Chapter 5

Designing Assessments with Structured-response and Constructed-response Items

5.1 Chapter Overview

The main focus of Chapter 5 is on how to craft items and tasks with highly structured responses or with response formats that are open-ended to different degrees. The item design formats highlighted in this chapter are best suited for tapping cognitive or proficiency-based constructs, and were introduced in Chapter 3 (see Table 3.5). As Chapter 4 made clear, for best results, item writing or selection should follow directly after specifying the indicators of the construct domain and sub-domains with sufficient clarity and specificity.

Specifically, the chapter addresses guidelines for developing multiple-choice, true-false, matching or completion (fill-in-the blank) items; context- or scenario-dependent item sets; and various constructed response tasks. Examples of good and bad items are provided, spanning levels from elementary to higher/professional education, and with a variety of examples from different disciplines.

Several existing textbooks and resources provide as many as 40 guidelines for constructing particular types of items (see for examples, Haladyna & Dowing, 1989; Haladyna, 2012, for multiple choice items). From such sources, this chapter extracts the most essential and useful rules that align with the latest know-how from the cognitive sciences on the how humans acquire, build and demonstrate knowledge and proficiency in different areas. Readers are encouraged to pursue more extensive treatments on these topics elsewhere, as needed.

Figure 5.1 highlights how this chapter connects with the rest of the book and the overall Process Model. Specifically, Phases II-III of the Process Model in black background are concerned with all the operations necessary to produce an assembled and finished instrument, starting with item design, the chief concern of Chapter 5.

Insert Figure 5.1 about here

5.2 Chapter Objectives

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

1. Evaluate the implications of the latest knowledge bases from the cognitive sciences for designing effective items and tasks in cognitive domains with structured or constructed response formats.
2. Design or select items with formats best suited for measuring cognitive outcome- and process-indicators, as defined in construct domains.
3. Apply/follow established guidelines for writing different types of structured response items with scoring keys: fill-in-the blank/completion, true/false, matching, multiple choice, and scenario-dependent items.
4. Apply/follow established guidelines for writing different types of constructed response and essay items, supported with appropriate scoring rubrics.
5. Conduct a content validation and critical review to improve the quality of items designed/selected for the specified domain, inferential needs and assessment purposes.
6. Apply relevant segments of the Process Model and assessment design

specifications to design/select items for compiling a test or assessment.

5.3 Why Item Construction Details Matter

To begin, consider two snippets of conversations from the *Peanuts* cartoon strip by Charles Schulz below (1968, United Feature Syndicate, Inc.; 1975, United Feature Syndicate, Inc., excerpted from Mehrens and Lehman, 1984, pp. 141 and 106, respectively). The quoted text is from the original cartoon strips.

Episode 1

Lucy: "*How did you do on your test?*"

Linus: "*Don't ask me...it was a disaster...*"

Lucy: "*Couldn't you even pass a true or false test? What happened?*"

Linus: "*I falsed when I should have trued!*"
....

Episode 2

Patty (Reads title of an examination in school.): *"History Test".*

(Reads first line of question): "*Explain World War II.*"

(Reads second line): "*Use both sides of the paper if necessary.*"

....

Dealing with the mechanics of item construction might seem like "much ado about nothing" until one appreciates how easily measurement errors can be introduced through the "How to Assess" design portal of the Process Model (See Figure 5.1, Phases II-III). Unlike

Linus' caricatured test-taking experience above, a good true-false item should yield sound information on the underlying construct indicator(s) in the cleanest, and most straight-forward, way possible. Test-takers should not feel that they are tricked into "falsing"—or responding in a way that inadequately reflected what they actually knew or could do, were unable to control, or with which they felt uncomfortable, due to some inherent properties of the items. Should that turn out to be the case, there may well be work left for the test and item designer to do!

The same principle applies to constructing open-ended tasks. Loose, undefined and ambiguous essay prompts--such as, "Explain World War II"--leave too many options open for respondents. Such unrestrained prompts invite highly variable and sometimes confused responses, with elevated probabilities for error, some of which may enter through examiners and examinees themselves.

The goal of the item designer should be to elicit the most valid and reliable response data possible from the respondents/examinees. Simultaneously, good item-writers must also bear in mind the current knowledge bases from the cognitive sciences that suggest how humans think, learn and develop proficiencies in given domains. The cognitive sciences are a diverse set of disciplines including the neurosciences, computer sciences, and cognitive and educational psychology.

Reflection Break

- Think of particular structured response items you have encountered as a test-taker. What interesting attributes, flaws or strengths did you find in these?
- Think of open-ended assessments from your test-taking experience? What interesting features stood out in these?
- Which item formats best captured what you knew or could do?

