Chapter 6

Designing Behavior-based, Product-based, and Portfolio-based Assessments

6.1 Chapter Overview

Historically, the term, performance assessment, refers to an array of different and specialized techniques (Berk, 1986). Chapter 6 provides background theory on, and specific guidelines for, designing three types of performance assessments, namely: **behavior-based**, **product-based**, and **portfolio-based assessments**. Broadly defined, interview-based assessments are also a type of performance assessment. These are discussed in Chapter 7 alongside **survey-based instruments**.

Looking back, Chapter 3 introduced the reader to the major types of performance assessments, outlining their general characteristics, advantages and limitations in Table 3.5. Chapter 6 builds on that information, showing how these assessment methods fill needs that traditionally designed, written assessments and other formats cannot. To bolster the information in this chapter, readers are encouraged to pursue relevant readings elsewhere. Several sources cited in the chapter could be useful. To connect this chapter with the rest of the book and the overall Process Model, see Figure 6.1.

Insert Figure 6.1 about here [Replicate and renumber Figure 5.1 as Figure 6.1]

6.1.2 Chapter Objectives

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

1. Identify the distinctive features of performance assessments—namely, behavior-based, product-based and portfolio-based assessments--describing their utility and applications in different disciplines.
2. Select performance assessment formats and modalities best suited to measuring indicators of cognitive, non-cognitive, behavioral, and health-related constructs.
3. Apply the Process Model and established guidelines/criteria for designing behavior-based, product-based and portfolio-based assessments with accompanying scoring rubrics.
4. Design different types of scoring protocols to suit assessment needs: analytic rubrics, holistic rubrics, rating scales, and checklists, including methods for obtaining "derived scores" from performance assessment data (e.g., norm-referenced scores; criterion-referenced scores and performance categories).
5. Evaluate the quality of performance assessments and scoring rubrics critically for the specified domains, populations, inferential needs and assessment purposes.

6.2 Behavior-based, Product-based, and Portfolio-based Assessments: Definitions, Examples and Origins

"Competence is a mental state which underlies and is enacted in performance......". (Mayher, 1990, pp 278-279).

"... performance assessments...require students [or other respondents] to generate rather than choose a response.....Exhibitions, investigations, demonstrations, written or

oral responses, journals, and portfolios are examples......" (Herman, Aschbacher, and Winters, 1992, p. 2)

6.2.1 Definitions

Cronbach as cited by Berk (1986, p.ix) defined performance assessments as procedures requiring: (1) the use of a variety of techniques, (2) a primary reliance on observations, and (3) the integration of information from more than one method of assessment or data source to draw inferences about the constructs of interest. Based on response modalities of examinees or subjects, this book has defined the three assessment methods that are the focus of Chapter 6, as follows (Chapter 3; see also Chatterji, 2003):

- **Behavior-based assessments-**
    - Assessment exercises where "live" or enacted behaviors, performances or demonstrations must be directly observed, recorded, and scored for obtaining evidence of some underlying skills, abilities, dispositions or other characteristics.
    - **Examples**: A ballet dancer's performance technique; a manager's negotiation skills with clients; behavioral symptoms of patients suggesting an underlying health condition; demonstrable aspects of professional attitudes towards any field of work; or a student's debating skills.
- **Product-based assessments-**
    - Assessments where tangible items or products that respondents create provide evidence of some underlying skills, abilities, dispositions or other characteristics.

  - **Examples**: A student's science laboratory report; published work of academics and scholars; a manager's annual financial report; a teacher's lesson plans.

- **Portfolio-based assessments-**
  - Assessments based on collections of work samples, creative products, recorded behaviors/performances, and/or personal reflections, gathered at a designated time point, or over a specified period of time, that together serve as evidence of some underlying skills, abilities, dispositions or other characteristics. Items collected by design in a portfolio could "showcase" a person's best work or document growth in given domains over time for individuals.
  - **Examples**: A photographer's portfolio of work; a student's portfolio of original compositions in poetry-writing and recitation, and so forth.

Performance assessments are useful for measuring several behavioral and non-cognitive constructs of interest in educational, workplace, clinical or community settings. For instance, several proficiency-based constructs we discuss in this chapter lend themselves to measurement with different performance modalities. In addition, clinical assessments often involve direct observations of human behaviors and functioning levels to screen or diagnose particular disorders. These assessments could be carried out with highly structured, behavior-based instruments or less structured, "anecdotal" observations. Both are discussed here.

6.2.2 Early Applications

Workplace Assessments Some of the earliest signs of performance assessments were recorded in governance-related workplace settings in China. Several accounts date these as far back as 1000-2000 BC (Du Bois, 1970; Thorndike, R.M., 1990; Ward et al. 1996).

The Chinese conducted very high stakes examinations for civil servants, where they employed results to select, recruit and promote public officials to high level ministerial positions. The subject area domains they examined included arithmetic, music, horsemanship, geography, civil law, writing, Confucian principles, and knowledge of public and private ceremonies (Allen & Yen, 2002; Cohen & Swerdlik, 2010). Of these, music, horsemanship, writing, and civil law suggest that different forms of performance assessments were a very likely part of their testing arsenal. Unfortunately, not much documentation is available on how the Chinese built the civil service examinations.

Educational Assessments There is also evidence that performance assessments were valued in early education movements. For example, "alternative" performance-based assessment ideas were prevalent among progressive education enthusiasts in America in the early 19th century. Historians note that there were many views and little consensus on the definition of **progressive education** when the concept first originated (Cremin, 1961). Generally, however, **progressivism** emphasized the principles of "learning by doing", focusing on building students' problem-solving, creative and critical thinking skills. Progressives also viewed education as a pragmatic means for developing social responsibility and democratic ideals in pupils—emphasizing skills and habits relevant for community-building, service and social reform (Dewey, 1897; 1902).

Student-developed projects and products typified progressive teaching and learning ideals. There was a clear leaning towards more individualized tasks and open-ended response formats. Important outcomes of the Eight-Year Study, an experimental project conducted from 1930 to 1942 by the Progressive Education Association, included what were labeled then as "innovative" student tests and "alternative" forms of assessment in participating schools (Aikin, 1942; Smith, Tyler & the Evaluation Staff, 1942; see also: http://education.stateuniversity.com/pages/1947/Eight-Year-Study.html).

Education reforms of 1990s revived a new brand of “alternative assessments” in American public education. This time the impetus was to counter the overuse of multiple choice tests of basic skills. There was also a recognized need to align testing and assessment methods to the then-emerging constructivist views of learning (Herman, Aschbacher, & Winters, 1992). Today’s standardized testing programs continue to incorporate constructed response and performance tasks, capitalizing on new technologies that offer optional assessment modalities to test-takers, such as, computer-based, interactive tasks (see item examples in SBAC at: http://www.smarterbalanced.org/assessments/sample-questions; and PARRC at https://parcc.pearson.com/practice-tests/).

Performance Assessments in Intelligence Testing Concurrent with the progressive education movement, Alfred Binet (1909) created one of the earliest performance tests of human intelligence, incorporating a combination of verbal and non-verbal tasks for multi-age populations of children. Three of the historically-recorded items are (Crocker & Algina, 2006, p. 9-10): Point to eyes, ears and nose (age 3); Indicate omissions in a drawing, and repeat with five figures (age 7); and Criticize sentences containing absurdities (age 11). The Binet (1909) test was intended to assess “educability” levels of children so that they could be placed in appropriately-matched education settings.

To design the assessment, Binet observed and recorded children’s performances directly on individually-administered tasks. He then compared the performances of his subjects against their chronological ages, interpreting the results as their “mental ages”. Binet’s methods involved numerous tryouts with tasks of varying difficulty on different-aged children. In this way, he was able to differentiate and sort amongst the capabilities of children at different developmental levels by age. His age-referenced **norms tables** for interpreting measures of human intelligence were a trend-setting marker in early test construction techniques (Crocker & Algina, 2006).

That early assessment became the precursor to the current Stanford-Binet Intelligence Scales. The most current version is also an **individualized assessment**, comprised of non-verbal and verbal performance tasks arranged by degree of difficulty. The modern test is administered by trained psychologists. In the main, it is used for the purposes of diagnosis and placement of individuals with special needs in intervention programs at school or clinical settings (Roid, 2003; Walsh & Betz, 2001).

Reflection Break

A. **Review each item below (a-e). Identify the construct, population, inferences and uses implied in each. Indicate which assessment method(s) you would prefer for each (a-e). Rationalize each choice.**
   - **behavior-based?**
   - **product-based?**
   - **portfolio-based assessment?**
   - **a combination?**
   - **Other?**

a) A mental health counselor's history-taking abilities with clients, before s/he provides services
   a. **Construct, population, inferences and uses:**
   b. **Choice of assessment method:**

b) The accuracy of a psychologist's formal written report following testing that will then inform career counseling services for a client.
   a. **Construct, population, inferences and uses:**
   b. **Choice of assessment method:**

c) A teacher's classroom competency level based on the syllabus s/he prepared for a science course
   a. **Construct, population, inferences and uses:**
   b. **Choice of assessment method:**

d) A manager's ability to lead ethnically diverse teams towards reaching an organization's goals.
   a. **Construct, population, inferences and uses:**
   b. **Choice of assessment method:**

e) A doctoral student's depth of knowledge of a topic based on the oral defense of his/her dissertation

a. **Construct, population, inferences and uses:**
b. **Choice of assessment method:**

B. **Identify an assessment scenario of your own where a performance assessment would be the most suitable method. Explain why.**

6.3 Distinctive Properties of Performance Assessments: More than a Change in Format

What are some distinctive properties of performance assessment modalities that are advantageous for assessment designers? These formats are unique in that they allow us to (a) “see” a construct differently, (b) tap into “situation-specific” aspects of a construct, and (c) incorporate important “facets” of a construct into the design and scoring structure of an assessment. In addition, these methods are often better suited for measuring complex constructs for which other assessment methods have proven to be limited. As the following examples illustrate, in some cases, well-designed performance assessments help assess the more demanding levels and types of mental processing than other formats. A discussion follows.

