CHAPTER 4

What to Assess?
Specifying the Domains for Constructs

4.1 Chapter Overview

4.1.1 Introduction

Chapter 4 marks the beginning of the methodological sections of the book. It discusses the theoretical foundations and methods for specifying construct domains. Well-specified domains help improve the quality of individual items and assessment instruments that we create, select, or adapt for our own specified assessment purposes, impacting the validity levels and overall quality of the assessment results. This chapter details the major techniques, with applied examples of the procedures from the literature.

Figure 4.1 shows how Chapter 4 connects with the rest of the book and the Process Model, with two boxes. The specific topics deal with the *What to assess?* portal under Phase I, linking directly with the first bullet under the *How to assess?* portal under Phase II: Specify the domain(s), subdomain(s), and indicators. Readers should refer back to Figure 1.4 in Chapter 1 for the complete model, as necessary.

4.1.2 Chapter Objectives

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

1. Explain the interrelationships between domain sampling theory, procedures for specifying domains, and the validity of construct measures.
2. Distinguish between instruments conceptualized with simple (nonstratified), stratified, ordered, and/or unordered domain structures.
3. Locate relevant theory, literature, data sources, or knowledge bases to specify domains, subdomains, and indicators for constructs of interest.
4. Write and organize indicator statements when specifying construct domains, using appropriate guidelines and conventions.
5. Apply suitable taxonomies to clarify, categorize, and organize observable indicators of construct domains (e.g., Bloom's taxonomy, the functional taxonomies, or the cognitive–affective–behavior–metacognitive [CAB-M] taxonomy).
6. Validate the construct domains as a step in Phases I–II of the Process Model (using the criteria of content relevance, content representativeness, organization, coherence, and clarity).

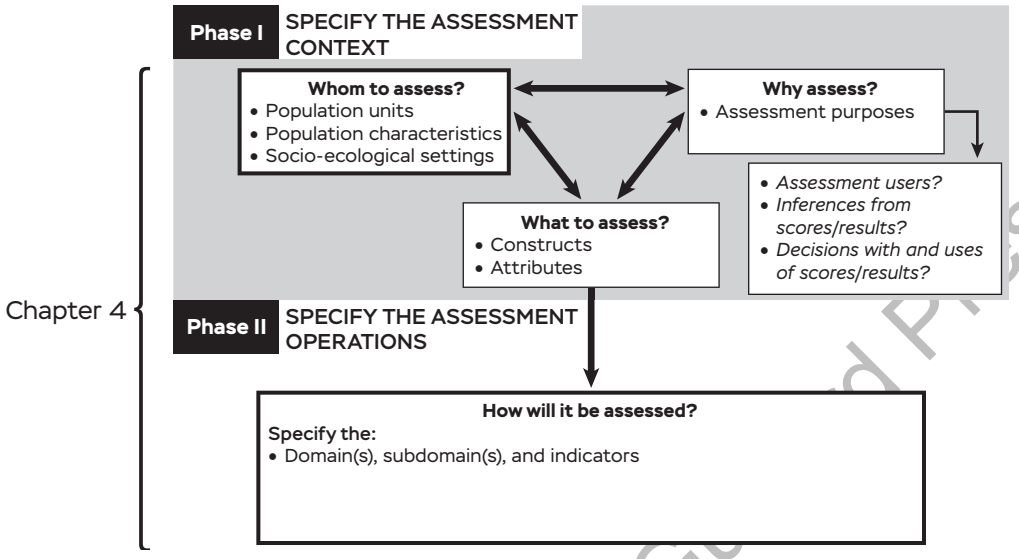

FIGURE 4.1. Connecting Chapter 4 to the Process Model and the rest of the book.

4.2 Domain Sampling and Domain Specification: Foundational Theory and Applied Illustrations

> 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.
>
> —EDWARD E. CURETON (1951, p. 152)

A critical requirement for sound assessment design deals with having clarity about "what" we would like to assess. Chapter 4 focuses on the theoretical foundations and selected techniques for specifying construct domains. Recall that, in the typical case, the constructs we desire to measure are hypothetical and unobservable characteristics of persons, groups, objects, or entities. By measurement convention, we rely on proxy "indicators" that are more directly observable to make inferences about their presence or absence, whether in absolute categories or to different degrees (after Cronbach & Meehl, 1955; see Chapter 1). Specifying construct domains thoroughly and defensibly with observable indicators should be a first concern of all assessment designers seeking to build validity into the instrument design process from the earliest stages. This chapter explains why this procedural step is also informative for adopters and users of preexisting instruments, whether the tools are implemented as is, or excerpted in parts.

4.2.1 Domain Sampling Theory and Beginning Illustrations

Figure 4.2 illustrates an approach for identifying observable descriptors of constructs drawing on a school of thought called **domain sampling theory** (Ghiselli, Campbell, & Zedeck, 1981; Nunnally & Bernstein, 1994; Tryon, 1957). Domain sampling, a recommended approach in this book, could be applied when measuring an array of attributes and constructs in any theoretical or applied field. As forthcoming examples will show, the approach is evident in a vast majority of test development projects in education, both past and present.

Domain sampling theory holds that any given assessment instrument is but a *sample* of all possible items, tasks, or exercises that could be linked theoretically or empirically to the constructs we attempt to measure. It posits that it is possible to operationalize a theoreti-

cally proposed construct as a collection of observable responses, behaviors, words, or actions that the respondents would likely display when given the items taken from a larger universe. The "universe" should ideally be grounded in an acceptable body of knowledge with a discernible boundary. It could be drawn from established theories, scientific research, practice-based experiences, or socially accepted norms, but, for the constructs to be measured credibly, it must be relevant, coherent, and defensible (Chapter 1). The observable indicators of a given construct must also share properties in common rooted in that same knowledge base.

For any given construct, this bounded universe of indicators is referred to as the **domain**. Guided by a domain specified with observable indicators, we could then develop or select matching **item samples** to build an assessment instrument.

These theoretical ideas on domain sampling are illustrated with a concrete example in Figure 4.2. We focus on the attitude suggested in the opening quote: *generosity* (a psychological construct) representing an individual's general predisposition and willingness to give to others. Initially, the concept of generosity might appear vague and abstract. There could be various competing theoretical conceptions of it in the literature, with an infinite number of potential indicators. We might question the few indicators offered in the quote (Cureton, 1951), such as "a person who gives more frequently or regularly than others." Importantly, however, we see three observable *actions* indicating what a generous person would likely do, such as the frequency of their giving. These actions serve as a useful starting point to operationalize the otherwise ambiguous construct. To lend credence, however, we must tie these initial indicators to an accepted knowledge base. For any construct, the choice of a knowledge base rests on the judgments of assessment designers or researchers involved. Once identified, we could expand or narrow the depth and breadth of the domain within some clearly set limits, as well as verify its merits with regard to the substance of the indicators. The next steps in defining the construct can then follow reasonably, systematically, and manageably.

The pictorial illustration shows how the domain sampling procedure would proceed for measuring *generosity*. The picture shows different kinds of circles to depict qualitatively different behavioral indicators. Notice the proportionality in the items matched to the distribution pattern of substantively different indicators for the construct domain. We could either create the set of items for inclusion in the assessment tool or select them from elsewhere. As we apply this procedure, it becomes clear why assessments are viewed as "behavioral samples" indicative of the targeted constructs. Because any given test or instrument can include only a limited sample of all possible items, there is always some amount of error in the test scores or construct measures we produce. Errors rooted in item sampling issues are referred to as *domain sampling errors*. During validation, we would evaluate the extent to which these and other sources of error interfere with the quality of the construct measures produced and our intended inferences and uses of the same.

Starting with the domain specification phases in Figure 4.1, how could we ensure we have sufficient levels of **content-based validity** in the measures? Indicators for a targeted construct should specify two components clearly, namely, the *content* and *behavior* dimensions. For example, one observable indicator of a generous person is someone, as Cureton posits, "who gives larger amounts or amounts which are larger in relation to his income or other obligations." Here, the verb "give" suggests an expected action; this is the *behavior* specified for an indicator of *generosity*. The phrase "larger amounts or amounts which are larger in relation to his income or other obligations" is the object of the verb in the sentence; it points to the attitude-specific *content* in that action.

Gives [behavior]	. . . ***in large amounts to others*** in relation to their income and other obligations [content]

Similarly, we could specify the attitudinal content and behaviors in the remaining indicators that Cureton's (1951) quote suggests, as follows:

- Persons who ***give*** [behavior] ***regularly to others*** in relation to their income and other obligations [content].
- Persons who ***give*** [behavior] . . . ***frequently to others*** in relation to their income and other obligations [content].

