discusses three data sources useful for domain specification purposes: (1) scientific research, literature reviews/syntheses, and documentary sources; (2) expert practitioner perspectives; and (3) direct observation methods and case studies.

Using existing theory, research, literature reviews and documentary sources. Such sources could include one of more of the following: documents published by national or

international professional associations in a relevant field; the mission, goal or policy statements of relevant organizations, institutions or programs; current textbooks or curricula in a relevant field for a particular level; and scientific, professional and academic literature in a relevant field. **Content analysis** of such documentary sources provides a way to identify indicators of a domain. Two illustrative examples follow.

Refer to Box 4.2. Madni, Baker, Chow, Delacruz and Griffin (2015) sought to identify and describe the *non-cognitive* domain of "social-psychological attributes" of effective teachers—that is, teachers who succeed in improving their classrooms and student outcomes. They identified four broad sub-domains through examinations of existing scientific research and literature, identifying general indicators to define the construct that are shown in Box 4.2. These were a teacher's personality characteristics, motivational attributes, intra-personal skills, and inter-personal skills.

As conceptualized, the domain had four broad strata (Madni, Baker, Chow, Delacruz & Griffin, 2015). The authors then further categorized each stratum with theory- and literature-based sub-indicators. For instance, the Personality sub-domain was operationalized based on the five factor theory of personality (Olver & Moorardian, 2003), which suggested that there are five dimensions of human personality that would also apply to effective teachers: Conscientiousness, Emotional Stability, Extraversion, Agreeableness, and Openness to Experience.

For a second example that applied this method, see Box 4.3. Here, the researchers were interested in operationalizing a cognitive domain given by the Accreditation Council for Graduate Medical Education (ACGME), labeled as Practice-Based Learning and Improvement (PBLI). The PBLI domain is a two-pronged, somewhat ambiguous competency area that was

not clearly delineated in medical education curricula and the medical literature at the time of this research project. It was expected to include core knowledge and skills in an area called "evidence-based medicine (EBM)" (Guyatt, Rennie, Meade & Cook, 2002), as well as practice-related behaviors of physicians relevant to lifelong learning with attention to patient care.

To define the domain shown in Box 4.3, the research team used a variety of existing documentary data sources, such as, established standards and guidelines in the medical profession, curriculum goals and criteria set by the accreditation agency, as well as established literature and EBM Users' Guides as the main data sources. In addition, they sought perspectives of medical and EBM experts to fortify the indicators and sub-indicators of the construct domain.

As the excerpted domain specifications show, the indicators are the expected competencies and sub-competencies of competent physicians. There were four action areas where doctors were expected to demonstrate these behaviors: "therapy", "prognosis", "diagnosis", and assessing potential "harm" or harmful side-effects of prescribed treatments (Authors, 2013; Authors, 2009). These action areas were treated as the sub-domains of the larger PBLI domain.

Insert Box 4.3 about here

Using Expert or Practitioner Perspectives. Sometimes, a construct that needs to be measured is either unknown or not yet formalized in the existing literature. For example, in light of recent episodes of terrorist and large scale violence in the U.S. involving internet-based or social media operators, a need arises to define and identify through assessments, what is and is not "deviant, anti- social behavior" on the internet. This still-undefined construct has relevance today for further study and policy-making.

In cases such as the above, several methods exist for eliciting knowledge through direct interviews and facilitated group processes with expert panels. Experts useful for domain specification purposes are individuals who are either highly educated, knowledgeable or otherwise experienced professionals in a topic and field, and deemed to have deeper insights of the issues related to the constructs of interest. Widely used methods in this category are the Delphi dialogue method, focus group interviewing, and the nominal group technique processes. A description of each follows.

The **Delphi method**, originally developed as a systematic, interactive forecasting method, relies on a panel of experts (Dalkey & Helmer, 1963).The experts answer questionnaires in two or more rounds on the hypothetical construct (or topic). After each round, a facilitator provides an anonymous summary of the experts' ideas from the previous round along with rationales for those judgments. In the second round, participants are encouraged to revise their earlier answers in light of the facilitator's feedback until the group converges towards consensus. Delphi is based on the principle that consensus-based decisions from a structured group of qualified individuals are more accurate than those obtained separately from each expert.