6.3.1 “Seeing” Construct Complexity

Performance assessments of different types afford us with new ways of seeing, or being able to see, constructs that we might otherwise not be able to access or measure. Behind each performance assessment is some competency area, or a non-cognitive, behavioral or health-related domain that becomes evident only when respondents behave, perform, engage actively, or execute multi-step tasks to produce something. With multi-dimensional and complex domains, more than one performance assessment method, or a combination of assessment methods, may become necessary. The performances could be measured in natural, or simulated, structured settings.

Consider measurement of teaching competence. This is an example of a complex construct domain where one assessment format is rarely sufficient for measuring the full breadth and scope of requisite indicators. In discussing the construct, Mayher (1992) observed that "...the structure of [teaching] competence is so complex, as are the contexts which influence and affect how it is employed,... [that one] cannot determine the competence by only observing performance" (pp. 278-279). Complex competency domains encompass a diverse array of skills, abilities, dispositions, or mind-sets. In such cases, performance assessment methods allow a window into those dimensions that traditional written tests are unable to reach. To measure such domains fully, assessment designers or concerned researchers must select the most apt formats for the given purpose, weighing validity, reliability and utility issues that apply.

Another setting where performance assessment formats facilitate our ability to "see" the construct is competency in a profession, such as, medicine. In the journal, Academic Medicine, George Miller (1990) delineated a hierarchical and cumulative framework for assessing different, but demonstrable levels of medical competence on a continuum of expertise. The levels progressed from a beginner level performance ("Knows", expected minimally of medical students at the undergraduate level), to the intermediate level ("Knows How" and "Shows How", expected in resident physicians undergoing graduate level training), leading up to an expert level ("Does", expected in fully-trained doctors in practice settings). He concluded that "no single assessment method could provide all the data required for judgment of anything so complex as the delivery of professional services by a physician" (Miller, 1990, p. 563), underscoring a need for multiple assessment modalities. Miller's pyramid is shown in Figure 6.2.

In Chapter 4 we encountered several taxonomies akin to Miller's pyramid that instructional and assessment designers could apply for better representation of the embedded components of complex constructs, starting with Bloom's taxonomy. Miller's framework is particularly relevant for designing performance assessments of competency-based domains, both in medicine and other professional fields. More follows on this taxonomy in a later section.

For a final example, think of "Management Competency" in workplace contexts. This is a very broad and multidimensional , according to the literature. It comprises a diverse array of skill-sets, abilities, attitudes and personality traits expected of leaders and managers of organizations (Freeman, 2010; Northouse, 2018; Sedera & Gable, 2010; Thornton & Byham,1982). Up-to-date job analyses could reveal sector-specific variations in definitions in different industries, but, the dimensions common to most comprise a long list: Communication Skills; Initiative; Leadership Abilities, Stakeholder and Customer-Orientation; Motivation and Energy Levels; Budgeting and Planning; Organization; Delegation; Negotiation; Nimbleness and Learning Agility; Knowledge/Information Management; Recruitment and Hiring; Resource Management; and Judgment and Decision-making. Clearly, the layers of this domain are many, and each dimension could be treated as a separate construct in and of itself.

The first sub-domain listed, Communications, includes a range of writing and speaking skills that effective managers should display at work. For instance, how skillfully managers can articulate a vision in a formal written statement, or how effectively they communicate that vision to employees, are best revealed through performance assessments. While written communication skills could also be assessed through essays, selection of performance modes provides a different lens for "seeing" the construct couched in authentic conditions of the workplace. Here, product and behavior-based assessments could together capture particular aspects of the domain that other modes cannot.

Insert Figure 6.2 about here [Miller's pyramid]

6.3.2 Situation-specific Performances

A second distinctive feature of most performance assessments deals with their situation-specificity. That means that some parts of the targeted domain are exhibited conditionally, or under certain stimuli specific to a particular setting or context, such as, schools, workplaces, homes or communities.

Continuing with Miller's example, suppose that doctors in residency training must "Show how" to perform a physical examination of patients in a hospital setting. Here, they must demonstrate they can perform the task amidst various hospital and infrastructure-related stimuli and conditions. The competency domain is thereby defined by physicians' abilities to provide the appropriate responses and take actions in that context. Situation-specific performances raise the demands and difficulty levels of tasks, as compared to assessments situated in artificially structured or simulated settings.

For an example outside medicine, consider the Performance Based Teacher Education (PBTE) program of the 1990s. Here, assessment designers focused on observable aspects of teaching competence demonstrated by teachers in live classroom contexts. Evidence of the targeted teaching capacities had to be obtained while teachers taught their own students situated under the natural conditions of their schools (Mayher, 1990). The real environments added context-based challenges and unpredictability to the construct definition, increasing difficulty levels of the tasks that were observed.

More recently, research has focused on situation-specific, performance assessments of teaching competence that reveal the subject matter expertise of teachers. Gitomer, Phelps and colleagues (2014) and Gitomer and Zisk (2015) specified what competent teachers with greater mastery of content knowledge should be able to do while teaching classes in mathematics and English language arts. Table 6.1 shows an excerpt of their domain specifications.

In the list of indicators is a teacher's ability to anticipate student misconceptions, challenges or questions in the subject area; explain concepts in multiple ways to improve student understandings; and perform the academic tasks they expected of students themselves, in an expert and fluent manner. Interestingly, the authors defined subject matter competence with the similar sets of situation-specific indicators in both math and English Language Arts, suggesting that these indicators of subject matter competence would apply generally to teachers regardless of their specialty (Gitomer and Zisk, 2015, pp.22-23). Note observational assessments would be used here to infer levels of subject matter knowledge in teachers.

Insert Table 6.1 about here [Gitomer et al table]

Assessments of constructs like "Management Competency", discussed earlier, may also call for situation-specificity, altering the inferences we would draw about the skills, behaviors or attitudes depending on where these are exhibited. It is therefore necessary to recognize whether particular settings or conditions are a part of the domain definition.

For instance, to assess how competently a manager communicates with a client during a complex negotiation process, an exercise must be situated in a live negotiation event at the workplace. Alternatively, we could capture such performances via a realistic, but simulated exercise. "Assessment Centers" of the early 20th century first introduced the notion of **situational testing** in the business management literature (Bass, Burger et al, 1979). Assessment centers utilized both real and simulated assessment conditions.

Simulated assessments are structured in artificial contexts where conditions mimic actual events but with controls instituted deliberately by designers. In education applications, a modern technology-based example of simulated, situational testing is called a "serious game". Serious games are interactive, video-based exercises accompanied with behavioral checklists or rating scales that are gaining utility in medical education environments (see for example, Blum et al, 2018).

6.3.3 Facets of Performance

Finally, performance assessments have distinctive properties that help designers think about the facets of performance that should be factored into the scoring procedures. For assessment designers, the term, facets refers to the structural parts of an assessment that interact together to generate the score for respondents or examinees. Collectively, the facets of the assessment help in operationally defining the underlying construct in a more complete form.

A typical paper and pencil test, for example, has two main facets--the items and respondents/examinees (Persons x Items). Here, the examinees interact with the items to produce the total scores. Because they require involvement of human judges, a typical performance assessment includes at least three facets: (1) the items or tasks, (2) the persons, and (3) the observers/judges. In this sense, a performance assessment can be thought of as a multi-faceted assessment system (Brennan, 1984; Linacre & Wright, 2002; Shavelson & Webb, 1989).

Recognizing the number and types of facets helps in the performance assessment design process. Let us consider a previous example. Suppose we are assessing resident physicians' competency levels in formulating patient treatment plans using portfolio-based assessments. Assume that two groups of raters are involved in observing and providing ratings on the junior-most residents at a medical school during their training years. The raters are the senior resident physicians and supervising, "attending" physicians. Let us also say that the culminating, performance outcome falls under the "Shows how" level of Miller's taxonomy, and that there are four specific indicators subsumed within the competency domain as shown in Box 6.1 The taxonomic levels of the culminating general outcome and embedded skills are indicated in italics.

Insert Box 6.1 about here [domain treatment plans]

To tap into the domain, let us say that assessment designers selected a portfolio-based assessment method as the best fit for the measurement needs in this scenario. In the design specifications, a collection of treatment plans (product-based assessments) combined with observations of residents completing the specified tasks in hospitals or clinics (behavior-based assessments), were deemed as appropriate means for chronicling growth of the trainees on the indicators of the domain.