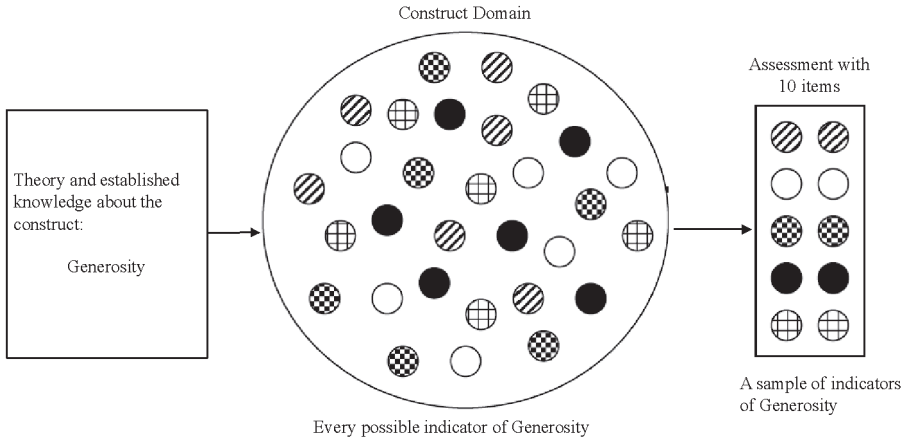

FIGURE 4.2. Domain sampling for content relevance and content representativeness. Adapted with permission from Chatterji (2003). Allyn & Bacon. © Madhabi Chatterji.

The items we design or select must be relevant and representative with regard to the underlying knowledge base for the underlying constructs, from which we would specify the *content* and *behaviors*. Once a domain is adequately specified, it is easier to arrive at a *content-relevant* and *content-representative* sample of items that match the indicators. This is an essential requirement for building content-based validity into the instrument, also referred to as *content-oriented evidence of validity* (AERA et al., 2014). What this brief opening example makes evident is that we often *cannot see* what we are attempting to measure until we have specified the construct domains with action-based indicators that are directly observable.

Adopting a domain sampling approach for instrument design calls for some up-front investments of thinking, time, and labor, but all the assessment operations that follow become easier. Identifying the construct-specific content and behavioral components of indicators, as shown, adds clarity to the indicators. One would typically need multiple indicators—often well more than three—to measure a construct well. The higher the clarity in the content and behaviors of indicators, the better informed the item construction or item selection steps will be during assessment design.

The parallels between the domain sampling theory applied to assessment design and general sampling theory in the social and behavioral sciences may be obvious by now. When applying domain sampling principles, the domain with the specified boundary is analogous to a bounded population of persons or objects, with the indicators replacing the elements within. Unlike probability sampling, however, assessment designers must rely on judgment-based, logical procedures for developing the indicators and then mapping and selecting items to match. In sum, we would apply **purposive sampling** procedures for domain specification and sampling purposes, guided by thoughtful judgments (Creswell & Creswell, 2023; Henry, 1998).

The theoretical ideas on domain sampling are aligned with a good deal of measurement theory on **reliability**. Thinking of the items that are included in an instrument as a random sample of some larger universe allows access to several useful ideas. Mathematical formulations for reliability estimation under classical test theory and generalizability theory draw on domain sampling theory (Crocker & Algina, 2006; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989). More follows in later chapters.

4.2.2 Applying Domain Sampling Ideas to Build Educational Tests

Table 4.1 illustrates the approach that Ralph Tyler (1949) undertook to map and sample the content and behaviors from a domain of **educational objectives** for

a test in a high school course on human biology. As evident, he organized the curricular material in a 2 × 2 table, parsing the instructional objectives into *content* and *behavior* components. The test items were then sampled in a systematic manner from selected cells of the table, marked with an X. Not every cell is marked, suggesting that some parts of the tabulated curriculum were not being tested. These emphases are usually determined by classroom instructors according to their own professional decisions during lesson planning.

Examine the two dimensions of the table in further detail. Like the indicators of generosity, the table specifies (1) a "content" component outlining topics to be taught, learned, and tested in the vertical header (e.g., nutrition, circulation, or respiration) and (2) a "behavioral" component indicating the targeted cognitive behaviors and mental processes in the horizontal header (e.g., show understanding, apply principles, or interpret data). Restated, the objectives from Tyler's (1949) table might look like the following.

Subsequent to instruction, students should be able to:

Apply [behavior] . . .	. . . ***principles in nutrition*** in the unit on foundations of human organization [content]
Interpret [behavior] . . .	. . . ***data on nutrition*** in the unit on foundations of human organization [content]

The parallels in the wording and formatting of indicator statements for the generosity domain and Tyler's achievement domain are not accidental. Formal conventions and widely accepted guidelines now exist for formulating indicator statements along the lines shown. (Readers may also notice the similarities in structure of the Chapter Objectives provided in this book.)

Notice that Tyler formulated educational objectives as the desired ends, or as *outcomes* of the teaching–learning processes, as demonstrable by students at strategic points after instruction. In his view, the objectives defined the curriculum for teachers; hence, they should serve as the common foundation for designing both the instructional plans and classroom assessments (Tyler, 1949, p. 50). His thinking of instructional objectives as **learning outcomes**, or statements that projected the achievement expectations for students, is common practice today, but was an innovation at that early point in the history of American education. An undesirable framing of an educational objective, in Tyler's viewpoint, might specify the activities to be performed by instructors instead. For example, the following would be a poorly framed objective: *Instructors will deliver lessons on nutrition*.

The principle of matching what is tested with what is likely to be taught is the main basis for establishing content-based validity of assessments of educational achievement domains. Tyler's approach set a historical precedence for educational test design efforts (1949,

TABLE 4.1. Domain Sampling with a Table of Specifications

		Behavioral aspects of the objectives						
Content aspect of the objectives	Foundations of human organizations	1. Understanding of important facts and principles	2. Familiarity with dependable sources of information	3. Ability to interpret data	4. Ability to apply principles	5. Ability to study and report results of study	6. Broad and mature interests	7. Social attributes
	1. Nutrition	X	X	X	X	X	X	X
	2. Digestion	X		X	X	X	X	
	3. Circulation	X		X	X	X	X	
	4. Respiration	X		X	X	X	X	
	5. Reproduction	X	X	X	X	X	X	X

Note. Reprinted with permission from Tyler (1949). © University of Chicago Press.

1951; original table in Table 4.1). He illustrated how **domain specifications** laid out in the form of a table could help in the systematic sampling of content and behaviors we desire to measure. That tabulated format is called a **table of specifications** or a *table of test specifications* today. It is an operational standard in all large-scale educational test design projects. For a recent application of similar domain sampling approaches in education guided by the Common Core State Standards reforms, see the mathematics domain with accompanying task and item specifications, released in 2015 by the Smarter Balanced Assessment Consortium (SBAC; see *www.smarterbalanced.org/wp-content/uploads/2015/08/Mathematics-Content-Specifications.pdf*).

4.2.3 Alternatives to Domain Sampling: Empirical Methods of Instrument Design

Another widely used but altogether different approach to instrument development originated in psychology. Referred to as the "empirical method of scale development" (Comrey, 1988, p. 754), the approach departs from domain sampling theory in distinctive ways. It relies on *items gathered from other tools*, and the results of empirical testing and research on those items, for instrument design purposes.

We often see new instruments fashioned with items, tasks, and exercises selected from elsewhere, guided largely by the common sense and interests of the researchers and designers. In the empirical method, the assessment design process starts by assembling or adapting items directly from existing tests and assessment resources based on the judgments of the assessment designers or researchers. Following that, items and instruments are empirically tested with typical examinees or respondents. The statistical properties of the data generated by the newly assembled items and tool inform decisions on which items will be retained or dropped in the final instrument. Many well-known psychological tests and scales were compiled in this manner in personality and social psychology (see studies on the Minnesota Multiphasic Personality Inventory as examples).

A number of statistics and analytic options are available to establish conceptual distinctions among similar and dissimilar sets of items in the empirical approach to instrument design (AERA et al., 2014; see also Briggs & Cheek, 1986; Comrey, 1988). For instance, empirical **correlations** of a newly compiled instrument's scores with external variables could be useful. Here, it is the correlations that designers and users rely on for scale design and evaluation purposes, not any prespecified domain structures. In efforts to identify the latent traits measured by specific groups of items that are compiled, various **factor analysis** techniques are another frequently applied technique.