Focus group interviewing is a form of qualitative research in which a carefully selected group of people is asked about their perceptions, opinions, beliefs, or attitudes about given topics, such as, deviant internet behavior (Greenbaum, 2000). The interviews are conducted through guided discussions. In market research or political analysis, a focus group is a small, but demographically diverse group of people whose reactions are studied to determine their reactions about a new product or policy so as to sense the possible reactions of a larger population.

Questions in focus groups are typically asked in an interactive group setting where participants are free to talk with other group members. During this process, the researcher either takes notes or records the chief points; the data are then categorized qualitatively by theme and sub-theme to generate answers. As such, the method can be useful in crafting construct domains in unknown areas.

The **Nominal Group Technique** is another group process involving problem identification, solution generation, and decision making in an area, to arrive at consensus-based indicators of a new or different construct (Delbecq & Vandeven, 1971). The method has been used in curriculum design and evaluation contexts, as well as in social policymaking. The expert groups may be of different sizes, but the interest is in making a decision quickly with everyone's opinions taken into account.

To start the domain specification process using the nominal group approach, every member of the group would give their view of the domain structure and indicators, with a short explanation to justify the recommendations. Then, indicators are reviewed so that duplicates are eliminated from the list. Next, the commonalities and hierarchies are organized in the form of a taxonomy. Members could also rank order the indicators in terms of priority or weight.

Here is an applied example. Graham and colleagues (2009) used a variant of focus group interviews and nominal group techniques to identify the key indicators of a relatively unexplored medical competency area called Systems-based Practice (SBP), also given by ACGME. Their goal was to operationally define resident physicians' competence in SBP in terms of observable roles, actions, and behaviors.

These researchers collected data from a series of group meetings using structured interviews of 88 health care professionals working in various roles in large hospitals and healthcare systems in New York City---from doctors, to nurses, to technicians and patient care support staff, and finally, to administrative staff. Their methodology involved coding of themes obtained from the two procedures, which were conceptually matched and organized to create a taxonomy of observable behaviors defining the SBP domain. The domain served as the basis for designing observational assessments to rate resident physicians' competency levels while they underwent training. Expertise-based methods were found to be especially helpful in this case, as existing data sources and literature were either lacking or nebulous on the SBP construct; however, for effective group processes, it was vital that members were authoritative and experienced experts.

Concept-mapping is another method that could be useful for grouping ideas on unknown or still-undefined constructs with facilitated expert group processes, followed by appropriate qualitative and quantitative analysis of the data (Rosas & Kane, 2012; Goldman & Kane, 2014). Here, expert participants could start by brainstorming a series of descriptive, representative statements on a construct-related question, such as, "What are the signs of socially-deviant behavior on the internet?". Next, all statements are clustered, sorted in piles, and quantitatively rated by experts to indicate similarities and dissimilarities. To verify commonalities among indicators based on the ratings, multivariate statistical methods like cluster analysis may be employed to identify similar clusters of statements. In a final step, the expert group helps interpret the clusters, thereby creating a domain and sub-domain framework for the construct.

Direct Observations, Critical Incident Techniques, and Case Study Research. Other ways to gather relevant data to define domains for constructs that are still relatively undocumented or unknown, involve observational and case study methods. Three common ones are discussed next.

The **Critical Incident Technique**, a method based on direct observations and reporting by key informants, is useful for identifying extreme behavioral indicators of a given construct on a performance continuum. A key informant is an individual with first-hand knowledge of a topic or situation. An early study by Flanagan (1954, as cited in Crocker & Algina, 2006, p. 68) employed the method by asking job supervisors (the informants) to define "critical behaviors" representing outstanding performance versus completely ineffective performance of workers on the job in given workplace settings.

Case study methods, involving direct observations, recording, and cataloging of symptoms and behaviors of patients suffering from still-undocumented diseases or disabilities, can be similarly useful in defining health-related construct domains. To enhance credibility, the observations must be made by clinically trained professionals and, in the ideal scenario, corroborated over multiple cases (Stake, 2013).