There could be five facets pertinent for designing the portfolio-based assessment in this scenario, as spelled out in Table 6.2.

- Facet 1: Tasks (A Given Range of Patients' Health Problems)
- Facet 2: Settings (Clinics, Hospitals)
- Facet 3: Sub-domains of competence (Indicators a-e in Box 6.1)
- Facet 4: Raters (Senior Residents, Physicians)
- Facet 5: Examinees (Resident Physicians-a cohort of, say, 20)

Insert Table 6.2 about here [facet Grid]

The five facets in the scenario have elements ranging in number from 2-20 shown in Table 6.2. Each resident's final score on the assessment could be produced through the interaction of some, or all, of these different facets of the assessment system. An important, guiding design question in such applications is: Which facets should count in producing that score? Such a decision must be taken by assessment designers or researchers involved, keeping in mind the desired inferences and uses to be made with the scores.

For example, a total score could be produced based only on Tasks and Examinees (Facet 1 and 5), regardless of the remaining facets. The question to be answered here would be:

- How did the residents perform on the entire domain of skills (a-e), across the specified range of patient problems, assuming that the raters, sub-domains of performance, and settings are similar and interchangeable?

If this approach is taken, priority is placed on residents' competence in treating a variety of patient problems, and only one score is produced from the assessment. The validation burden is to show that the different raters, settings, and sub-domains do not yield systematic variations in the scores (or construct measures), and are indeed inter-changeable. This is the most common design approach.

Alternatively, we could decide to produce separate scores for Tasks and Examinees sorted by Setting (Facets 1, 2, and 5). Here, the question would be:

- How did the examinees perform on the domain across the full range of patient problems, sorted by type of setting, assuming any variability on other facets is controlled or negligible?

In the latter approach, designers would expect different scores/construct measures for residents by domain/sub-domain and setting.

Taxonomic levels of the different indicators by sub-domain could also be treated as another facet of an assessment, depending on user needs. Miller (1990, p. 563-565) observed that medical assessments often fall short in sampling the three highest cognitive levels of his pyramid adequately, and in practice settings, the total number of observations is too few on medical students and residents. The inclusion of a higher number of facets would add greater complexity to the design, no doubt, and pose implementation barriers. On the other hand, facet grids like the one in Table 6.2 could also help in limiting the design from becoming too unwieldy or impractical.

Facets selected could vary by type and number, and include demographic variables of the population like gender or year of training. Because the facets affect the number and types of scores eventually produced, the information that users seek should be factored into the decision on which facets should matter. Ideally, the facets in a performance assessment must be specified at the start.

From a validity standpoint, content-relevance and content-representativeness is contingent on the adequacy of the sample of observations we take from the cells of a many-facet grid. Returning to Table 6.2, during the design process, we could ask: How should we sample observations from the cells to optimize levels of validity, reliability and utility in residents' scores? Review the sampling decisions to be made in Table 6.2. To obtain valid scores for competencies a-e by setting, for example we should collect observations of each resident in both clinics and hospitals, across the specified range of patient problems, performing all the tasks implied in a-e. To obtain reliable scores, the total number of observations on each resident physician should be large enough to obtain a reliable reading of how competent they are at specific points of their training.

There is no fixed "rule of thumb" on the ideal number of observations necessary in assessment cases like the one above. In the end, validity and reliability are empirically or formally ascertained through appropriate examinations. Regardless, our design decisions must be made judiciously, accounting for the number of facets critical to the construct definition and of value to potential users of the assessment system. The resources available for implementation of the assessment, the intended inferences and assessment uses, and the stakes to be tied to decisions made with results must all be borne in mind. A larger sample of observations is always recommended when the stakes tied to score-based inferences are higher.

Finally, a facet grid (Table 6.2) combined with the domain (Box 6.1), could together serve as **design specifications** for assembling performance assessments. That is, they could provide a blueprint to guide the design or selection of the assessment. Recall that Phases I-III of the Process Model deal with specifying the assessment context and operations before assessments are designed or selected (Figures 6.1, and 1.6 in Chapter 1).

Reflection Break

- **Would you agree that Miller's taxonomy could be useful in designing performance assessments for competency domains outside medicine or medical education? Why or why not?**
- **Identify types of constructs where Miller's taxonomy is inappropriate. Give examples.**
- **Identify examples of performance assessments that should be situation-specific to (a) schools, (b) workplaces, (c) health care settings, and (d) elsewhere in the home or community. Which context factors are critical to defining the construct in each example?**
- **Identify the facets you would value in a performance assessment you would design. Give three reasons to explain why.**
- **In your opinion, how many facets are relevant to designing a sound performance assessment system for assessing how well a group of musicians are able to interpret and play written musical scores at (a) rehearsals and (b) during formal concerts? Results will help select members for the local orchestra.** Name each facet. Draw the grid that could serve as part of your assessment design specifications. How would you sample from each facet to maximize (a) construct validity (b) reliability, and (c) utility, of the scores?

6.3.4 An Applied Example: Behaviour-based Assessment of a Neurological Disorder

Performance assessment modes are not simply useful for measuring competency-based constructs. See Box 6.2 for a case showing how clinical researchers are designing and validating performance measures of a health construct in very young children. Hypertonia is a behavioral symptom of a variety of pediatric neurological conditions. Most common among these is Cerebral Palsy (CP). Children with CP develop movement and posture disorders of different types, causing activity limitations that are attributed to disturbances that occurred in the brain when they were infants or fetuses.

The Hypertonia Assessment Tool is a behavior-based observational procedure intended to measure hypertonia (movement disorders) more generally, but mainly to help diagnose three subtypes of movement disorders in children with CP called dystonia, spasticity, and rigidity.

Children with both “positive” (with CP) and “negative” (normal or without CP) symptoms were in the sample for the validation studies, and the authors reported on agreements found with both the positive and negative cases. Once designed, items were content-validated by experts. Percent agreement in classifying children from several samples was the main statistical method of analysis.

Results of their validation study are excerpted in the table in Box 6.2. Item validation and reduction eliminated seven of the 14 original items. On validity, the percent agreement level was adequate with adjusted kappa values ranging from low (0.30) to excellent (1.0) based on expert agreement levels in classification. Test–retest reliability based on agreement of classification rates on two occasions, two weeks apart, was poor for dystonia items at .30, but 1.0 for other subscale items. Interrater reliability showed agreement rates from .30-.91. Agreement level was lower for the dystonia items, with kappa ranging from fair (0.30) to good (0.65).

If we were applying the Process Model to the case, we would find that authors specified the instrument was intended for clinical intervention or research purposes, where researchers wish to infer levels of hypertonia in children observed (Why?). The construct is hypertonia, a pre-existing health condition in children (What?). The population that the researchers focus on consists of North American children from two months to three years in age (Who?). This is Phase I.

The authors reported on their systematic item design work (Phases II-II). This was followed by validation studies examining content and convergent validity, along with internal consistency, test-retest and inter-rater reliability investigations (Phase IV). Results on the items and overall tool are excerpted in Box 6.2. Chapters 9-10 will detail each of these psychometric procedures the authors applied. To forward the researchers' goals, what next steps would you recommend if you were following the Process Model?

Insert Box 6.2 about here [HAT case]

6.4 Selecting Appropriate Performance Assessment Modalities

We turn now to selecting the best-matched assessment method or mix of methods to measure given constructs, a key part of the design process. We will consider five main guiding principles, or tasks, for selecting the assessment modalities best-suited to given contexts. They flow naturally from the domain specification stages of assessment design and are as follows. The tasks need not be performed as steps in a sequence.

(1) Clarify and understand the sub-domains and types of indicators that define the construct in terms of the embedded content, taxonomic levels, and conditions (i.e., situation-specificity).

(2) Identify the process- and/or product-indicators embedded in the domain.

(3) Match particular indicators or sets of indicators with best-fitting performance modalities.

(4) Weigh the performance modality chosen against assessment purposes specified by users.

(5) Evaluate how well the selected assessment methods uphold validity, reliability and utility in the user contexts.

For continuity, we will review and apply these guidelines to three already-familiar constructs: medical competency in patient treatment planning; managerial competency in communications; and hypertonia in children.

6.4.1 Clarify and Understand the Construct

To begin, it is necessary specify the domain for a construct, and if ambiguous, to break it down further in terms of sub-domains, each delineated clearly with component indicators and sub-indicators. As a part of this step, a taxonomic analysis of indicators and sub-indicators aids in clarifying What it is that we are trying to measure. Refer back to Chapter 4 for more on domain specification procedures.

Most construct domains call for multiple levels and types of performance, as shown in Box 6.1. Taxonomies help designers sort the various processing demands placed on respondents or expected of them, so that tasks or exercises that are optimally aligned can be devised. Sometimes, two or more taxonomies could shed light on nuanced aspects of the domain, such as, specific types of cognitive, affective, or behavioral processing that we would like to measure, and situations where they would be manifested.

For example, review Miller's (1990) pyramid again in Figure 6.2 in tandem with Box 6.1. The framework categorizes four progressive levels of cognitive complexity, each placing different performance demands on physicians. Although originally intended for assessing the performance of doctors-in-training, the taxonomy translates usefully to competency measurement in other professions or educational settings, as well.