Although popular, there could be limitations to taking a purely empirical approach to instrument design. A first concern is that designers rely too often on superficial characteristics of items for devising new instruments, such as the wording or language in the original items. Second, when the source instruments are designed for, or previously tested on, very different samples of respondents, the item and score properties often fail to generalize in new contexts. Finally, if we start with originally *untested* items and tools, we could obtain flawed assemblages of items. Other issues with empirical instrument design methods could be:

- Lack of a strong conceptual foundation for the constructs to be measured by items
- Badly designed items to begin with, in the original source instruments
- Improperly balanced item composition in the new instrument
- Inadequate respondent samples or data-collection designs in the original instrument try-outs
- Inherent limitations of the data analytic techniques relied on to assemble items

4.2.4 What Is the Recommended Approach?

This book's approach is to combine the most valuable elements of domain sampling theory with those of empirical instrument development methods. In the absence of a theory-guided domain framework against which empirical results can be referenced or interpreted, errors could enter the instrument design and assembly process—including subjective biases of researchers and assessment designers themselves. The recommended sequence for instrument design is Phases II–III for domain specification and item sampling, and Phase IV for performing the content and empirical validation efforts.

We begin building instruments by specifying the construct domains guided by the principles and procedures of domain sampling theory. The item design, item sampling, and content validation procedures follow accordingly. We continue the iterative design process with appropriate empirical validation studies. Importantly, the design and validation efforts should fit the assessment user contexts. By evaluating the empirical validation results against *a priori* domain frameworks, the aim is to optimize the functioning of instruments and measures in the specified user contexts. For the complete Process Model, see Figure 1.4, Chapter 1.

QUICK RECAP: Domain sampling theory offers useful principles and systematic procedures for specifying domains for hypothetical constructs we might desire to measure. Domain specification should be the first step in operationally defining constructs; it helps create instruments with higher levels of content validity. Regardless of construct type (attitudinal and educational achievement domains were illustrated), we apply similar steps and procedures for specifying the indicators of a domain. Well-specified indicators make the construct "observable" and should be tied to defensible knowledge bases. Empirical methods of instrument design rely on items from preexisting instruments and statistical analyses of item response data for item selection and instrument assembly. Each approach has its merits, utility, and limitations. The Process Model integrates key elements of both, undergirded by the logic of user-centered design.

Reflection Break 4.1: Objective 1

- With an example of an assessment in your area, distinguish between the *theoretical definitions* versus the *operational definitions* of the constructs measured.
- For designing or selecting your own assessment, identify two to three strategies from *domain sampling theory* that you find useful. Justify your answer.
- Compare the *pros* and *cons* of the *domain sampling approach* versus *empirical methods of scale development*, taken individually.
- Outside the two main approaches in Section 4.2, have you seen *other approaches* to assessment design? Describe how the design process was carried out. What are the advantages and disadvantages of that approach?

4.3 Construct Types, Domain Conceptualizations, and Structures

In the social and behavioral sciences, the nature and type of construct often dictate the domain conceptualizations and structures we can reasonably adopt for instrument design purposes. This section provides an overview of the main construct types and domain structures in the literature as a preamble to the sections that follow. See Boxes 4.1–4.3.

4.3.1 Types of Constructs

What are we interested in measuring? Constructs we choose to measure could fall under one or more of the following four categories.

1. **Cognitive constructs:** These constructs were historically classified under the *cognitive domain* by Bloom and colleagues (1956) and deal with one's mental abilities and intellectual capacities in a specified area, both inborn and learned. Indicators in cognitive domains could deal with specific human capacities like concept recall, retrieval and understanding; application of concepts and principles; other types of information processing; problem-solving skills; and proficiency levels in performing mentally demanding procedures. *Examples:* practical intelligence, quantitative skills, decision-making skills, language proficiency, scholastic achievement in science.

2. **Social-behavioral, personality, attitudinal, and affective constructs:** This class of attributes was historically classified under the *affective domain* by Bloom et al. (1956) and are often labeled broadly as "noncognitive" attributes. These are all psychological constructs dealing with one's social-emotional mindsets, proclivities, or dispositions; opinions and values;

personality characteristics; or perceptions, feelings, and social behaviors related to some topic, object, or area. Noncognitive attributes could also be innate or learned tendencies of human beings. *Examples:* attitudes, satisfaction, ethical behaviors, emotional volatility, political beliefs.

3. **Health-related constructs:** Broadly, these constructs deal with various aspects of one's health, well-being, and daily functioning, including one's physical abilities and motor skills; physiological well-being; mental health conditions; or conversely, one's disabilities, diseases, or disorders. Bloom et al. (1956) labeled motor skills learned in school under a *psychomotor domain*. *Examples:* physical strength, endurance, clinical depression, cerebral palsy, sleep apnea.

4. **Sociodemographic constructs:** These constructs deal with societally defined and agreed-upon background and demographic characteristics of individuals, groups, or entities. *Examples:* nationality, socioeconomic status, religion, race/ethnicity, private versus public sectors.

Sometimes, individual constructs we desire to measure may be determined to belong in more than one category, or a new category that lies outside this taxonomy based on the latest literature or societal perspectives. For example, research in the brain sciences suggests that both cognitive and noncognitive characteristics, as defined above, arise from activities and functions in different regions of the human brain, calling into question the biological basis for the first two category distinctions (see, e g., Naliboff & Mayer, 2006). Readers should treat the four classes of constructs as a practical categorization system useful simply for initiating the instrument design process.

4.3.2 Ordered and Hierarchical Construct Domains

Once we pin down the construct to measure, a first question is: What is the best way to conceptualize and organize the domain's indicators, given the underlying theory? Ordered domains are useful for measuring a person's growth or change on constructs; unordered domains are useful for measuring a person's status on the construct at a given time point.

In one measurement tradition, the construct domains are envisaged as *ordered* and *hierarchical*. Researchers following this tradition have attempted to measure the underlying attributes with unidimensional linear scales similar to a meter rule for measuring the length of objects. In applying this tradition, the domain's indicators and matching items are located in a hierarchy, increasing based on their judged levels of difficulty, value, importance, or intensity on an underlying continuum.

For an example of an ordered domain for a cognitive construct, examine Box 4.1. Here, we see a sample of progressively more difficult indicators in a mathematics learning domain. The construct deals with student abilities in recognizing and reasoning with mathematical patterns. This domain, and others like it, were specified for designing developmental assessments for multi-age groups of students. The students were being schooled in nongraded environments where they were allowed to progress at their own pace through the instructional units in different subject areas. As shown with the examples, the mathematics tasks were developed to match each competency level in that hierarchal arrangement of indicators (the curriculum objectives), clarifying the expectations for both teachers and students in nongraded classrooms (Banerji, 1999; Banerji, Anderson, & Kerstyn, 2000; Banerji & Ferron, 1998; Goodlad & Anderson, 1987).

In music education programs, for another example, we might encounter hierarchical domains underlying assessments to measure levels of competence in playing a musical instrument, like the piano. The expected performances are graduated by difficulty from beginner to intermediate to more advanced levels. For other educational applications, think of hierarchically linked, ordered domain specifications underlying mathematics assessments taken by students at the end of elementary school (grade 4), middle school (grade 8), and high school (grade 10). Ordered cognitive domains are useful in either educational or workplace training contexts to map how individuals develop over time in areas that become progressively more difficult or challenging. Most tests based on hierarchical domains are conceptualized to yield a common multilevel, ordered scale structure that increases in difficulty or intensity level.

Note that sometimes, separately constructed assessments designed for different grade level curricula are "vertically linked" mathematically after the fact (see,

BOX 4.1 Ordered Domain Specifications for a Classroom Assessment: Mathematics Indicators and Items at Three Levels of Difficulty

Assessment Use Context

What to Assess?

Learning progressions in skills and concepts on mathematical patterns

Whom to Assess?

Third- to fourth-grade students in public schools in the United States

Why Assess?

Classroom assessment

- *Measure-based inferences?* Learning gaps and learning progress along the mapped curriculum
- *Specific uses?* Classroom-level formative decisions
- *Primary users?* Teachers, instructional aides, learners, and families

Domain with Indicators and Tasks at Three Levels

General Indicator

- Given problems arranged by difficulty, students will identify and explain mathematical patterns represented by symbols, basic number concepts, and arithmetic operations. (Taxonomic level: *Higher-order thinking*)

Specific Indicators

- **Level 1 Indicator:** *Identify, continue, and explain simple repeating patterns involving shapes, numerals, or letters.*

  Matching-Item Example:

  What is the pattern that you see below? Fill in the last *three* blanks to continue the pattern. Say why your answer is correct.

  ABC, ABC, ABC, _____, _____, _____?