See Box 4.4 for an example of how this might occur. Mukherjee (2010) describes how the study and deeper understandings of leukemia—the construct, in this example-- evolved from the direct observations of patient cases recorded by highly committed doctors and cancer researchers. Doctors and scientists learn about diseases—the construct here-- by observing, documenting, cataloging, and classifying the symptoms. Such collective learning and documentation over time helps knowledge about a disease evolve. In essence, the work of

specifying a construct domain is similar to identifying, classifying, and verifying such observable symptoms.

Once catalogued, the symptoms serve as the indicator framework that can then be used to design or select optimal measurement methods. In the case described, counts of white blood cells in the blood surfaced as the most useful observable indicator of the disease, providing the gateway to assessment, and subsequent measurement and diagnosis of the condition in patients.

Reflection Break

If you are designing an instrument to measure knowledge and skills of *effective teachers* after they complete a training course, which data sources would you seek to specify the domain?

If you are designing an assessment of *sleep apnea* in 5 year old children, a health condition where they have breathing difficulties when they are asleep, which data sources would you use to specify the domain?

What are the advantages and disadvantages of each of the data sources and methods discussed, for deriving domain indicators? Which would you apply in your own assessment design project? Explain.

4.3.2 Step 2. Derive, Write, and Organize Indicators using Guidelines

Deriving and Writing Indicators. Because most constructs are typically big concepts comprised of multiple, sometimes layered components and dimensions, we may have to start with a general and vague collection of brainstormed ideas and themes extracted from the data sources we review and analyze in Step 1. The task then is to bring clarity to these themes by restating them as indicators using established guidelines and tools, and then organizing these statements into some reasonable form that is useful for assessment design/selection. The same guidelines apply when specifying the domains for all types of constructs: cognitive, non-cognitive, health-related, or social/demographic.

The first general guideline is that indicators of a given domain are stated and organized from general indicators that may be less directly observable, to more and more specific ones that are. Any vagueness inherent in the definition of a construct domain is thereby removed. It is therefore always a good idea to state indicators starting with "actionable" verbs representing the behaviors, words, actions that we could expect to see from individuals when they respond to items or tasks. For example, while "Communicate", "Perform", or "Differentiate" convey directly observable behaviors that can be tapped through items, "Know", "Understand" or "Feel" are less so.

Indicators could be parsed into four key parts: Process/behavior, Content, Condition, Criterion performance. For example, with a cognitive construct like mathematics proficiency, an indicator could include all four:

> *Students calculate* (Process/behavior) *the sum of a given sequence of whole numbers, fractions, or decimals* (Content) *without the use of calculators* (Condition) *with 80% accuracy* (Criterion performance).

Each indicator should minimally clarify at least two of the above dimensions: (a) the underlying process/behavior to be captured, and (b) the content. With a non-cognitive construct, such as teacher attitudes towards students, indicators are typically specified with the first two essential components. Here is an indicator to measure teacher beliefs about student learning:

> *Teachers communicate or endorse beliefs that* (Process/behavior*) all students can learn, develop, and grow* (Content)

In Box 4.2, we see the general indicator for another non-cognitive domain for teachers: "In professional contexts, teachers exhibit/display personality attributes of conscientiousness" (first bullet). Here, "exhibit/display" delineates the attitudinal *processes* or *behaviors;* "conscientiousness" identifies the attitudinal *content*. Additionally, note that this indicator also specifies a *condition*, "in professional contexts", indicating situations where the behaviors would likely be observed during assessment.

Should the specification of the teachers' non-cognitive attributes stop with only the general indicators as given in Box 4.2, the domain would be incompletely and inadequately specified. There is still much ambiguity in concepts like "conscientiousness" that would benefit from further breakdowns with specific indicators. What would a conscientious teacher be likely to specifically say or do in professional contexts, based on the literature? These indicators would need specification for facilitating assessment design/selection.

Organizing Indicators. The purpose for this is step to designate indicators in a coherent hierarchy or related clusters, guided by the literature or data sources. This book suggests organization of indicators from general to specific in tree-diagrams, as shown in Figure 4.3. However, concept maps or other methods that show inter-relationships among similar versus dissimilar clusters of indicators, may also be a useful alternative.