At the bottom of the pyramid is the "Knows" dimension. Through tasks at this lowest level, medical or other professionals would demonstrate that they have the mastery of knowledge necessary for executing their job-related functions competently. The "Knows" domain tends to receive the most attention in existing medical assessments (Miller, 1990). Knowledge-based domains are assessed easily and efficiently with traditional, written item formats discussed in Chapter 5, whether administered via paper and pencil or technology-based media.

The next level above in the pyramid is the "Knows how" level. Indicators at this second level are less frequently tested, and already we find that written formats might be too limiting to apprehend them. A competent physician's task at the "Knows how" level would involve multi-step tasks requiring complex procedural skills specific to their job functions and roles. Refer back to the definition of Complex Procedural Skills in Table 4.2, as needed.

We also see that the higher levels of Miller's pyramid ("Knows How", "Shows How", "Does") call for complex processes that are revealed typically through demonstrable behaviors rather than written tests. In addition, some of these mental processes could also generate products. For example, a patient history-taking exercise entails complex cognitive processes, but preparing patient-specific treatment plans supported with documentation are the resultant products. One or both types of evidence could be the focus of assessments we design.

The two most demanding performance levels in Miller's pyramid (1990) are "Shows how" and "Does" in Figure 6.2. Note the situation-specificity at the higher levels. The "Shows how" level deals with physicians' capacities to demonstrate competencies with real patient cases, but while still in supervised, educational settings. The "Does" level is evidenced in real hospital care and clinical environments unsupervised, after doctors are fully trained. For assessing these two highest levels of the Miller pyramid, we need direct evidence of the requisite knowledge and cognitive processes at work while physicians engage with patients in the desired settings. The best avenue, therefore, may be through direct observations of physicians' behaviors in action. That is measures must be taken while doctors are actually examining patients and negotiating various hospital- or clinic-related conditions in situ.

We must recognize that the levels are cumulative in Miller's pyramid. So, the "Knows how" level involves knowledge at the "Knows" level, but also engages added and deeper cognitive processes in the defined contexts. Several higher order thinking and decision-making skills at the "Knows how" level are testable through scenario-based, written items as well (see Chapter 5). Product-based or behavior-based modes, however, and would enable users to "see" other aspects of the competency domain through situated testing.

6.4.2 Identify Process- and/or Product Indicators in Domains

Also helpful during selecting assessment methods, is obtaining clarity on whether the domain, once specified, includes any product or process indicators. By definition, product indicators are formally stated expectations where the best evidence of construct-related processes at work is obtained through tangible items examinees/respondents produce. A process indicator, in contrast, is one where we require a demonstration of multi-step behaviors and formal processes. Usually, the latter must be directly observed for obtaining evidence of the underlying capacities (Escalas & Luce, 2004; Fink, Ward & Smith, 1996; Simon, 1978).

The constructivist movement in cognitive psychology, concerned with how a learner actively constructs meanings about this world, holds that it is beneficial to assess both the processes and products of learning through applied activities (Herman et al., 1992; Resnick & Resnick, 1992). In scholastic learning contexts, process outcomes deal with how an individual completes a task or solves a problem. They involve cognitive or other multi-step procedures that students employ to find solutions. Using a scientific method to conduct a research study is an example of a process outcome.

Product outcomes, on the other hand, focus on the final result or solution to a task. Calculating the area of a room, or writing a book report are examples of product outcomes. In some school-related tasks, one could emphasize both process and product outcomes during assessment, such as in the "use of a writing process (process outcome) to compose a story (product outcome)." In the event that both the process and product are valued parts of the curriculum, we may have to make a decision as to whether we wish to assess both or just one of the two targeted outcomes.

Examine Box 6.3 now, which demonstrates how this principle could be brought into play with Communication Skills, the first dimension of “Management Competency”. The taxonomic classifications are at the Higher Order Thinking and “Does” levels for the two illustrative indicators, “Write a vision statement..” and “ Communicate through speeches...”. Honing in on these as process versus product indicators makes clear the different performance modalities that would be necessary to measure each at the taxonomic levels specified.

Usually, a product-based assessment format is the superior assessment method from a validity standpoint when application of given capacities yields a tangible product. These indicators are also referred to as product outcomes in educational or training contexts, as indicated. In contrast, a behavior-based assessment becomes the better choice to capture actual performances-on-the job by doctors or professional trainees, and others in educational settings. These indicators are also called process outcomes in educational settings, drawing on the constructivist school of thought (Escalas & Luce, 2004; Fink, Ward & Smith, 1996; Simon, 1978).

Insert Box 6.3 about here

6.4 3 Match Indicators to Best-fitting Performance Modalities

Review again the applications in Boxes 6.1 and 6.2. Fit of particular indicators or sets of indicators to performance modes is ascertained based on the (a) content, (b) taxonomic level, or (c) conditions specified in indicators, or a combination of these factors. If the construct domain is clarified with observable and clearly specified indicator statements, it will aid in executing this step.

Let us reconsider the case of assessing hypertonia in young children. Two indicators of dystonia reported in the paper (not shown in Box 6.2) dealt with dystonia symptoms that children would exhibit when sleeping or when awake but at home that caregivers are best-equipped to report. These were (Jethwa et al, 2010):

- History of variability in (muscle) tone with sleep compared with awake time
- History of an increase in (muscle) tone with activity⁄movement

Given the situation-specificity of the indicators, the researchers opted for an interview-based assessment format with parents or care-givers. For the remaining, they decided on a behavior-based format performed by clinicians. Again, where more than one assessment mode is necessary, the final decision should account for validity, reliability and utility issues, taken together.

6.4.4 Consider the Assessment Purposes

The specified assessment purposes could also guide the choice of appropriate assessment methods. For example, with the scenario in Box 6.1, supposing assessment users express a need to track how well doctors-in-training show improvements on the entire skill set over time, or with different patient problems and clinical challenges. Here, a **portfolio-based assessment** format might be more suitable, as opposed to any one type shown in Box

6.1. Portfolios would offer a mechanism for collecting multiple, patient-specific treatment plans that doctors prescribed from year to year. Collectively, this evidence could be used for the purposes of point-in-time job performance reviews, or for assessing growth of physicians still undergoing training on an annual basis.

6.4.5 Evaluate Choice of Performance Mode using Validity, Reliability and Utility Criteria

Which validity, reliability, and utility threats should designers worry about when making a choice of assessment method(s)? This step involves ensuring that the selected assessment formats will be functional, but also hold up to standards of quality, especially when measures are evaluated psychometrically (AERA, APA & NCME, 2014).

Performance assessments present unique measurement challenges not encountered with other assessment methods. Depending on one's perspective, we could view the properties of performance assessments simultaneously as their strengths or weaknesses. A main strength is that observers are an integral part of the assessments we design. Human judges could add unique insights to, and "ways of seeing" constructs. On the other hand, we must now contend with unknown levels of subjectivity and sources of error that could impede the quality of measures, outside challenging utility

To begin, a key to assuring overall construct validity of scores from performance assessments is a well-specified construct domain. The domain supplies the essential framework of indicators and sub-indicators for designing or selecting tasks, behaviors and exercises that match, as well as, for developing scoring rubrics.

Tasks, behaviors and exercises must reflect the targeted content, taxonomic levels, and conditions specified in the indicators. To be defensible, again, the domains and scoring criteria must be consistent with established theory and reasonable knowledge bases and data sources (Chapter 4). Miller's (1990) pyramid may serve as a useful general resource outside medicine or medical education, but it should be applied with thought and judgment.

Second, identifying the relevant facets, and sampling the observations appropriately from a many-facet grid will enhance quality of measures. To be defensible, the facets selected—tasks, persons, sub-domains, levels, settings, raters, or other conditions--must fit accepted definitions of the construct and user-specified inferential needs and uses.

Issues of reliability stem from random or unknown sources of variance in scores. An unduly small sample of observations on subjects is a commonly encountered handicap of performance assessments. In a multiple choice test, 10 questions sampled from a homogeneous domain is often a large enough set of observations to generate a reliable score. With performance assessments, making 10 observations on each person is a prohibitively high number, and may be onerous in practice or research settings. Various psychometric tools and techniques can guide decisions on the optimal number of observations necessary for particular assessment purposes, some of which we will consider in Chapters 9-11 (Brennan, 1984; Linacre & Wright, 2002; Shavelson & Webb, 1989).

Added sources of error could lower reliability in performance assessments. These include loosely applied scoring rubrics, inconsistent behaviors of raters and observers, or uncontrolled assessment settings. Commonly applied strategies to avert or minimize these threats involve (a) observer and rater training procedures, (b) standardization of rater instructions; (c) standardization of materials, conditions, and administration protocols, as well as, (d) post-hoc statistical treatments of the data to improve the quality of information on individuals or groups of examinees/respondents. Addressing such issues proactively during the design process usually improves the quality of measures we obtain in the end.

With regard to utility and efficiency, performance assessments typically demand more time, materials and human resources than structured response tests or other group-administered assessment methods. These assessment modes should therefore be adopted when we cannot measure the constructs of interest satisfactorily with other methods. When utilized, we must ensure there are sufficient resources available to maintain and implement the assessment systems. Stakes tied to decisions in applied contexts is always a useful guiding criterion on when to adopt performance assessment methods.