5.4 The Human Brain and Implications for Item Design

Most of the item formats we examine in this chapter will appear traditional on the surface (e.g., multiple choice, essay). But, new knowledge in hand compels us to rethink and retool these old item formats in innovative ways so as to design more cognitively-informed assessments. We begin this section by recapping some specific implications of this new information for item and task design in cognitive domains. Following a review of the main tenets and concepts, Table 5.1 and Boxes 5.1-5.4 allow us to compare and contrast the merits of old versus newer item design approaches using the basic multiple choice, true false, completion and short answer formats. In the illustrations the focus is on a mathematics domain for Grades 3-5 learners: Long Division.

Research from the cognitive sciences and brain science literature suggests that human **cognitions** (what individuals know about a domain--say, a math topic like long division) and related **emotions** (whether they feel confident or anxious when doing math) are not innate, fixed attributes that are genetically wired into the human brain from birth. Rather, over time, humans develop in various cognitive and non-cognitive domains. Such learning is constructed and re-constructed through numerous life experiences, and the brain's structure continually changes as a result of new learning. Indeed, humans learn, build and rebuild cognitions and emotions throughout their lifetimes (Barrett, 2017; Buehl, Alexander & Murphy, 2002; Greeno et al, 1993; Pellegrino, 2002; Pellegrino et al, 2014; NRC, 2012; NRC, 2013; NRC, 2001; Seaton, Parker, Marsh, Craven & Yeung, 2014; Stiles, 2000; Usher & Pajares, 2008; see also, Bandura, 1997; 2006).

Once we appreciate the above, the ramifications for item design are six-fold. These deal with the notions of: knowledge structures in the human brain, metacognition, contextually-situated cognition and learning, transfer of learning, varied pathways to expertise development, and reciprocity of cognitions and emotions. Each has implications for how we envisage the domain and the most apposite item features (after Barrett, 2017; Pellegrino, 2002; NRC, 2001; NRC, 2012; 2-13).

5.4.1. How we Construct Cognitions and Emotions

"Knowledge Structures" in the Human Brain. First, we know today that the human brain's architecture is made up of both the short-term, working memory and the long-term memory. We store most of our organized knowledge about different objects, experiences, people or phenomena in the long-term memory.

With every learning experience, new information is processed, organized in our brains, and retrieved, as needed, using neural networks. Naïve understandings of a child are built on intuitively-gained knowledge structures; these change as new learning experiences accumulate, leading to assimilation, adaptation, and modification towards more formalized knowledge structures in the brain. When faced with new situations or problems, our level of success at adapting to new environments and in finding solutions to problems is contingent on how well we can *retrieve and use* appropriate information from our long term memory.

According to two National Research Council committees (NRC, 2001; Pellegrino et al, 2014), assessments should therefore hone in on mapping the development of individuals' knowledge structures in specific domains and in different contexts of learning. For example, use of **probing questions** can help make visible where a person's knowledge levels and thinking lie on a topic. Such items can also help reveal deeper levels of understanding, as well as, gaps in knowledge and misconceptions.

"Metacognition". Second, people have the capacities to reflect on their own learning processes, including their problem-solving successes, missteps and failures in a domain. They can thereby self-regulate and correct course to reach an end goal. This capacity is called **metacognition**. Tasks that help assess differences in metacognitive skills in a given domain will enable differentiation among persons at varying levels of expertise in an area, ranging from novices to experts on a continuum of development. Experts engage in greater levels of metacognition and self-correction than do novices.

"Contextually-situated Cognition and Learning". Third, much of what humans know, learn and can do is through interactions with others in particular social and cultural contexts, including the classroom, community and family. Deeper understandings of the world develop *in context*, and the meanings particular things carry for individuals may be different in new contexts. Therefore, assessment tasks should situate knowledge-testing in contexts applicable to given domains, with attention to the meaningfulness of those contexts for examinees. Traditionally-designed items often test decontextualized pieces of information. By contrast, cognitively-informed items contextualize such assessments in situations to which the targeted population can relate.

"Transfer of Learning". Fifth, individual capacities to transfer knowledge and skills can vary from situation to situation. Transfer is not automatic. However, transfer of knowledge to new situations can be facilitated through "scaffolds" whereby prior knowledge and new information are connected through deliberate learning experiences and opportunities. Scaffolding could occur in formal schooling contexts through explicit teacher guidance, or through informal experiences and incidental situations outside school.

Assessment designs that embed scaffolding techniques make use of items that *teach as they test*. Such items could help guide cognitive development and promote transfer of knowledge and skills to more complex problems situated in alternative settings.

No "One-Size-Fits-All" Way to Learn and Build Expertise. For different people, the brain's structures may develop through different pathways, and at different paces in different domains. Assessment design should attend to factors that influence learning and building of expertise in an area.