  Explanation:

- **Level 2 Indicator:** *Identify, continue, and explain mathematically increasing patterns using basic number sense and arithmetic concepts.*

  Matching-Item Example:

  What is the pattern in the series of numbers below? Predict the last *three* numbers in the series. Explain the mathematical rule. Show all your work.

  1, 10, 100, 1000, _____, _____, _____?

  Math Rule:

  Explanation:

- **Level 3:** *Identify, continue, and explain mathematically increasing patterns, using intermediate-level arithmetic concepts––for example, multiplication, division, fractions, prime numbers, and squares.*

  Matching-Item Example:

  See the pattern in the series of numbers below. Predict the last *two* numbers in the series. Explain the mathematical rule that helped you find the answers. Show all your work.

  2, 4, 16, _____, _____?

  Math Rule:

  Explanation:

Answers.

1. Rule: repeating series of same letters. ABC; ABC; ABC.
2. Rule = x*10. 10,000; 100,000; 1000,000.
3. Rule = x^2. 256; 65,536.

e.g., Strachan et al., 2020). Typically, these instruments are not based on hierarchically conceptualized domains. Standardized test developers often perform this type of test-linking procedure. But the tests are usually not built from ordered test design specifications, nor administered to large, multigrade samples of test-takers. Hence, the inferences possible from the vertically linked construct measures are limited.

4.3.3 Unordered Domains: Homogeneous and Heterogeneous Structures

In an alternative tradition, the domain is conceptualized as *unordered*, where indicators are not hierarchically arranged, but highly similar and interchangeable with respect to the content and behavior to be measured. Unordered domains often have a **simple (nonstratified) domain structure.** Think of a domain tapping into only addition skills in mathematics with one-digit numbers: 5 + 4 = ____, 3 + 3 = ____, 7 + 8 = ____, and so on. Such a domain would have a simple structure because the items or tasks would focus on one tightly defined indicator, with items linked to that narrowly specified range of content and behavior: *Add* [behavior] *single-digit numbers* [content] *correctly*. Here, we would expect items to be replicable with respect to difficulty level. By design, a simple, unordered construct domain is homogeneous in nature—a property that can be tested empirically with a statistical index, an *item homogeneity index* (more on that in Chapters 10 and 11).

By contrast, think of a construct domain tapping into a wider range of arithmetic skills requiring separate proficiency levels to be measured in addition, subtraction, multiplication, and division. Given the four qualitatively different indicators, this domain would have a **stratified domain structure**. Contrasted to the addition only domain, there would be four heterogeneous strata here. Under a general domain of arithmetic skills, we would have four subdomains corresponding with each stratum. Within each stratum, we could still have unordered **subdomain** structures. Once the indicator-referenced items are compiled, each subdomain could potentially yield a **subscale** or subtest with a separate score. With sufficient overlap in content and behaviors across strata, we could tie together the measures under a common domain framework representing the overall construct.

Graphically, for generosity, Figure 4.2 suggests a similarly stratified, and unordered domain structure. There are five hypothetical strata. The content of indicators is shown graphically (with differently filled-in circles). Implicitly, each subdomain would be defined by observable indicators of generosity that were substantively different. For example, "acts of generosity" as demonstrated in five separate arenas of life could be specified, such as generous acts in the family, the workplace, the neighborhood and community, religious organizations, and political/governmental organizations. As with all unordered domains, the indicators (and items generated) would be treated as equivalent in terms of difficulty, value, or weight, with no implicit hierarchy in the strata.

Take a look at two real-world applications now. Respectively, each served as the foundation for designing cognitive and noncognitive assessments.

The first domain framework is shown in Box 4.2 (Chatterji, Graham, & Wyer, 2009). This domain was specified for designing multiple-choice tests to measure competency levels of resident physicians while they were under training at university hospitals. Similar to Tyler's structure, this domain was also conceived as two-dimensional. Physicians under training were expected to demonstrate their skills in practicing evidence-based medicine (EBM). The four "content" strata are therapy, diagnosis, prognosis, and harm (refers to evaluating harmful side effects of treatments). The content strata are crossed against "behavior" strata corresponding to four EBM skills: ask (asking clinical research questions to gather evidence), acquire (acquiring the evidence), appraise (appraising the evidence), and apply (applying the evidence to patient cases). Items were developed to populate the cells of this table of specifications with weighting assigned to items in the cells, as shown. This is an example of a heterogeneous, stratified domain that is unordered.

In Box 4.3, we see a domain for a noncognitive construct to measure the personality characteristics of effective teachers. This domain is based on a literature review conducted by the authors (Madni, Baker, Chow, Dellacruz, & Griffin, 2015). It is a similarly stratified and heterogeneous domain with four strata: Conscientiousness, Extraversion, Emotional Stability, and Openness to Experience. The developers recognized separate strata that define the overall personality construct, identifying added dimensions to be described

BOX 4.2 Unordered and Stratified Domain Specifications for a Cognitive Competency Assessment

Assessment Use Context

What to Assess?

Competencies in practicing evidence-based medicine

Whom to Assess?

Resident physicians undergoing training at university hospitals

Why Assess?

Program evaluation and accreditation

- *Measure-based inferences?* Competency levels of trainees and overall program performance
- *Specific uses?* Evaluating average trainee competency levels for program accreditation
- *Primary users?* Faculty, program directors, trainees, and mentors

Domain Specifications

General Indicator

Resident physicians will formulate clinical questions to locate, appraise, and apply the best-available research evidence for making clinical decisions relevant to patient needs.

Specific Indicators

- "Ask" skill: Formulate questions to guide information searches on therapies and clinical courses of action for patients.
- "Acquire" skill: Select evidence-based courses of clinical action for patients.
- "Appraise" skill: Critically appraise the quality and applicability of the evidence for individual patient cases.
- "Apply" skill: Integrate the evidence into clinical decisions while taking into account individual patient values and circumstances.

Table of Test Specifications

Behaviors:	*Content:* **THERAPY**	*Content:* **PROGNOSIS**	*Content:* **DIAGNOSIS**	*Content:* **HARM**
"Ask" Skills	10%	5%	5%	5%
"Acquire" Skills	10%	5%	5%	5%
"Appraise" Skills	10%	5%	5%	5%
"Apply" Skills	10%	5%	5%	5%

Note. All items carry equal point weights. Percentages denote item weighting and distribution plan for designing a multiple-choice test matched to the domain.

Source: Adapted with permission from Chatterji, Graham, & Wyer (2009). ©Accreditation Council for Graduate Medical Education.

later. Although not in the original article, each stratum could be represented by a general indicator using the conventions we applied for *generosity*; specific indicators are also added here for illustration purposes. Once specified, all domains should be validated independently to add rigor to the domain specification procedures.

QUICK RECAP: How we conceptualize the domain structure depends on the type of construct, the supporting literature, traditions of measurement that influence the design process, and the assessment purposes. The construct type could be cognitive, noncognitive, health-related, or sociodemographic. Domains could be conceptualized as ordered or unordered. For both, we could opt for either homogeneous or heterogeneous structures. These design decisions should depend on the purposes for assessment, as given by the primary users and assessment designers, and other contextual factors.

Reflection Break 4.2: Objective 2

- Define each broad type of construct introduced in this section: cognitive, noncognitive, health-related, and sociodemographic constructs. Give *new examples* in the four categories.
- How will you deal with any construct ambiguities you find during the instrument design process?
- Distinguish between the following: ordered domains; unordered domains; simple homogeneous

BOX 4.3 Stratified and Unordered Domain Specifications: Assessing Personality Traits of Effective Teachers

Assessment Context

What to Assess?

Personality characteristics of effective teachers

Whom to Assess?

Teachers in public education systems in different localities

Why Assess?

Formative decisions in workplace contexts

- *Measure-based inferences?* Strength of different personality characteristics
- *Specific uses?* Professional development, mentoring, and goal setting
- *Primary users?* Teachers, supervisors, mentors

Domain Specifications

General Indicator. In professional contexts, teachers exhibit the personality characteristics of:

- *Conscientiousness*
  Specific Indicators *(Examples):*
  - Teachers complete all their school-related obligations on time.
  - Teachers prepare lessons carefully to meet diverse students' needs.
- *Extraversion*
  Specific Indicators *(Examples):*
  - Teachers behave in a friendly manner toward colleagues and coworkers at school.
  - Teachers reach out voluntarily to support colleagues and coworkers at school.
- *Emotional stability*
  Specific Indicators *(Examples):*
  - Teachers show calmness in their classroom demeanor.
  - Teachers speak in a well-modulated tone at school.
- *Openness to experience*
  Specific Indicators *(Examples):*
  - Teachers show openness to new methods of teaching.
  - Teachers express open mindsets when teaching diverse students.

domains; stratified, heterogeneous domains. Give examples of each from your experience.