See Figure 4.3 for an illustration on how to organize the broad cognitive learning domain, Historical Thinking, with general and specific indicators in the form of a **tree-diagram**. There is a main branch and five sub-branches to the tree. Each indicator delineates the cognitive process/behavior (notice the verbs, such as, "Analyze") and (b) the content (see the objects of the

verb, such as, “Multiple causation”). The sub-branches may need further expansion with added levels of branches, if tasks or items cannot be constructed to tap into them.

In sum, the number and levels of branches and sub-branches necessary in a tree-diagram for specific domains, may vary. For example, with a mathematics proficiency construct, a general indicator in the main stem could be:

- *Solve problems dealing with mathematical patterns and sequences.*

An embedded, second-level indicator in the first branch could be:

- *Apply appropriate arithmetic operations to continue a given mathematical sequence that contains whole numbers, decimals or fractions.*

A third-level indicator in the next sub-branch, could be:

- *Calculate the sum of the terms of a given mathematical sequence that contains whole numbers, decimals or fractions.*

More Examples. For deeper examination of details, let’s refer back to the example of an ordered domain in Box 4.1 again. Here, we see the general indicator, an expected, culminating learning outcome for students, stated as: “Identify, generalize, and explain mathematical patterns and relationships, showing understandings of number sense, arithmetic operations, and geometric principles at intermediate levels of difficulty”. That general indicator encompasses a wide variety of still-unspecified knowledge and skills that may be hard to capture with items, including well-designed ones. In contrast, the more specific indicator at Level 1, representing an embedded competency, provides more clearly defined boundaries of the patterns concepts and

skills expected to be mastered and demonstrated by students at the specified levels of proficiency in that domain.

The verbs "identify", "continue" and "explain" in the specific indicator at Level 1 in Box 4.1, indicate the cognitive *processes* or *behaviors* to be tapped; the "simple repeating patterns" in mathematics, identifies the *content*. As evident from the examples of all the items, such details in indicator specifications make item-writing or selection much easier. Together, the clarity of indicators specified at Levels 1-3 help us envisage the particulars of tasks that would fall at each level of the domain. Once matching tasks or items can be produced, no further clarification is necessary.

Box 4.5 offers six criteria to evaluate the clarity of indicator statements in a domain**.** The goal should be to bring clarity to a previously unfocused, incoherent or ambiguous construct, so as to facilitate effective item-writing or selection. Assessment designers can then opt to tap into indicators individually or collectively with different operations and task formats (see Chapter 3, Table 3.7).

Insert Figure 4.3 and Boxes 4.4-4.5 about here

Reflection Break: Application Exercise

What are the advantages to organizing indicators in a tree-diagram? Disadvantages?

Write three indicators to assess the construct, *automobile driving proficiency*, with:

- content and process/behavior specified
- content, process/behavior, and condition specified
- content, process/behavior, condition, and criterion performance specified

Write two indicators for the construct, *attitude towards automobile driving*, with:

- content and process/behavior specified
- content, process/behavior, and condition specified

4.3.3. Step 3. Finalize Types or Levels of Indicators to Define Domain and Sub-domains

Using Taxonomies to Classify Indicators. A taxonomy of behaviors/indicators is a classification system that assessment developers could use to analyze and obtain a deeper understandings of the types of behaviors or levels of cognitive complexity that they will target for assessment. Classifying indicators using taxonomies helps in the selection of assessment methods and item formats that are the best "fit" for the parts of the domain in which we are most interested. Memorization of facts can be appropriately assessed by a multiple choice item; reasoning and critical thinking, on the other hand, would call for an open-ended assessment method; attitudes and social behaviors might require a different assessment method altogether.

Taxonomic analyses also help in prioritizing and assigning weights to indicators, so that items can be written or selected for indicators that are more heavily weighted than others based on judgments of designers or researchers. In some cases, all sub-domains and indicators carry equal weight; in others, the weights assigned could vary.

The literature offers several taxonomies to choose from as we go about the work of assessment design, each with different degrees of usefulness for different constructs. Table 4.2 provides three alternative taxonomies that may be useful for cognitive, non-cognitive, health-related, and social/demographic constructs, respectively.