Reflection Break

- Think of a competency domain in a profession or discipline of your choice (e.g., education, counseling, mental health, nursing, social work, medicine, business or other). In your selected domain, identify 1-3 indicators at each of these levels from Miller's pyramid, and state them using standard conventions (see Chapter 4):
  (a) Knows
  (b) Knows how
  (c) Shows how
  (d) Does.
- See above. Identify individual or sets of indicators best-measured with one or more of the following formats: behavior-based, product-based, or portfolio-based assessments. Justify your selections.
- Identify another domain (or parts of a domain) that is not a competency-based construct, but would be best-measured with one or more of the following formats: behavior-based, product-based, or portfolio-based assessments. Which taxonomy from Chapter 4 (Table 4.2) did you use to classify the indicators? Justify your choices.
- For students in K-12 or higher education contexts, which of these is a (a) process indicator, (b) product indicator, or (c) a combination? Justify.
  - Conducting an experiment in chemistry.
  - Writing the report of a scientific investigation.
  - Writing a novella in creative writing course, using an iterative writing process.
  - Cooking a dish in home economics class.
  - Employing iterative steps of outlining, drafting, editing, rewriting to compose an essay in final form.
- Give examples of a **process indicator** and a **product indicator**. State these in measurable terms. Distinguish them from **process/product outcomes**.
- How would you prioritize among validity, reliability and utility when selecting the best performance modality? Explain.

Insert Box 6.3 about here

6.5 Applying the Process Model to Design or Select Performance Assessments

Building on the previous concepts, three examples now follow to demonstrate how to apply the Process Model to design or select a performance assessment for a given user context. In order, we review the steps for designing a behavior-based assessment in a professional training scenario in detail. Next, we review a product-based assessment for a health information training scenario in the workplace, and a portfolio-based assessment in K-12 education in abbreviated form. We conclude the chapter with criteria to guide content validation reviews of scoring rubrics and performance assessments.

6.5.1 Designing Behavior-based Assessments

For this example, consider Boxes 6.4 A-B and Table 6.3 together. The construct is a competency domain for trainees participating in specialized, hospital-based, physical therapy educational programs, where the patients are still recovering and housed in cardiac care units. The instrument in this case was developed by a faculty member affiliated with such a hospital-based educational program unit at Mount Sinai Medical Center in New York. Phase I begins by answering the Why?, Who?, and What? questions of the Process Model, followed by specifying the assessment operations in Phase II, which then leads into the design of the instrument in Phase III.

Phase I. Specify the assessment context

Box 6.4A delineates the assessment context in detail, indicating the score–based inferences to be made with the results at programmatic and individual trainee levels. Multiple observational assessments were to be made on trainees during the course of their training, before and after instructional exposure. The results were meant for making proficiency-based inferences vis-à-vis the domain, and for fulfilling formative and summative decision-making needs at the individual trainee level, with averaged "total scores" used to evaluate program effectiveness by faculty and program-level leaders.

Phase II. Specify the Assessment Operations

The domain, as represented with a "general targeted outcome" in Box 6.4B, is specified at the Complex Procedural Skills level with three, embedded, performance-based indicators 1.1-1.3 subsumed. The expectation was that trainees would perform the culminating outcome at the "Shows How" level of Miller's taxonomy towards the end of the training program, or under instructor supervision. To specify the domain, the designer used several relevant documentary sources that provided the current, best practice guidelines in physical therapy. The domain overall specifications in Box 6.4 A-B provided the framework for assessment tool and scoring protocol produced, a portion of which is depicted in Table 6.3.

It should be noted here that the complete set of domain and design specifications are not shown. The author broke down each specific indicator further into several enabling skills, for example, which facilitated the design of the scoring protocol we see in Table 6.3. For example, two sub-indicators under Indicator 1.1 were as follows; these were further broken down to derive Items 1-10 in Table 6.3.

1.1 Investigate the medical condition of patients (Complex Procedural Skills)

> 1.1.1 Apply information on individual history and physical findings of patients to make decisions—pre-admission notes, medical history, consultations with doctors and medical staff (Application)
>
> 1.1.2 Evaluate the pathophysiology of the disease process of patients to make decisions, using relevant information (Higher Order Thinking-Evaluation)

A review of even the excerpted domain in Box 6.4B suggests that a behavior-based assessment mode is clearly a logical match. Details on facets, score types and bases for weights and point allocations, the situation/conditions under which data were to be gathered, and so on, complete the assessment design specifications.

A number of decisions were made by the designer. The Table of Specifications shows that the scoring placed an exclusive emphasis on three kinds of mental processing: Complex Procedural Skills, Application, and Higher Order Thinking. Does that mean that measuring physical therapists' knowledge is not important to the profession? It may not be quite that simple. Based on the judgment of the assessment designer and domain experts he consulted, we see that those taxonomic levels were a greater priority in the curriculum. The behavioral tasks in Items 1-10 in Table 6.3 are also better-tapped by the assessment modality they chose. Finally, the prioritized facets and score types suggest a value for demonstrated mastery by skills in the domain and sub-domains, but allowing for some variability on other facets, including instructors who rate. A provision is made to ensure that instructors are trained in the domain-referenced rubric and rating protocol.

Phase III. Design the Instrument

Table 6.3 shows an excerpted portion of the final instrument. The assembly and compilation of a behavior-based assessment exercise requires attention to four needs:

(a) writing the **items** or **tasks** tied to indicators, sub-indicators and enablers in the domain, as shown, or selecting the same;

(b) developing standardized guidelines and a system for collecting **observations** systematically to support user-specified needs—here, classroom instructors in the main would be making both formative and summative uses with the results;

(c) developing the **scoring rubric and rating scale(s) or checklist(s)** to generate the scores desired, with appropriate levels of content validity and precision;

(d) identifying the conditions, materials or equipment necessary for performance of tasks, and

(e) writing a set of **directions** to assessment administrators and raters/scorers to set up the assessment environment and engender the desired levels of accuracy and consistency.

The scoring rubric in Table 6.3 is **analytic**, and incorporates a **rating scale**. It is analytic because instructors could generate a separate score on examinees for each item and sub-domain, allowing instructors to tailor their teaching to skill profiles by sub-domain from the earliest stages of training, or as needed. This type of scoring protocol is ideally suited for formative decision-making, an intended use (see design specifications).

This particular analytic rubric also allows for **holistic scoring**, producing an overall score for the domain. That is, all the sub-domain ratings for an examinee could be summed to produce one score. While this is not what is known as a **holistic rubric** (treated later), it still yields a "total score".

See Table 6.3. As there are 25 items in all, with a maximum rating of 2 on each item, the maximum "total score" on the entire domain would be 50 (or 2 x 25). Say, an examinee's total score at the end of the training program, per the designer's specifications, is 20. As it is to be reported as a percentage of the maximum score possible, the final score is 40/50 or 80%. Those numbers could vary in a cohort of trainees. The percent scores, say 50-95%, is to be interpreted as "level of mastery".

Not also that two criteria help define the rating scale: competency and consistency. How was the 3-point rating scale defined, and what instructions did the observers need to follow? Here is an excerpted segment from the original source which defined the three levels on the rating scale.

Performs Competently and Consistently (Rating =2) means that during observation–

- a behavioral task, when performed, complies with standards and guidelines of the American Heart Association, American Physical Therapy Association's Standard of Practice Act and Code of Ethics and the Guide to Physical Therapy Practice of American Physical Therapy Association/Cardio-pulmonary section (2nd edition, 2001)
- a behavior has been observed often in a clinical setting (>70%)
- a clear pattern of learned behaviors has been seen
- the behavioral task was done independently
- only a few reminders or leading cues were given during task execution

Performs Competently but Inconsistently (Rating=1) means that during observation –

Student demonstrated this behavior in compliance with standards and guidelines but was inconsistent, or required help repeatedly to demonstrate the behavior consistently. A pattern of behavior is developing but not yet completely consistent (50-60% of the time).

Not observed or performs both inconsistently and incompetently (Rating=0) means that during observation -

- Fails to comply with 2+ standards and guidelines on two or more occasions of five
- a behavior has been observed only 5 or fewer times out of 10 in a clinical setting (<50%)
- an inconsistent pattern of a learned behaviors is seen over multiple observations
- frequent assistance, reminders and verbal cues are needed
- cannot demonstrate behavioral tasks frequently despite guidance or supervision

Some guidelines, along with some levels of rater training, are essential with performance assessments to control for accuracy and consistency levels during scoring. As trainees progressed through the curriculum, three levels of variability were expected on each item, leading to a distribution of percent mastery scores by sub-domain and domain.

Could the designers have selected an existing instrument instead? Indeed. Classroom teachers/instructors and researchers often rely on pre-existing assessments, constrained as they often are for time and resources. In ideal applications of the Process Model in such scenarios, one would start with design specifications as shown in Box 6.4 A-

B. Next, select the best-fitting items or available instruments that show the closest match, while making adaptations to suit local conditions or for meeting specific user needs.

Phase IV Validate the Items, Instrument and Construct Measures

The first iteration of assessment tool must undergo some level of checking, field-testing and/or validation before it is ready for use. In the illustrative case dealing with classroom assessment and program evaluation uses, the researcher began with an "internal" content validation. The domain specifications, the initial draft of the instrument, the scoring rubric and items were reviewed by classmates and the measurement faculty of the graduate level course in instrument design at a North American university in which the designer was enrolled. Following that, there was an "external" content validation of the tool that was performed by other clinical faculty at the Mount Sinai Medical Center. The "try-out" of the tool with two trainees helped standardize the instructions and scoring rubrics further.