- Describe how you plan to conceptualize the construct domains(s) for your own instrument.

4.4 Domain Specification as a Part of the Process Model: Steps, Techniques, Guidelines, and Conventions

See Table 4.2 next in conjunction with Figure 4.1 presented earlier in this chapter. There are five steps to the domain specification process. This section offers some established guidelines and techniques for specifying construct domains, situated in the Process Model. As we proceed, we refer back to Boxes 4.1–4.3, with added illustrations and discussion.

4.4.1 Five Iterative Steps for Specifying Domains

Specifying Phase I of the Process Model is fundamental to user-centered assessment design. By aligning the indicators with the population characteristics and relevant socioecological factors mindfully, we could specify construct domains in a more contextually and culturally responsive manner. Similarly, keeping the user path, intended inferences, and uses of construct measures in mind alerts designers to the assessment stakes, as well as the interests of the primary users. As a general rule, the higher the stakes tied to assessment-based actions, the more rigorous the design procedures should be—starting with the domain specification and initial content validation processes. Good practice demands that we subject the initially drafted domains to at least one critical review and revision, even for informal assessments. The overall process is iterative rather than sequential. Refer to Tables 4.1 and 4.2 next showing the steps that would follow.

TABLE 4.2. Five Iterative Steps to Specify Construct Domains

Step	Domain specification procedures
Step 1	Identify constructs or attributes to measure with reference to Phase I of the Process Model (*What to assess?*). • Label the construct(s) and set the theoretical limits. • Take into account the population specifications (*Whom to assess?*). • Take into account the assessment purpose specifications (*Why assess?*).
Step 2	Locate appropriate knowledge bases, data sources, or resources for deriving indicators. • Literature reviews, scientific research, curricula, and/or documentary sources. • Expert knowledge and specialist perspectives. • Direct observations, critical incident techniques, or case studies.
Step 3	Write and organize indicators using established guidelines and conventions. • Use standard guidelines to specify observable aspects of indicators. • Organize indicator statements in useful ways.
Step 4	Use appropriate taxonomies for classifying and clarifying the behavioral dimensions of indicators. • Sort and separate indicators in cognitive, noncognitive, health-related, or sociodemographic constructs. • Identify homogeneous clusters of indicators by "behavior" dimensions to be measured. • Develop or select item or task examples to match selected indicators.
Step 5	Validate the specified domains using internal and/or external reviews. Iterate and revise, as needed.

4.4.2 Locating Appropriate Resources for Specifying Domains

Assessment designers should try to locate credible sources from which to derive the construct indicators. For most informal assessment design efforts, a quick Google search might suffice. In more formal applications, added research and alternative data sources will likely be necessary. This section discusses three commonly used data sources and associated research methods for domain specification:

1. **Using existing theory, research, literature reviews, and documentary sources:** Common resources for domain specification are scientific, professional, and academic research articles in relevant fields; documents published by national or international professional associations; mission and policy statements of

organizations, institutions, or programs; and current textbooks or curriculum resources in a given field and education level.

How are indicators extracted? A useful way to identify indicators is by conducting a thematic content analysis of the articles gathered. For example, Madni and colleagues (2015; Box 4.3) desired to measure noncognitive traits of effective teachers. Through a review of the research literature dealing with effective teacher characteristics, they identified four general areas (domains): personality characteristics, motivational attributes, intrapersonal skills, and interpersonal skills. Guided by the five-factor theory of personality (Olver & Moorardian, 2003), they further broke down personality characteristics of teachers into five dimensions: conscientiousness, emotional stability, extraversion, agreeableness, and openness to experience. Box 4.3 is developed from that work, with specific indicator statements added to illustrate how the domain could be specified in further detail.

Similarly, the cognitive competency domain in Box 4.2 was operationalized as shown based on an extensive literature review performed by the authors (Chatterji, Graham, & Wyer, 2009). The research team drew on a variety of documents for the literature review: established standards and guidelines in the medical profession, curriculum goals and criteria set by the accreditation agency for graduate medical programs, and medical research literature and users' guides on EBM (Guyatt, Rennie, Meade, & Cook, 2002). In addition, they sought perspectives of medical and EBM experts to fortify the indicators and subindicators of the construct domain. The table of specifications shown was compiled after the indicators of the domain and subdomains were validated by experts. To build the competency tests, the team designed a large pool of multiple-choice items to match the indicators in the cells. Items were sampled proportionally from each cell based on percentage weights assigned by medical experts on the team.

2. **Using expert opinions and perspectives:** In situations where the construct to be measured is very new, or the attributes are little known and yet to be formalized in the existing literature, we could utilize the opinions of experts, scientists, or experienced practitioners for specifying the domain's indicators.

How are indicators extracted? There are several qualitative research methods and facilitated group processes for eliciting knowledge from individual experts or expert panels. During the data-collection process, researchers could either take notes or record the chief points conveyed by experts on domain-relevant topics. The data could then be categorized qualitatively by theme and subtheme using content analysis methods. Multivariate statistical methods, like **cluster analysis**, may also be used subsequently to analyze and further interpret the qualitatively coded data we secure. Selection of the "experts" is key for domain specification purposes with these methods; the individuals solicited should be highly knowledgeable and/or experienced professionals, with the necessary insights into the constructs and populations of concern. Widely used in this category are the Delphi method, focus group interviewing, nominal group technique, and concept mapping. Each is described briefly below.

The **Delphi method**, originally developed as a systematic, interactive forecasting method, 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 (Dalkey & Helmer, 1963). Here, the experts would answer questionnaires in two or more rounds on the hypothetical construct (or topic). After each round, a facilitator would provide an anonymous summary of the experts' ideas along with rationales for those judgments. In the second round, participants would be encouraged to revise their earlier answers in light of the facilitator's feedback on the overall group's perspectives, until the group converges toward consensus.

Focus group interviewing is a form of qualitative research in which a carefully selected group of people is asked for their perceptions, opinions, beliefs, or attitudes about given topics, such as the manifestation of COVID-19 as a disease in different social groups. The interviews are conducted through guided discussions. For market research or political analysis purposes, 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 (Greenbaum, 2000). For domain specification applications, the interviews could attempt to extract opinions and points of view on the indicators of a construct.

The **nominal group technique** is another group

process involving problem identification, solution generation, and decision making in an area. This approach could be similarly useful for arriving at consensus-based indicators of a new construct (Delbecq & Van de Ven, 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.

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 (Goldman & Kane, 2014; Rosas & Kane, 2012). Expert participants start by brainstorming a series of descriptive, representative statements on a construct-related question, such as "What are the signs of severe COVID-19 in the elderly?" 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 subdomain framework for the construct.

An applied example with a combination of the above methods can be found in the work of Graham and colleagues (2009). This research team 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 practices (SBP), given by the Accreditation Council for Graduate Medical Education. Their goal was to operationally define resident physicians' competence in SBP in terms of observable roles, actions, and behaviors.

The researchers collected data using structured focus group interviews of a total of 88 health care professionals working in various roles in large hospitals and health care systems in New York City—doctors, nurses, technicians, patient care support staff, and administrative staff. Their methodology involved gathering data from group meetings, coding themes obtained from the consensus-seeking procedures, and then conceptually matching and organizing the indicators to define the SBP domain. The domain served as the basis for designing observational assessments to rate physicians' SBP competency levels during their residency training.

3. **Direct observations, critical incident techniques, and case study research:** Other ways to define domains for constructs that are relatively undocumented or unknown involve observational and case study methods. Three commonly employed practices are discussed next.

How are indicators extracted? 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. A key informant is an individual with firsthand 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 key informants) to define so-called critical behaviors representing outstanding performance on the job in a given workplace setting, contrasted with completely ineffective performance of workers. The identified behaviors served as indicators of the construct.

Case study methods involve detailed reviews and analysis of case records for identifying and cataloging behavioral indicators of constructs. Case study methods are particularly useful in defining health-related constructs for still undocumented diseases or disabilities. Constructs could be defined comprehensively by studying symptoms and behaviors of patient cases via direct observations supplemented with other forms of case data. To enhance credibility, the case data must be collected by appropriately trained professionals and corroborated over multiple cases (Stake, 2015).

See Box 4.4 for an illustration of how case studies could be used to specify domains. The illustration draws on a historical narrative on cures for cancer, describing how early understandings of leukemia evolved from recordings and direct observations of patient cases by committed doctors and cancer researchers (Mukherjee, 2010). Careful observations, documentation, cataloging, and classification of the symptoms from multiple cases led to collective learning about the disease. The most useful observable indicator of the disease was determined to be the count of white blood cells in the patients' blood. This discovery provided the gateway toward developing protocols for measurement and diagnosis of the condition in patients.