A Cognitive Taxonomy. The cognitive taxonomy provided in Table 4.2 builds on the classic *Taxonomy of Educational Objectives* (Bloom et al., 1956), better known as Bloom's taxonomy, and some more recent ones (Gagne, 1965; Marzano, Pickering and McTighe, 1994; Author, 2003; NRC, 2012). In Bloom et al's (1956) system, learning outcomes and educational

objectives are classified into three major areas or "domains" (not to be confused with our use of the term "construct domain."). Bloom's "domains" were broadly construed areas of cognitive, affective, and psychomotor learning. Bloom's *cognitive domain* included educational objectives dealing with the development of intellectual skills and understandings. The *affective domain* dealt with objectives focusing on the development of attitudes, values, and appreciations. The *psychomotor domain* included all objectives dealing with physical and motor skill development. In an early analysis, Krathwohl et al (1964) determined that the biggest percent of objectives in educational curricula of that period fell under Bloom's cognitive domain.

In a more current National Research Council publication discussing 21st century knowledge and skills necessary for college and career readiness (NRC, 2012), the committee distinguished between three broad domains of **cognitive**, **interpersonal** and **intrapersonal** competence. They defined each "cluster", as follows, drawing on advances in research in the cognitive and brain sciences. Cognitive competencies in this framework include those of knowledge acquisition, critical thinking, reasoning, argumentation, innovation and creativity. Interpersonal competencies include teamwork, collaboration and leadership, including communication and conflict resolution skills. Intra-personal skills include intellectual openness, work ethic and conscientiousness, self-evaluation and metacognition skills.

Now, consider the cognitive taxonomy given in Table 4.2, that draws on the preceding sources. This taxonomy was applied for assessing cognitive proficiency constructs in both K-12 educational contexts and in professional education curricula (Authors, 2013). It recognizes four types of cognitive processing capacities, each requiring different types or levels of mental demands: concept recall and understanding is at the lowest level; application is at the next

higher level; and complex procedural skills and higher order thinking and problem-solving skills are the two most demanding levels of cognitive processing.

The categories are cumulative—hence, application level tasks will typically also call for concept knowledge and understanding. Likewise, higher order thinking and problem-solving tasks will typically call for both concept recall and understanding as well as application skills. The suggested verbs for indicators illustrate the types of performances associated with each type/level.

A taxonomic analysis of the indicators of a domain similar to that in Box 4.1 follows as an illustration. To assess cognitive proficiency levels in mathematical patterns in high schoolers, a general indicator might be:

Solve problems involving mathematical sequences using suitable operations and strategies.

This would involve higher order thinking and problem-solving skills, according to the taxonomy.

For students to demonstrate this proficiency, however, some specific indicators subsumed must be considered during instruction and assessment. These are at lower cognitive levels, such as:

Define a "mathematical sequence" (Concept recall and understanding)

Give an example of a "mathematical sequence" using whole numbers (Concept recall and understanding)

Calculate the sum of the terms of a given mathematical sequence (Application)

A Taxonomy for Non-Cognitive Constructs. Non-cognitive constructs are defined as dispositions of human beings towards some attitudinal object, such as, a place, person, event, or an experience, and could include interpersonal and intra-personal indicators. The psychological literature on the attitudinal constructs dating back to the 1960s offers a useful, tripartite taxonomy to help organize, classify, and label indicators in non-cognitive domains (Eagly & Chaiken, 1993; Hovland & Rosenberg, 1960). This taxonomy is updated based on new research and thinking in Table 4.2 (after NRC, 2012), and borrows from the multi-component model of attitude.

Human attitudes are viewed as learned mind-sets based on a person's evaluations of an object that have "cognitive" (note here, this term refers to one's beliefs or perceived knowledge and awareness), affective, and behavioral components. These components–associated with the mnemonic, "CAB" (as in a taxi that will get you where you want to go) are updated here. The taxonomy we will follow is comprised of the the following.

- **Cognitive component of dispositions.** This component represents what individuals perceive they know or hold to be true about some attitudinal object. It involves one's beliefs, perceptions, values, opinions associated with an object, including self-beliefs like self-efficacy (Bandura, 1997). An indicator tapping into the cognitive component of a person's attitude towards a political party, for example, would look like this: *Expresses or communicates values reflecting political party positions on social policies.* An item to match, with a positive-to-negative response continuum, is: *I support a woman's right to choose when it comes to child-bearing*.