Phase V. Evaluate Readiness of Tool for Use

Albeit conducted on a small scale, several issues surfaced as a result of the validation, for which revisions were necessary. These included a review of new standards for added items and indicators, alterations to the system of data collection and scoring, fine-tuning of the rating scale, and rewriting of standardized guidelines for instructors. The revised tool is shown in Table 6.3, and was released for use in the program, deemed ready for dissemination to all clinical instructors involved.

Reflection Break

- **Are the design specifications for the instrument adequate for the purposes? See Boxes 6.4 A-B. What items are missing, if any?**
- **Create a set of questions to guide a thorough content validation of a behavior-based assessment. Prioritize the questions. Which are the top three criteria, in your view, that will assure a high quality assessment tool?**
- **Why is a content validity evaluation a recommended part of applying the Process Model? Give three reasons.**

6.5.2 Designing a Product-based Assessment

This example is situated in a national project, which again involved designing classroom assessments tied to curriculum modules in health information technology (HIT). In 2009, the passage of the American Recovery and Reinvestment Act in the U.S. was accompanied by the Health Information Technology for Economic and Clinical Health initiative. The latter called for all hospitals and healthcare systems to begin maintaining and using Electronic Health Records (EHRs) for patients.

To help healthcare and information technology professionals implement EHRs meaningfully, the federal government funded a national effort to design workforce education programs. To this end, between 2010 and 2012, the Office of the National Coordinator for Heath Information Technology funded five Curriculum Development Centers. The goal of the centers was to design instructor-friendly curricular materials, instructional modules and assessments for training healthcare professionals in implementing EHRs. The courses were to be delivered through various universities or community college programs.

Tables 6.4-6.5 show selected materials from one of several modules developed at a northeastern university-based center under the above initiative. This particular module was intended for individuals who would be responsible for training doctors, nurses and other medical professionals on data use standards and patient privacy acts at hospitals. The Process Model was applied to guide the overall design processes for the curriculum and assessments (see Authors et al, 2018).

Phase I. Specify the assessment context

In this application, the first element of the assessment context is defined by a competency-based construct in health IT (What is to be measured?). The population comprised healthcare professionals that currently formed a part of the U.S. workforce

(Who?). The projected uses of assessment results were for the purposes of classroom assessment and related formative/summative decision-making by instructors. The inferences to be made from results or scores were at the individual trainee level, and dealt mainly with domain-referenced learning levels of trainees during, and after delivery of instruction (Why?).

Phase II. Specify the Assessment Operations

Table 6.4 shows the domain for the particular module, with the exiting performance outcome and targeted, taxonomic levels of three embedded indicators. In this instructional module, as seen, several assessments were multiple choice tests of knowledge level indicators in the domain. The culminating outcome in the domain, however, was a "product indicator", targeting a higher order level of performance. Specifically, the learning outcome required student-trainees to be able to show knowledge of data standards and guidelines, as well as, convince healthcare professionals to use EHRs in accordance with guidelines. This requirement was translated into an assessment task.

Phase III. Design the Assessment

The "capstone" task in the instructional module was a product-based assessment exercise, whereby students created "training materials" in the form of a persuasive statement on data standards that they could subsequently incorporate into their own training modules at work. Through the product they developed, trainees were required to show their knowledge and understanding of HIT data standards and legal requirements for protecting patients when using EHRs. They also needed to demonstrate their capacities to persuade the audience with specific examples of non-compliance or oversight, so that consequences for patients and hospital staff in charge were clear to participants.

Many variations of this task could have been designed to tap the culminating outcome in the domain specifications shown. This particular product-oriented task focused on six specific data standard topics. The task was untimed, and designed as a “take home” exercise. The directions for potential examinees were designed to help limit the boundaries of the task. For example, to encourage examinees in providing “on point” responses, the instructions set out clear content parameters and page limits. Directions given to both students and instructors were intended to help control errors during task administration and use. The potential users included all instructors in similar programs, only a handful of whom were involved in the design of assessments. For instructor-raters and examiners, task- and rubric-based guidelines and parameters were meant to allow for efficient sorting and separation of work products obtained into those that merited higher versus lower scores.

When rubrics tied to classroom assessments can be tightly aligned with domain specifications and classroom instruction, it raises content validity levels of results. Reviews of the research literature recommend sharing of scoring rubrics with students, right alongside the classroom assessment tasks. When such practices become a part of instructional routines, they have been shown to improve student learning, motivation and self-regulation levels. Such practices also help in communicating expectations of performance clearly and transparently to students and examinees, and can help contain error variability when summative decisions are made (Andrade & Cizek, 2010; Brookhart & Chen, 2014).

See Table 6.5. A **holistic scoring rubric** was designed for the task shown, that separated four, ordered performance levels, with scores ranging from 1-4. Note the differences between the holistic rubric in Table 6.5 and the analytic one shown earlier in Table 6.3.

Both are indicator-based and delineate the criteria of performance at different levels. But, while the descriptive statements at each performance level in Table 6.5 are linked to the indicators in the domain generally, indicators individually are not rated. Trainees receive scores depicting the overall quality of the product instead. Unlike their analytic counterparts, holistic rubrics do not permit diagnosis of specific weaknesses or learning needs. Because they provide a global summary of proficiency or performance in the domain, holistic rubrics are often applied for summative decisions.

Finally, examine the directions given to instructors who were to rate student-examinees using the rubric. Each rating category includes a series of descriptors that are attentive to the content, taxonomic level and conditions specified in the general outcome of the competency domain that is being measured. Such a focus also lends content validity to the rubric, improving how it is interpreted by users during use.

Phase IV Validate the Tasks and Construct Measures

Content-based validity is of the highest priority when designing classroom assessments (AERA, APA, & NCME, 2014; Author, 2003; Shepard, 2000). It involves a critical review process performed by experts so as to ensure that a match exists between the targeted learning outcomes and content covered on assessments.

Validation in the project began with selecting and training subject area experts and classroom level stakeholders in field settings to serve as teams of internal and external validators. These experts then conducted structured reviews to validate the content and overall quality of the assessment tasks, certifying that all the curriculum-based assessments were ready for dissemination to other educators. Two rounds of validation were performed, one by each team, followed by revisions to improve the assessments iteratively.

Phase V. Evaluate Readiness of Tool for Use

Key users of classroom assessments include teachers and students. In the project described, the validators' comments and quantitative ratings suggested positive perceptions of assessments and curricular products overall. During reviews, the content validators placed a high value on the following aspects of assessment quality: Content Relevance, Usability, Clarity, Connectivity with Curriculum, Link of Theory to Practice and Link to Workforce Needs. Following a third round of revisions, products were deemed ready for use in classroom contexts. Psychometric validation efforts lay beyond the scope of the project, but were recommended.

However, narrative commentaries revealed several specific areas that needed detail, correction, or clarification, requiring specific changes to the product-based assessment presented. The initially published version of the scoring rubric thus went through another modification. Tryouts on a small scale revealed that variance was limited to upper levels of the rubric because the initial version focused primarily on lower cognitive levels ("Describe" or "identify"). No one scored at Level 0-1 in the tryout with students! The difficulty level of the task and rubric was thereby raised to be more consistent with authentic training settings and variability found in responses.

Contrasting Rating Categories 1 and 4 of the Initial Rubric with the Final Version

Prior Rating category 4

Accurately describes ALL aspects below and relates each to the importance when implementing an EHR.

- Describes the different types of standards required (clinical data representation and medical terminology)
- Describes the importance of context regarding data standards
- Identifies the importance of grammar data standards for communicating in public health informatics
- Identifies patient rights under the Notice of Privacy Practice

- Identifies and summarizes each HIPAA security requirement (administrative, physical, technical)

Prior Rating category 0-1

- Describes 2 or less than 2 aspects below and relates aspects to the importance when implementing an EHR.
- • Describes the different types of standards required (clinical data representation and medical terminology)
- • Describes the importance of context regarding data standards
- • Identifies the importance of grammar data standards for communicating in public health informatics
- • Identifies patient rights under the Notice of Privacy Practice
- • Identifies and summarizes each HIPAA security requirement (administrative, physical, technical)

Reflection Break

Compare the earlier descriptors for the highest and lowest performance categories in the rubric given above, with the revised versions for the product-based assessment task in Tables 6.4- 6.5. What changes do you see, and are these useful improvements to the original, in your view? Explain.

6.5.3 Designing a Portfolio-based Assessment

The final example illustrates how the Process Model was applied to develop a portfolio-based assessment for school-going children of ages 5-7 in Florida in the 1990s, a period when portfolio-based assessment became a "buzz word" in the standards-based education reform movement (Arter & Spandel,1992). The state-funded effort, implemented in several school districts, was called Project CHILD. It reflects a number of the defining features of well-designed portfolios as an assessment modality in instructional contexts (Butler, 1997; Santos, 1997).

1. Portfolios require collections of work, products, or behavior samples that are systematically gathered and scored but could serve many different assessment purposes—instruction, assessment, evaluation of programs, or administrative.