BOX 4.4 Using Case Studies to Derive Observable Construct Indicators

Assessment Use Context

What to Assess? "Leukemia," an unknown disease at the time

Whom to Assess? Adult patients

Why Assess? Clinical diagnosis and treatment

- Measure-based inferences? Identification of disease symptoms and characteristics
- Specific uses? Diagnosis
- Primary users? Oncologists, cancer researchers, patients, and families

Patient Case Studies for Indicator and Domain Specification

Case 1: An early case study dated 1845 presented observation records of a 28-year-old slate-layer. The observations were made by John Bennett, as described in Mukherjee (2010, pp. 12–13).

Early stages of disease showed the following external signs in the patient:

- "a mysterious swelling in his spleen . . ."
- "great listlessness on exertion," which continued over a period of 20 months
- "tumor in his abdomen which gradually increased in size," becoming stationary 4 months afterward

In the next few weeks:

- rapid disease progression
- patient had "fevers"
- "flashes of bleeding"
- "sudden fits of abdominal pain, gradual at first, then on a tighter, faster arc . . ."
- then "more swollen tumors sprouting in his armpits, his groin, and his neck . . . ," leading eventually to patient's death

An autopsy revealed:

- patient's "blood was chock full [of] white blood cells . . . [which is] a principal constituent of pus . . ." [but] "Bennett could not find a source for the pus" at that time.

Case 2: Observations of another concurrent case were conducted by Rudolf Virchow. This case concerned a cook in her mid-50s, as presented by Mukherjee (2010, pp. 13–16).

- The patient showed "striking similarities" with the first in that she had a "massively enlarged spleen."
- The patient also presented with "white blood cells . . . explosively overgrown [in] her blood."
- Again, there was a mysterious "absence of any wound" as the source of the excessive white blood cells or "pus."

Inferences

Based on the similarities in indicators in both cases:

- Virchow recognized "the blood itself was abnormal. . . . The blood cells had overgrown in a distorted, uncontrolled fashion with "millions of white blood cells . . . seen under his microscope" after the patient's death.
- Virchow named the disease "leukemia."

Implications for Assessment Design

Based on multiple case records, "leukemia could be counted . . . by drawing a sample of blood or bone marrow and looking at it under a microscope" (Mukherjee, 2010, p. 19). The counts of white blood cells of normal versus diseased individuals became a key diagnostic indicator of the condition. Assessment techniques could be built using this and other indicators derived from added case studies.

QUICK RECAP: Table 4.2 recommended five iterative steps for specifying domains. Once the construct types are identified, specifying indicators immediately follows. Locating the right resources to identify domain-relevant indicators typically requires reviews of existing research or documentary resources. For unknown or novel constructs, researchers could seek expert opinions, or conduct small-scale exploratory studies using key informants, focus group interviews, direct observations, or case study methods to generate construct indicators. These steps fall under Phases I–II of the Process Model, under the *What?* and *How?* portals.

Reflection Break 4.3: Objectives 1–3

- What are the *benefits* of specifying Phase I prior to specifying the domain?
- Compare *three different methods* for locating resources to specify construct domains. List the advantages and disadvantages of each for your own instrument design project.
- Suppose you wanted to measure *teacher attitudes toward technology* (or another construct).
  - Identify the *construct type* based on the four-category taxonomy given.
  - How would you specify the *domain* and *subdomains*? What resources would you use to derive indicators and items? What steps would you take to locate resources? What would be your domain structure?
  - How would you establish the credibility of your domain?

4.4.3 Writing and Organizing Indicator Statements

Because most constructs are big concepts comprising multiple and sometimes layered components, we often start the domain specification process with a collection of brainstormed ideas and broad themes extracted from the data sources we select. Restating the initially brainstormed ideas as formal indicator statements helps bring clarity and preciseness to the themes. Organizing the indicators in coherent clusters adds theoretical integrity and meaningfulness to the assessments we design. Three guidelines now follow building on the earlier indicator illustrations on generosity and human biology.

1. Use action-oriented verbs to write indicator statements. Action words convey the directly observable behaviors, acts, or responses that are easily measurable through the tasks or items. As mentioned, by using the verb "give," Cureton (1951) offered action-oriented indicators of generosity. Similarly, for indicators in cognitive domains, we would prefer to use the verbs "write," "state," or "apply" instead of verbs like "know," "understand," or "discuss" that are relatively vague. In the same vein, to measure hand–eye coordination in children with cerebral palsy, a health-related construct, a researcher in biobehavioral studies framed an indicator with the action verb shown below (Sarafian, 2020). Cerebral palsy stems from a neurological condition that begins in the prenatal stages, leading to disabilities in the hand–eye coordination of children.

> *The child picks up* [action-oriented behavior]
> *objects of different sizes with one hand* [content].

2. When appropriate, specify the content, behavior, condition, and/or performance criterion in indicators. Minimally, well-written indicators specify the content and behavior components, as demonstrated earlier. Depending on the assessment purposes and construct type, we might specify *four* distinguishable parts in an indicator: behavior, content, condition, and a criterion performance. The condition in an indicator helps place the behavior in an exact setting where it will be performed, observed, recorded, and measured. The criterion indicates the performance expectation or a mastery level that should be demonstrated by the respondent or examinee.

Continuing with the earlier example, children with cerebral palsy are often placed with occupational therapists who help build their hand–eye coordination skills. To design assessments to measure the progress of children receiving therapy, the domain's indicators might specify the condition and a criterion level of performance, as shown below:

When prompted by the examiner during individual assessments [condition], the *child picks up* [action-oriented behavior] *objects of different sizes with one hand* [content].

When prompted by the examiner [condition], the *child picks up* [action-oriented behavior] *objects of different sizes with one hand* [content] *on three separate occasions* [criterion].

For another example, we could write an indicator in the mathematics patterns proficiency domain with all four parts, as follows:

Students will be able to calculate [behavior] *the sum of a sequence of whole numbers* [content] *without a hand calculator* [condition] *showing 80% accuracy* [criterion performance].

An indicator for measuring the teacher personality attribute of conscientiousness that includes a condition could look like this:

In classroom teaching contexts [condition], *teachers exhibit* [behavior] *conscientious behaviors to ensure student learning* [content].

3. Organize the indicators from general to specific. To bring a domain into sharper focus, indicators are typically organized from general to more and more specific statements, until the domain is as clear as possible for item design or selection purposes. Indicator clustering should be coherent and meaningful, guided by the literature or data sources used. In a domain's schematic representation using tree diagrams, the general indicators are more broadly stated and placed at the upper levels of the domain. Further breakdown of the general indicators to specific indicators that are increasingly more concrete clarifies the definition of the construct. For organizing and presenting indicator statements from the general to more specific levels, Figure 4.3 shows a **tree diagram** arrangement.

Examine the concept of *conscientiousness* with and without the specific indicators provided to see why second- or third-level indicators are often necessary to remove any inherent ambiguities in constructs. Box 4.3 provides only two as illustrations for each general indicator, but the need for more may be obvious. To derive the specific indicators, guiding questions might be helpful, such as "What would a conscientious teacher tend to say or do in professional contexts? How would they act, based on the literature or practitioner reports?" The embedded specific indicators add detail to the content and behaviors to be measured.

The tree diagram in Figure 4.3 depicts a domain that aims to measure historical thinking skills in secondary school students. Graphically, the domain shows one main branch with five subbranches. The general indicators illustrated are rather broad and specify the content dimensions only (not desirable!). But, as each is broken down further with a second-level indicator delineating the behavior (such as "analyze") and the content (such as "multiple causation") of focus, the construct is clarified. Each such indicator could be further broken down by adding branches to the tree, until the operational features of tasks and items are evident for assessment design purposes. The number and levels of branches and subbranches necessary in a tree diagram could vary for different construct domains.

For the mathematics patterns proficiency construct (Box 4.2), a general indicator in the main stem of the tree diagram could be broken down at three levels, with each level further detailing the content and task-specific behavior to be assessed, as follows. Mapping out the domain coherently helps retain the logical, internal consistency of the theoretical foundation on which the assessments and construct measures are built.

General indicator (main branch of a tree diagram):

Solves problems dealing with mathematical patterns and sequences.

Specific indicator (second-level branch):

Selects appropriate arithmetic operations to continue a given mathematical sequence.

Specific indicator (third-level branch):

Calculates the sum of the terms of a given mathematical sequence containing whole numbers, decimals, or fractions.