To grow, individuals need chances to learn from mistakes, coupled with opportunities to consolidate new learning. Research shows that learning and development in an area can be shaped by identifying specific learning needs and following up with focused feedback, timely instructional interventions, and targeted practice. Strategies that people practice become more automatic and efficient with time.

Human Cognitions and Emotions. Non-cognitive development, having to do with our feelings, beliefs, and dispositions, occurs simultaneously with cognitive development in a domain. If we feel good about what we already know in a domain like mathematics, we will likely choose to undertake more difficult tasks without fears of failure, and will also be more likely to succeed in progressively more difficult tasks in that area. And, as we gain greater mastery of a domain, reciprocally, our attitudes and self-beliefs in our capacities for future success also increase.

The reciprocity principle has implications for the culture we create for assessment design and use. Cognitively-informed assessment design approaches, in contrast to traditional approaches, utilize continuous progress designs that acknowledge the reciprocity of cognitions and emotions. Such designs may foster positive feelings in examinees as they improve and gain mastery in given domains.

In recommending what Next Generation Science Assessments (NGSS) should look like in K-12 education settings, a 2014 National Council Report strongly endorsed the cognitively-based assessment principles just discussed (excerpted from Pellegrino et al, 2014, Conclusion 4-1, p. 3, Summary). In their view, soundly designed science items would:

- Integrate science learning across disciplines, or "....include multiple components that reflect the connected use of different scientific practices in the context of interconnected disciplinary ideas and crosscutting concepts";
- Map "learning progressions", or "...address the progressive nature of learning by providing information about where students fall on a continuum between expected beginning and ending points in a given unit or grade"; and
- Be instructionally-useful, or "...include an interpretive system for evaluating a range of student products that are specific enough to be useful for helping teachers understand the range of student responses and provide tools for helping teachers decide on next steps in instruction" .

The committee recognized two approaches--namely, evidence-centered assessment design and construct modeling--as incorporating the fundamentals of cognitive research and theory (see Mislevy et al, 2006; see also, Bennett, 2011; Stevens, et al, 2011).

Reflection Break

- Traditional test results often communicate a sense of permanence about one's assessed mental abilities in a domain (i.e., you're either good or bad at something). Which cognitive science ideas do you find most useful in countering this mind-set? Explain.

- To apply the cognitive science ideas you find most useful, what would you do differently when:
    - **specifying the domain** for constructs you measure?
    - **creating items or tasks**?

5.4.2 Traditional versus Cognitively–based Item Design Strategies

Now, examine Boxes 5.1-5.4 and Table 5.1. These provide contrasting illustrations of item design strategies in long division, showing how cognitively-informed approaches could differ from traditional item design approaches.

Every primary school student is expected to demonstrate proficiency in long division by the time they reach middle school, typically, Grade 6. Teachers expect that student abilities in long division are automatic and efficient by the time students begin to work on fractions, decimals, or manipulating algebraic equations that call for more complex arithmetic operations.

The tasks in Boxes 5.1-5.4 were designed by a team of researchers and teachers as a part of 2-year research and development project, guided by such educational expectations. The project was supported by the National Science Foundation and designed to detect and close students' learning gaps in selected mathematics units, using a teacher-mediated, diagnostic teaching and assessment technique. The assessment design exercises drew on the cognitive science principles just discussed, supplemented with theory on classroom assessment and pedagogy (Chatterji et al., 2009b;Chatterji, 2012).

First, examine Table 5.1. It presents a cognitively-informed domain in long division. Outcome 1.0 states that, at the end of the unit, all students must demonstrate the ability to solve "real life" problems using long division and other necessary operations. They must also be able to explain their solutions in their own words. Why and how is the domain "cognitively-informed”? First, it recognizes that students will vary in how they arrive at the culminating target. Second, the embedded indicators were revised and updated based on an analysis of typical errors students made *en route* to mastering long division skills. To begin, the team employed a “backwards design” procedure, starting with the culminating outcome that was targeted for students by participating teachers; next, they made several iterations to the indicators as teaching and assessment processes got underway, using student responses gathered from classrooms to elucidate the pathways to mastery (Chatterji, 2003,pp 124-140); see Wiggins & McTighe, 2005 for more on the “design downwards” strategy).

Table 5.1 about here

Chapter 4, Table 4.3 provided the cognitive taxonomy used to classify indicators and sub-indicators in the long division domain in Table 5.1. In terms of the taxonomic levels, we see a range of embedded indicators—the competencies and sub-competencies--from basic recall and understanding, to application, to complex procedural skills and on the end-outcome at a higher order thinking level.

In Boxes 5.1-5.2, we see both traditional multiple choice items, contrasted with a cognitively-informed item series. Both item sets tap into different portions of the domain in Table 5.1. What are the key differences in item design features?