- **Affective component of dispositions.** The affective component of attitudes represents feelings and emotions evoked in individuals in relation to the object. An indicator tapping into the affective component of attitude towards a political party, for example, would say: *Responds emotionally to party positions on policies.* A matching item, similar to the above, would be: *People that oppose my party's position on abortion make me angry*.
- **Social-Behavioral component of dispositions.** The social-behavioral component of attitudes represents what individuals might actually do or have done in the past that reflects their underlying stance on the object in social or societal contexts. An indicator tapping into the social-behavioral component of attitude towards a political party, for example, could say: *Supports a political party with personal contributions (monetary, service or time).* A matching item, similar to the above, would be: *I contribute my time to spread word on my party's position on abortion*.
- **Metacognitive component of dispositions.** This is a new dimension based on the cognitive sciences, and involves the skills to self-evaluate one's actions and behaviors and self-correct course to achieve certain ends. This set of skills could manifest while engaging in formal learning at school, or in sundry social-cultural or workplace contexts.

Taxonomies for Health-related and Social/Demographic Constructs. Along similar lines, Table 4.2 provides two taxonomies that could be applied for designing assessments for health-related or social/demographic constructs, respectively, as we have defined these two terms. As health-related constructs deal with one's state of health or well-being, classifying indicators in terms of type of symptoms of a health condition, would assist in item or task writing

and selection. Social constructs may be tapped through indicators dealing with demographic, social class, economic status, religious or geographic/regional background characteristics.

Developing or selecting item or task examples to match indicators. Returning to the questions in Box 4.5, a good check to evaluate how well the domain specified is to see whether the indicators allow us to write/select matching items or tasks when taken either individually or in coherent clusters. If this essential goal is not achieved, then the domain needs further clarification.

Validating domain specifications. The last step involves validation of the domain using informal peer review or more formal expert feedback processes, followed by revision of indicators using an iterative process suggested in Figure 1.7, the Process Model. The purpose of this step is to evaluate and correct for any major gaps or ambiguities that might persist in the domains before embarking on item design or selection. Depending on the formality of the assessment design endeavor, this step can be performed in a more or less extensive manner. For example, content validation of domains for classroom assessments designed by teachers may entail brief peer reviews; for large scale assessment design projects, the step might involve gathering of data more formally from external panels of judges, followed by appropriate analyses. More of techniques of content validation will follow in Chapter 9.

Reflection Break: Application Exercise

Apply the appropriate taxonomy provided in Table 4.2 to classify the construct type and level/category of each of the indicators in italics below. Justify your classification.

- *Compose a story for children.*
  Construct type: Taxonomic category:

- *Behave ethically in the workplace.*
  Construct type: Taxonomic category:

- *Follow rules while driving on main city roads and highways.*
  Construct type: Taxonomic category:

- *Demonstrate 20/20 vision at night.*
  Construct type: Taxonomic category:
- *Subscribe to a particular religious sect.*
  Construct type: Taxonomic category:

- Which of the above indicators are too vague or broad to allow sound item design or selection. If so, how would you clarify it further? Explain.

4.4 Summary

Chapter 4 introduced you to domain sampling theory as the main framework for specifying construct domains and sub-domains before beginning assessment design or selection. It presented several guidelines, taxonomies, and conventions to help operationally define constructs with observable indicators. The chapter demonstrated these methods with examples falling under cognitive, non-cognitive, health-related, and social construct categories.

Domain sampling theory espouses the notion of an assessment as a sample of all potential items and indicators that could theoretically represent a given construct domain. Well-specified domains allow items to be created or selected matched to indicators through logical design processes. When well-executed, this process maximizes the overall construct validity of the items, instrument and construct measures produced, and overall, the content-related validity and reliability of the measures. Domains are specified with indicators using relevant knowledge bases and data sources; these could be literature-based, theory-based, expert-based, documentary or research-based sources from case observations and studies.

Domains can be conceptualized with simple, stratified, ordered or unordered domain structures. The main tasks in defining domains for constructs are: writing and organizing indicator statements following guidelines; applying suitable taxonomies to clarify the substantive nature, levels, or types of indicators and sub-indicators in a domain; writing samples of items to test the clarity of indicators; and evaluating and revising the indicator statements and domain specifications based on peer or external validation and reviews.