QUICK RECAP: Domain specification is necessary to operationalize the construct with the requisite levels of clarity, focus, and theoretical coherence

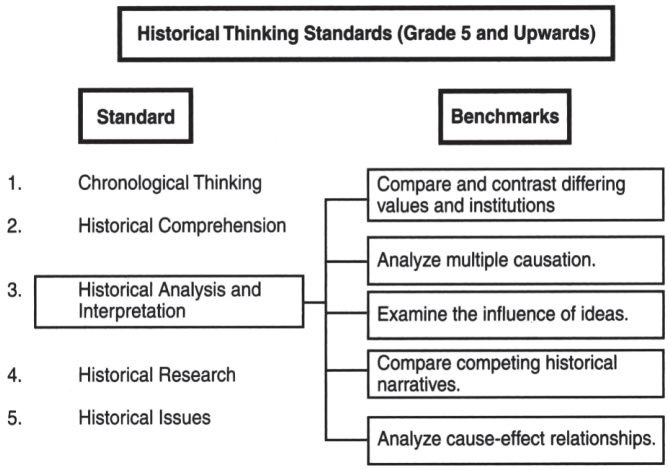

FIGURE 4.3. A two-level tree diagram: A specified achievement domain in history. Adapted for illustration purposes from Chatterji (2003).

to guide item design or selection. By convention, indicators could delineate the content, behavior, condition, and/or the expected criterion performances to be measured. Minimally, indicators should specify the content and behavior dimensions; decisions on the added components should be guided by the larger assessment context, purposes, and construct type. Tree diagrams are useful for organizing indicator statements in a domain from general to more specific levels. Other useful mechanisms for organizing domains are tools like concept-mapping diagrams.

Reflection Break 4.4: Objectives 3 and 4

- For this exercise, think of measuring abilities and attitudes on *driving automobiles* in adult U.S. populations. When specifying the domains, identify the advantages to (1) specifying the construct types, (2) locating appropriate and defensible sources and documents, (3) stating indicators clearly, and (4) organizing indicators from general to specific.
- Write four indicators to assess the construct *automobile driving proficiency.* Specify the content, behavior, conditions, and criterion levels of performance you desire to measure, as applicable.
- Write four indicators to assess the construct *attitude toward driving.* Specify the content, behavior, and conditions in the indicators that you desire to measure, as applicable.

4.4.4 Using Taxonomies to Classify Indicators

Benjamin Bloom and colleagues (1956) began a helpful tradition in assessment design, where the indicators are classified using a selected taxonomy of behaviors that then guides item design or selection. Within the domains and subdomains that define a construct, recall that domain sampling theory assumes that items are similar and homogeneous. A taxonomic categorization of the "behavior" component of indicators enables the production of closely related items tied to the desired indicators, helping achieve homogeneous properties of the items. Classifying indicators using behavioral taxonomies has added benefits in facilitating homogeneous selections of appropriate items, assessment modalities and item formats for differently classified indicators as well. Furthermore, when designing scoring procedures for assessments, the taxonomic level of indicators helps designers with decisions on how much

weight (importance in terms of points) to allocate to different groups of items that measure, say, simple recall versus application or higher-order thinking skills.

The literature offers several alternative taxonomies for these purposes, old and new. Each is applicable to a different construct type. This section offers a series of *functional taxonomies* that build on that literature base as optional tools for aiding the assessment design or selection processes for different construct types (Bloom et al., 1956; Chatterji, 2003; Eagly & Chaikin, 1998; Gagné, 1965; Katz, 1960; Krathwohl, Bloom, & Masia, 1964; Marzano, Pickering, & McTighe, 1994; National Research Council [NRC], 2012). Readers are encouraged to review the classic *Taxonomy of Educational Objectives* (Bloom et al., 1956) and more recent taxonomies (NRC, 2012) as a supplement to this section.

4.4.4.1 A Functional Taxonomy for Measuring Cognitive Constructs

The cognitive taxonomy, shown below, recognizes four separate types of cognitive capacities that we could aim to measure (after Bloom et al., 1956; Chatterji, 2003; Krathwohl et al., 1964; NRC, 2012). Each is expected to impose different mental demands on the examinee or respondent. Indicators measuring concept recall and understanding are at the lowest level in terms of expected cognitive demands; application is at the next higher level, with complex procedural skills and higher-order thinking skills at the most demanding levels of cognitive processing. The verbs suggest the cognitive behaviors that the indicators and items would be designed to match. For each level, either structured-response or more open-ended problem-solving tasks could be designed, as appropriate. (See the table at the bottom of this page.)

The categories are cumulative. Application-level tasks will typically also call for concept knowledge and understanding. Likewise, higher-order thinking tasks will typically require both concept recall and understanding, as well as application. A demonstration follows on how to apply the taxonomy for indicators of the mathematical patterns proficiency domain shown earlier dealing with mathematical sequences:

Define a "mathematical sequence." [concept recall and understanding]

Calculate the sum of the terms of a given mathematical sequence. [application]

Solve real-world problems involving mathematical sequences. [higher-order thinking skills]

Construct type	Levels or types	Definition	Examples of indicator phrases
1.0 Cognitive constructs	1.1 Concept knowledge and understanding	Requires examinee to recall, retrieve, and/or show comprehension of basic concepts, facts, and principles.	State a definition, law, or principle. Describe in your own words. Paraphrase the meaning of. Give examples of. Distinguish between.
	1.2 Application	Requires examinee to apply concepts, rules, principles, tools, or formulas.	Calculate. Solve a problem. Use a tool.
	1.3 Complex procedural skills	Requires the execution of a complex task, involving multistep mental skills and integrative thinking processes, usually following accepted conventions or standards in a field.	Employ a writing process to create a story. Write a laboratory record. Conduct a research study. Develop a blueprint.
	1.4 Higher-order thinking skills	Requires higher intellectual skills involving analysis, synthesis, and/or evaluative judgments.	Analyze, explain, create, compose, compare, contrast, critique, evaluate, defend, or justify.

Apply a multistep procedure to check answers to problems involving mathematical sequences. [complex procedural skills]

4.4.4.2 A Functional Taxonomy for Measuring Noncognitive Constructs

Recall that noncognitive constructs could include attitudinal, personality, and social-behavioral constructs, as well as various interpersonal and intrapersonal domains, including metacognition (NRC, 2012). In the psychological literature, there is a tripartite taxonomy to help organize, classify, and label indicators on attitudes, viewed to have "cognitive," "affective," and "behavioral" (CAB) components (Eagly & Chaiken, 1998; Hovland & Rosenberg, 1960). Note that for noncognitive constructs, the term *cognitive* refers to one's beliefs about something. While this usage is consistent with the literature on attitude measurement (Eagly & Chaikin, 1998), *cognitive* in the context of attitude measurement should not be confused with mental abilities and skills in the cognitive taxonomy given earlier. Rather, it should be interpreted as what a person holds to be true about something. This tripartite taxonomy is broadened below with a metacognitive component with the mnemonic CAB-M. In the demonstrations, the attitude-specific content focus is on childbearing and abortion topics. (See the table at the bottom of this page.)

For more concrete demonstrations, suppose you wish to apply the noncognitive taxonomy above to measure attitude toward childbearing and abortion in adults. The indicators next listed are matched to survey items assuming a Likert response scale: Strongly agree through Strongly disagree. Each item example is designed to measure a different CAB-M dimension.

- **Cognitive (C)** component of attitude toward childbearing and abortion:
  - C-Indicator. *Endorses a woman's right to choose the circumstances for bearing a child.*
  - Item. *A woman should have the right to choose the time when she has a child.*
- **Affective (A)** component of attitude toward childbearing and abortion:
  - A-Indicator. *Indicates feelings about a woman's right to choose on matters of childbearing.*
  - Item. *I am frustrated that my state's laws prevent women from making their own choices on childbearing.*

Construct type	Levels or types	Definition	Indicator or item examples
Noncognitive constructs (affective, attitudinal, personality, or social-behavioral domains)	2.1 **C:** "Cognitive" component of dispositions	What a person *holds to be true* about something; their opinions, beliefs, values, or perceptions about it, including experiences, objects, events, places, or persons	Endorses or communicates beliefs (regarding childbearing). Endorses or communicates perceptions of (childbearing).
	2.2 **A:** "Affective" component of dispositions	What a person *feels emotionally* about social issues, present or past experiences, events, places, persons, or objects	Endorses or communicates emotions or feelings about (childbearing).
	2.3 **B:** "Behavioral" component of dispositions	What a person *would do*, or a person's attitude-governed action and practices would be, in relation to given experiences, events, places, persons, or objects	Communicates doing something or engaging in practices, actions, or behaviors (related to childbearing).
	2.4 **M:** "Metacognitive" component of dispositions	A person's stance on self-reflection and self-evaluation of their own behaviors, actions, mindsets, aimed to self-correct a position.	Communicates self-introspective and self-correcting stances on issues (like abortion or childbearing).