2. They can be designed to provide documentation of the processes that candidates employ to develop any products that are included as work samples.
3. They are unique tools for chronicling growth over time in particular skill areas or domains.
4. In instructional contexts, portfolios can be easily integrated with instruction in the classroom, so that the assessment is not a one-time procedure, but an ongoing event.
5. Portfolios can be deliberately designed with involvement of persons assessed as well as, others relevant to the assessment context, whether they are students, professionals, parents or others involved in instructional, workplace, home or clinical settings.
6. Finally, portfolios facilitate ongoing metacognition and reflection by persons assessed and others relevant to the assessment process.

In Table 6.6, we see an excerpt of a language arts portfolio used in Project CHILD. In this illustration, the domain focuses on handwriting skills.

Phase I. Specify the assessment context

The specific sub-domain for which we see the materials in Table 6.6 deals with Handwriting Skills (What is to be measured?). At the time, the population comprised Grade K-2 students in public or private schools that were implementing the Project CHILD curriculum (Who?). The projected uses of assessment results were for the purposes of (a) classroom assessment and related formative/summative decision-making by instructors, as well as, for performing (b) ongoing curriculum evaluations. The inferences to be made from results dealt mainly with children’s growth in the delineated domain of skills (Why?).

Phase II. Specify the Assessment Operations

The domain is not shown but it specified the expected long-term learning outcome for students with several sub-domains of skills, as follows: Language expression skills (e.g., write simple, descriptive sentences, such as, games I like to play); Editing skills (e.g., check to see if you used a capital letter to start writing your name); and handwriting skills (e.g., draw lines with help from the teacher, print the alphabet in upper case, or write words and short sentences in cursive form). Guided by teachers, children also were expected to engage in self-reflection (e.g., review their portfolio and express verbally what they did well). The taxonomic levels integrated motor skills with targeted cognitive skills and concept knowledge. Each sub-domain was to be rated with a separate scoring rubric.

Students were involved in the assessment process by design. At the end of each semester, students selected what they thought were their two best work samples for a summative evaluation by teachers, who then assigned a grade. They also wrote a brief "letter" to their next teacher on what they did well and what they would like to learn in the following year. The last piece was a guided, metacognitive exercise led by the teacher.

Phase III. Design the Assessment Instrument

Several specific design-related decisions are unique to portfolios. The design of the Project CHILD portfolio-based tool required several clarifications with respect to the following.

- Who will participate in the assessment design procedures?
- Which work samples are to be included, what is the number and types of work to be sampled, when must these be gathered?
- How many rating protocols or rubrics should accompany the portfolio?
- Who will participate in the selection of work or behavior samples to be rated—only students, only teachers, or other?

- Who will rate the work or behavior samples, and when will the scoring occur to meet assessment needs of users?
- Under what conditions will the student samples be generated?
- How can we assure that the portfolio-based assessment is not a burdensome process for teachers, administrators or others?
- Is any kind of training or orientation necessary for teachers, assessment administrators or other users to ensure reliability?

The decisions in Project CHILD were made by teachers who were directly involved in the pilot program. The team reached a consensus that the portfolio was to be an integral part of their instructional tools for the overall curriculum. A separate rubric was deemed necessary for each skill area in the language arts portfolio, only one of which is illustrated next.

In Table 6.6, we see how records were kept by teachers to map handwriting development of children. Of all the writing exercises completed at the end of each semester, two student-selected writing samples were scored by the teachers. In this example, the handwriting rubric is a developmental checklist. It expects a development sequence of changes in handwriting skills of K-2 children. The binary scoring (observed or not), yielded a pattern of checks as each child developed in the area. Each check was given a point, and a total score as well as the profile of checks was maintained in the child’s portfolio going forward.

Phases IV-V Validation and Evaluating the Readiness of the Tool for Use

The guidelines for the portfolio and full set of scoring rubrics, once produced, were reviewed and content-validated by co-teachers from the participating school districts and an educational assessment specialist from a local university. All the materials were tested by teachers with their students and revised, as needed. As the intended uses were “low stakes”—or primarily to nurture growth and development of students in the early years of schooling, no further validation was deemed necessary.

Insert Table 6.6 about here

Reflection Break

- **Review the three performance assessments in Tables 6.4-6.6. Compare an analytic rubric with a holistic rubric with respect to:**
    - **How to design each**
    - **When to apply each**
    - **Advantages and disadvantages**

- **What are the advantages and disadvantages of a rating scale versus a checklist to score performance based assessments?**

6.6 Scoring Rubrics, Rating Scales and Checklists: More on How to Develop Them

A critical structural component of a performance assessment is the scoring rubric. The term, “rubric” refers the set of criteria and guidelines that designers specify to ensure that judgments of open-ended responses from subjects are accurate, consistent, and fair. Without rubrics, the assessment tool is only partially developed. Rubrics provide us with a means to make finer discriminations in performances that vary by degree of quality.

6.6.1 Analytic and Holistic Rubrics

In the analytic scoring system in Table 6.3 from a Physical Therapy skills domain, the response is broken down into relevant parts, and each part is assessed separately and assigned a separate score. Further, the score for different items may be weighted differently in an analytic rubric, as the Physical Therapy assessment also showed. The number of scores or ratings produced from an analytic rubric is equal to the number of parts we wish to separately assess. Often, an examinee or respondent may show a strong performance on some indicators of a construct, but much weaker performance on others. An analytic rubric will be able to pick up on such differences in performance by indicator or sub-domain.

In a holistic scoring system, in contrast, the response is judged more or less as a whole by the scorer/rater, resulting in an overall rating of its quality. A single score or rating is generated from a holistic rubric as shown in Table 6.5 for the health information technology domain. Holistic scoring is also called global scoring, because of its focus on the whole response rather than its parts.

It is possible to have a performance exercise that is scored both analytically and holistically. For example, consider an assessment requiring students to both develop and present a research report in social studies. The report could be scored analytically, while the oral presentation delivered in class may be given an overall holistic score. Decisions should be prompted by the assessment purposes and the substantive nature of the domain or tasks. Several large-scale assessment programs that incorporate performance assessment components opt for the more efficient, holistic approach to serve summative scoring needs. If the purpose is to diagnose strengths or weaknesses in pupils, or to facilitate formative decision-making, we would lean towards an analytic approach.

6.6.2 Rating Scales versus Checklists

A key decision in developing a rubric deals with the type of point-allocation scheme or "scale" we will use. Depending on the task and degrees of variability that can be reasonably expected in the responses, we could use either a checklist or a rating scale. Checklists have a binary point range. Rating scales can vary with 3-point, 4-point, 5-point, 7-point, 10-point or an even wider range of points on the scale.

A functional checklist must be linked to very specifically framed items tied to the domain. The items could define the observable criteria for acceptability of a product or performance. They could together identify some behavioral syndrome (say, for a disease). The items could also be arranged in a progressive sequence, defining criteria for development in a domain, as in the handwriting summary in Table 6.6. Checklists could also include items depicting an ideal order for carrying out a formal task or procedure.

For example, a research group is presently developing the Harvard Assessment of Anesthesia Resident Performance tool (Blum et al, 2017). The authors specified indicators for observing resident anesthesiologists' as they performed tasks in simulated, patient care settings. Consider the expected, "ideal" behaviors below. One domain is excerpted as an illustration (p.11, Appendix 2, with minor adaptations):

Domain Indicator: Physician Implements a Plan Based on Changing Conditions.

Items/Descriptors of Ideal Actions:

- Shows situational awareness
- Performs rapid and frequent re-assessments
- Is adaptable
- Prioritizes multiple tasks
- Manages flow
- Is decisive
- Manages time, personnel, and resources

During observation, each item above allows for a dichotomous classification of performance with a checklist because of the specificity. Checks may be totaled to obtain an overall score on all the items, as well/ Common checking categories could be: Yes = 1, No = 0; Present =1, Absent = 0; or Observed = 1, Not observed = 0.

Rating Scales, in contrast, are a set of ordered categories denoting different degrees of quality. Here is an example of a 4-point rating scale suitable for a behavior-based assessment similar to that in Table 6.3.

Not performed or omitted=0

Performed in partial compliance with professional standards, and fails to meet minimum expectations=1

Performed in compliance with professional standards and meets minimum expectations=2, Performed in full compliance with standards, and exceeds minimum expectations=3.
How do we decide on the range of the scale for allocating points? The best way is to try out the task designed with a small number of typical subjects to evaluate how well a rating scale performs with all the observed levels of responses.

To enable consistent and meaningful scoring by different raters or observers, it is necessary to define each scale point with clear descriptors that reflect how typical responses would actually vary. See the two student responses to an open ended graphing task from a science laboratory exercise in Figure 6.3 next. For the task, the following indicator is being assessed: “Plotted all X, Y coordinates accurately”.

A close reading of the questions vis-a-vis the two answers shows that Student A omitted one part of the task dealing with rationalizing the choice of graphing technique. Further, although the task is correctly executed by A, the graph breaks standard conventions as it is presented in two segments. Student B has a complete and correct answer, on the other hand, meeting all criteria in the domain.