Construct type	Levels or types	Definition	Indicator or item examples
3.0 Health-related constructs	3.1 Physiological indicators of a health condition	Physiological signs or symptoms of a state of well-being or an illness or disorder	Items on heart rate, blood pressure levels, body temperature, etc.
	3.2 Behavioral indicators of a health condition	Behavioral signs or symptoms of a state of well-being vs. an illness or disorder	Items indicating what a person will say or do when well vs. ill (e.g., stutters, forgets details, stumbles, sleeps too much, reports low or high energy levels)
	3.3 Physical appearance indicators of a health condition	Appearance-related indicators of a state of well-being vs. an illness or disorder	What a person looks like when well versus ill (e.g., redness of eyes, paleness of skin, gaunt)

- **Behavioral (B)** component of attitude toward childbearing and abortion:
  - B-Indicator. *Reports practices or acts related to personal stance on childbearing.*
  - Item. *I have participated in anti-abortion protests.*
- **Metacognitive (C)** component of attitude toward childbearing and abortion:
  - M-Indicator. *Indicates engaging in self-evaluations on issues of childbearing.*
  - Item. *I rethink my positions on childbearing based on information I read.*

4.4.4.3 A Functional Taxonomy to Measure Health-Related Constructs

Along similar lines, consider the taxonomy and specific examples next for designing assessments for health-related constructs. (See the table at the top of this page.)

Consider next the health condition of sleep apnea, a sleeping and breathing disorder found in children and adults. The item examples, designed for a caregiver interview for a child, demonstrate application of the behavioral dimension from the above taxonomy.

- Indicator. *Parent or caregiver reports on their child's sleeping habits at night.*
- Item. *How often is your child restless while sleeping at night?*
- Response Options.
  - *Every night of a week—Most nights in a week—Some nights in a week— Rarely—Never*
- Indicator. *Parent or caregiver reports on their child's breathing behavior while sleeping.*
- Item. *How often does your child snore while sleeping at night?*
- Response Options.
  - *Every night of a week—Most nights in a week—Some nights in a week—Rarely—Never*

4.4.4.4 A Functional Taxonomy to Measure Sociodemographic Constructs

Finally, the next taxonomy focuses on attributes and characteristics that are implicitly agreed-upon, societal constructions, typically employed by governments, organizations, and institutions to group individuals broadly. The definitions may vary depending on the cultural, regional, or national context; hence, indicator or item writing should reflect the contextually relevant social norms and construct definitions in given localities where the instrument is applied. Item examples are suggested. (See the table at the bottom of page 117.)

4.5 Content-Validating Specified Domains

Alongside Table 4.2, Box 4.5 provides 10 questions to guide evaluations of initially specified domains for constructs we desire to measure. Content validation is a last step of the domain specification process, accomplished using critical self-reviews, peer reviews, or

more formal studies with expert feedback processes. Minimally, the criteria we apply are:

- **Content relevance:** Are indicators well matched and relevant to the construct theory and knowledge bases?
- **Content representativeness:** Do indicators cover all pertinent content strata and levels tied to the construct theory and knowledge bases proportionately?
- **Organization and coherence:** Are indicators organized in a reasonable manner consistent with the construct theory and knowledge base?
- **Clarity:** Are indicators written clearly enough to allow the easy design/selection of items and remaining assessment operations?

Within given domains and subdomains, a *few* items ought to be designed to test the measurability of indicators either taken individually or in coherent clusters. Should assessment designers face barriers with this step, the likelihood is that the domain needs further breakdown or clarification. As provided in Box 4.5, a revision of the indicators and items should follow, using an iterative process, to finalize the domains and continue the design process.

QUICK RECAP: A taxonomic analysis of indicators allows designers to predesignate the exact dimensions to be measured by each indicator or groups of indicators, and as needed, to reorganize the domain accordingly into more homogeneous and logical arrangements. The assessment design literature offers several optional taxonomies for this step. Applications with four *functional taxonomies*, derived from old and new literature sources on the topic, were developed and demonstrated in this section; each taxonomy applies to a given construct type, assuming multidisciplinary assessment design projects.

The final steps in the domain specification process are evaluative, guided by 10 criteria (Box 4.5). The purpose is to correct for persistent ambiguities or major gaps left in the domains (such as oversights in relevant literature or mechanics of indicator specification). Depending on the formality of the endeavor, this step can be more or less extensive.

Reflection Break 4.5: Objectives 4–6

- Classify the construct type for the indicators below. Fill in the blanks with the appropriate functional taxonomy to categorize each type or level of the indicators in italics below. Justify your classification.
  - *Compose a story for children ages 5–7.*
    Construct type: _____
    Taxonomic category: _____
  - *Behave ethically in the workplace.*
    Construct type: _____
    Taxonomic category: _____
  - *Follow rules while driving cars on main roads and highways.*
    Construct type: _____
    Taxonomic category: _____

Construct type	Levels or types	Definition	Indicator or item examples
4.0 Sociodemographic constructs	4.1 Demographic factors	Construct categories based on biological, morphological, or physical characteristics	Sex Race Ethnicity
	4.2 Social class	Construct categories based on one's wealth, income, education, and/or occupations	Education level, socioeconomic status, "white- collar" vs. "blue-collar" workers
	4.3 Geographic, regional, or organizational membership	Construct categories based on membership in a defined country, region, or organized group	Nationality, religion, political party

BOX 4.5 Checklist for Evaluating Construct Domains

Ten Criteria for Domain Specification

1. *What to assess?* Are the different types of constructs identified separately?
   - Cognitive
   - Noncognitive
   - Physical or health-related
   - Sociodemographic
   - Other (clarify)
2. *Phase I specifications?* Are the construct(s) situated in the assessment use and user context?
   - Connected with *Whom to assess*?
   - Connected with *Why assess*?
3. Defensible data sources: Were appropriate data sources used to specify the indicators of the domain?
   - Existing literature and research
   - Curricula
   - Documentary sources, websites, and other knowledge sources
   - Expert viewpoints
   - Case studies, observation records, or other qualitative studies
   - Other (clarify)
4. *Organization:* Were indicators organized from general to more specific statements using a tree diagram format (or another reasonable organization tool)?
5. *Clarity:* Were indicator statements clear? Did they specify the necessary components clearly for each construct: behavior, content, condition, criterion performance?
6. *Coherence and homogeneity:* Was a suitable taxonomy of behaviors (or other appropriate taxonomies) applied to categorize similar types of behaviors to be measured in related groups?
7. *Item design:* Were a *few* examples of items or tasks created or selected to match indicators?
8. *Critical review:* Was the domain validated by the designers themselves, or more formally by peers and external experts?
9. *Revision:* If underspecified or poorly specified initially, was the domain revised using results of reviews?
10. *Quality of domain:* Is the final version of the construct domain defensible, observable, and measurable?

- *Demonstrate 20/20 vision, per optometric criteria, when driving a car at night.*

  Construct type: _____

  Taxonomic category: _____
- *Subscribe to the views of a religious sect.*

  Construct type: _____

  Taxonomic category: _____

- Which of the above indicators are too vague or broadly stated to allow sound item design or selection? If so, how would you improve it further? Explain.
  - Specify the domain for a construct you wish to measure, using the procedures discussed in Chapter 4.

4.6 Summary

Domain specification is a key step in sound instrument design. Chapter 4 introduced you to domain sampling theory as the main framework for specifying construct domains and subdomains. It presented several guidelines, taxonomies, and conventions to help operationally define unknown, complex, or ambiguous constructs with observable indicators. The chapter demonstrated these methods with examples falling under cognitive, noncognitive, health-related, and sociodemographic construct categories.

Domains could be conceptualized with simple, stratified, ordered, or unordered structures. Five main steps in operationally defining constructs involve locating appropriate construct theories and data sources; writ-

ing and organizing indicator statements; applying suitable taxonomies to designate a level or type to indicators and subindicators; creating samples of items to test the quality of indicators; and content-validating and refining the domain prior to item design or selection.

Domain sampling theory espouses the notion that all assessments yield behavioral samples from respondents or examinees, as tied to a theoretically grounded domain. Well-specified domains allow items to be matched to indicators through logical alignment processes. When executed thoroughly, this process builds content-based validity into the items and instrument, maximizing overall validity and reliability of the eventual construct measures. Empirical methods of validation should follow to verify the overall construct validity and quality of the measures.