Should both A and B receive the same rating if the rubric and rating scale overlooked possible omissions by students? Would that approach yield valid data on task mastery levels? Could rater subjectivity enter the scoring process, obstructing fair and consistent results? Such sources of error variability, once detected, must be evaluated with care. The information gained can then be used to recraft the rubric and descriptors tied to the scale points.

Ambiguous rating scale descriptors add potential measurement errors to the data, but are rather common in practice. Here are two rating scales for the task in Figure 6.3, with weak descriptors:

Rating scale: 0 = Poor, Satisfactory=2, Good =3, Excellent =4

Rating scale : 0 =Below Average 1 = Average 2= Above Average 3 = Outstanding

The problem in both is that broad terms like "Good", “Average” and "Excellent" are vague and connote different meanings to different raters. By carefully reviewing the domain, indicators that apply, the task, as well as, typical errors and omissions found in sample answers, we could concretize the rubric and rating scale descriptors further. The goal should be to reduce errors related to subjectivity in rater interpretation of rubrics.

Insert Figure 6.3 about here

6.6.3 Ten Guidelines for Rubric Design

1. Design a performance task or exercise linked to targeted indicators in the domain and overall assessment specifications.
2. Write the task instructions and prompts; set the conditions, listing all equipment or other materials necessary for task performance; provide instructions for subjects/examinees, assessors and others.
3. Specify whether an analytic or holistic scoring rubric will be used, or some combination of the two.
4. Specify the point-allocation scale to be used: Checklist, rating scale, or both?
5. Identify weights and point values to be allocated to different domains, items and task components in rubric, as applicable. This step is particularly relevant for analytic rubrics.
6. Develop a draft of a scoring rubric linked to the targeted indicators in domain. Write observable descriptors to define different levels of performance. Operationally define each scale point.
7. Try the task out on a sample of typical respondents. (Alternatively, perform the task yourself.) Observe or gather data on the possible range of responses to the task(s).
8. List common errors, omissions, or inaccuracies that you find in typical responses. Based on an error analysis, revise the descriptors of on your rubric and scale.
9. Check back to make sure that rubric matches the indicators originally specified in the domain, as this will ensure content-based validity of the results.

10. Select sample answers or videotaped performances at each score point (level of performance on the scale) to be shared as "anchors" with raters during scoring. This will enhance rater reliability.

6.5.4 Using Rubrics: Threats and Sources of Error

Human judges serve as an integral part of a performance assessment tool. Like all human beings, they are vulnerable to error during scoring. Random errors in scoring occur even when there are well designed rubrics to guide raters in their work. Large-scale assessment programs, as a result, invest in formal rater training programs and maintain "banks" of assessment raters who are likely to rate more consistently than others. In local assessment applications, we can take similar steps to ensure that human lapses in judging are not adverse influences on the ratings.

Based on research on rater errors mainly with essays, there are five common sources of rater error:

1. Halo effect
2. Item or task carry-over effect
3. Test or performance carry-over effect
4. Order effect
5. Writing/language mechanics effect

Halo Effect The *halo effect*, a kind of subjective bias, is evident in circumstances where the rater's impression of the respondent on characteristics unrelated to the performance affects the rating. This influence is usually in a positive direction, although the reverse could also be true. In an Olympic figure-skating championship, we would conclude that the halo effect was contaminating the assessment environment if we found that the U.S. judge on the panel was awarding higher ratings to the U.S. skaters, irrespective of their actual performance. Similarly, teachers are vulnerable to the halo effect whenever their favorable impression of particular students leads to correspondingly favorable ratings, regardless of performance. Conversely, a very poor impression of a student's behavior in the classroom, might lead to unnecessarily strictness in ratings on academic work. Controlling the halo effect is possible by keeping the scoring processes anonymous.

Item or Task Carry-Over Effects This error is caused by the rater's judgment of the quality of the first item/performance carrying over to the next item for that same examinee. In the Olympics figure skating championship, if a rater's judgment of an ice skater's performance on the short program influenced the judgment of that skater's performance on the following long program, we would have an instance of an *item* or *task carry-over effect*. In classroom assessments, if the rating on the first problem influenced the scores on the problems that followed next for a student, we would see another example of an item or task carry-over effect. To counter such effects, we should rate the same task for all performers or students before beginning to rate the next task.

Performance Carry-Over Effect Educational researchers have found that essays of poor quality tend to be rated much higher when they are rated after two badly written essays, than when the two preceding essays are well-written papers (Hales & Tokar, 1975; Hughes, Keeling, & Tuck, 1980). This type of contamination is called a *performance carry-over effect*, because the judged performance level of one examinee carries over and influences the score on the next examinee's paper. With behavior-based assessments, a performance carry-over effect would occur if judgments of the first person's performance colored the rater's judgment of the performers that immediately followed. To combat the effects of the test or performance carry-over effect, we should periodically shuffle papers in a random order, or randomly distribute observation sessions for individuals observed.

Order Effects The order in which papers are read or individuals observed has also been found to result in inconsistent rating patterns. Prolonged observation, recording, or scoring sessions can lead to a gradual decline in the quality of ratings on papers at the bottom of the pile. This "slide effect" occurs because of rater fatigue, resulting from the repetitive work. To reverse the impact of order, raters should take breaks to rejuvenate themselves during long scoring or observation sessions. Again, random observations of examinees can also help counterbalance the order or "slide" effect.

Writing, Speaking and Language Effects In performance assessments, better speaking or writing skills or better language usage often influences the score on the performance or product, even when the targeted domain is not related. Research shows that raters tend to be influenced by spelling, vocabulary, punctuation, grammar, length of the answer, neatness, or effective presentation skills of the performers. Test-wise students capitalize on this vulnerability of raters when they think they can "bluff" their way through essay exams.

The solution is to craft a tightly defined tasks and scoring rubrics that are closely linked to the assessment specification and indicators in the domain. Instructions to the assessment should help direct the examinee to the expected response. In the same vein, raters should also have explicit scoring criteria tied to the valued indicators from the domain. Rubrics and accompanying instructions should be designed to help raters maintain validity during scoring. Use of anchor responses/products are another useful way to keep scorers focused on expectations of performance.

Other Biases Several sources of systematic bias apply to all educational assessments, including performance assessments. These are: readability bias, where the reading and vocabulary levels of the materials used in the task directions or materials exclude some respondents from being able to participate in the assessment exercise; inflammatory bias, where particular words or examples inflame or disturb test-takers of some identify or demographic groups, adversely affecting their performance; and opportunity-to-learn bias, where a systematic lack of curricular exposure leads to poor performance in some groups.

Two additional sources of bias apply to performance assessments. The first, called a "conditions/materials bias, sterns from a mismatch of assessment materials or technologies with some subjects or examinees who may have learning barriers or physical disabilities. Individual raters being consistently strict or consistently lenient in their ratings could also be an issue. All of these errors tend to be systematic and predictable, thus affecting validity rather than reliability of results.

The answer lies in standardizing, field testing, validating and carefully designing the assessment materials and equipment to suit the population specifications. The developmental level, age, or special needs status of individuals must be taken into account from the start.

6.7 Summary

This chapter focused on three types of performance assessments, namely: behavior-based assessments, product-based assessments, and portfolio-based assessments. Each type was defined with several applied examples.

Performance assessment modalities are more than a simple change of format in that they allow assessment designers to (a) see a construct more deeply and from different vantage points, (b) measure situation-specific aspects of a construct, and (c) approach the design and scoring of assessments with a many-faceted mind set. The assessment methods treated in this chapter are useful for measuring a wide range of complex constructs that other assessment methods cannot measure well. With some performance assessments, specifying the context--in terms of the setting, materials and conditions of performance--helps elicit the targeted behaviors and responses under designer-controlled circumstances. Depending on the assessment purposes, populations and domains, particular settings could vary, ranging from schools, workplaces, clinics or elsewhere in the home or community.

To select the most appropriate performance modality for a construct, we must take into account the construct domain, the population and the assessment purposes overall. Five specific criteria to guide us are: Clarifying indicators that define the construct in terms of content, taxonomic levels, and conditions (i.e., situation-specificity); Identifying any process- and/or product-indicators in the domain; Matching indicators with best-fitting performance modalities; Evaluating the choice against population needs and assessment purposes; and Evaluating how well the assessment methods chosen will uphold validity, reliability and utility in contexts of assessment use.

A single performance modality may not be adequate for assessing complex and multidimensional construct domains. The chapter introduced a new taxonomy of performance levels given by Miller (1990) that was applied for designing assessments of competency-based domains in different fields. It recognizes four levels: Knows, Knows How, Shoes How, and Does.

Raters, observers or examiners are a key facet of all performance assessment systems. Three illustrative cases detailed how Phases I-V of the Process Model were applied to design a behavior-based, a product-based and a portfolio-based assessments, respectively. Issues related to task or exercise design, validation and use for each were discussed.

A necessary component of all performance assessments is the scoring rubric. The chapter concluded by discussing different types of scoring rubrics, how to develop these tied to different assessment scenarios, and errors related to their use. Distinctions between holistic and analytic rubrics were discussed, along with the advantages and disadvantages of checklists and rating scales. Types of random and systematic errors that designers must take into account were discussed. The construct, assessment purposes, and population parameters (Phase I, Process Model) should govern the degree of structure and standardization we need to build sound performance assessment systems.