Traditionally-conceived items, whether of a structured response or constructed response format, tend to:

- Focus on measuring the *products of learning*—that is, the outcome or anticipated result of the learning process. In this case, that desired outcome is proficiency in long division computation and interpretation, presented alone or as a part of scenarios. There is little interest in assessing the underlying thinking or learning processes involved, or to uncovering any existing learning gaps in examinees that could be addressed through instruction along the way.

- Make the implicit assumption that all examinees will “know” or have mastered all the underlying knowledge and skills when they are tested. Items tap into outcomes that are expected for a particular stage or grade level. In the illustrative case, it is grade 4.
- Place an emphasis on whether the answers are right or wrong, typically using binary scoring mechanisms. Item design is geared toward summary inferences (pass/fail), and summative decision-making.
- Define item quality based on (a) the match of the items with the targeted content and cognitive levels in outcomes, and (b) application of item-writing guidelines applicable to the selected multiple choice format, discussed in a later section. In the illustrative case, the two traditionally-designed items match the main computational indicators and the expected outcome quite satisfactorily.

By contrast, the cognitively-informed structured response items differ on several counts, and tend to:

- View examinees as learners at different points on a continuum towards building proficiency in a domain. In the example, a developmental range of the population is identified from grades 3-5.
- Use *item series* to probe into deeper understandings, reveal misconceptions and errors along a learning continuum. Item sets aim to map knowledge structures of examinees as they progress towards higher levels of proficiency in the domain.

- Use both structured response and constructed response items to effectively tap into both *products* and *processes* of learning.
- Use scoring methods that yield detailed diagnostic information on domain-referenced profiles of indicators, as well as right/wrong summary scores. See the analytic checklist for student self-assessments, for an example of a diagnostic assessment.
- Design items tied to specified indicators in the domain, just as in traditionally-designed items, but ensure that domains define continua of expertise-building more broadly and deeply than with indicators emphasizing outcomes alone.

Importantly, *domains are specified differently* in the two different approaches to item design. In the research project, the “black box” of the long division domain was opened up to reveal the implicit learning processes *and* products that students needed to reach the final outcome in Table 5.1. Instead of breaking down a domain with indicators stressing products of learning in the main, cognitively-informed domains are conceived with fine-grained elements that assist in formative diagnosis, feedback, instruction, and practice leading towards higher proficiency levels (see also, NRC, 2001; 2012; Pellegrino et al., 2014)

Boxes 5.1 - 5.4 about here

Continuing a review of the items in Boxes 5.1 -5.4 in tandem with Table 5.1 now, we see that while real world “story” problems may be used in both traditional or cognitively-informed approaches to item design, the better items would be situated in real world scenarios likely to be developmentally appropriate and thus, carry meaning for the population.

Boxes 5.2 and 5.4 demonstrate how metacognitive skill-building might be built into item design. The item series in Box 5.2 is intended to facilitate error analysis, self-correction, and reasoning by students. In Box 5.4, we see an accompanying, analytic scoring checklist for guiding students in domain-referenced self-assessments. This was intended to facilitate metacognitive skill-building as well as continuous development. At the bottom of Box 5.4, three non-cognitive indicators encourage particular habits of mind and dispositions in relation to mastering long division, acknowledging the reciprocal nature in which cognitive and emotional development occurs.

Reflection Break

- Should traditional approaches to item and test design in cognitive domains be discarded, in light of the latest research from the cognitive sciences?
- Which item design approach, cognitively-informed or traditional, would be better for an instrument you might like to design/use? Why?
- Distinguish between indicators that are “learning products” and “learning processes”. Specify each in a cognitive or proficiency-based domain of your choice.

5.4.3 Which Item Design Approach is Better?

There are diverse perspectives on the place and value of traditionally-designed items today, and given the new information from the cognitive sciences, the key question is: what kind of “test” is the ideal for cognitive and proficiency constructs ? (Mislevy, 2006; Gordon et al, 2014). This book takes the position that the **assessment purposes—**intended inferences and uses of assessments and the data these generate**--** should guide the **item design approaches** we adopt. A fundamental principle of engineering design applies to assessment design using the Process Model, as well: function should precede form (Roozenburg & Eekels,1995).

When well-designed and tied to underlying domains, traditionally crafted items continue to yield useful information for particular interpretive needs and uses. Typically, these needs are for summative and point-in-time decisions. Such items are still prevalent in a vast number of group-testing instruments and individualized assessments that support widespread applications in education, psychology and professional fields.

Cognitively-informed item design formats, in contrast, aid in formative decision-making and fine-grained score interpretations by users along developmental continua. As shown, these techniques have high utility in supporting diagnosis, feedback, and instructional needs of educators. In in health, mental health or therapeutic contexts, such item designs might be similarly useful for mapping progress in functioning levels of patients who may suffer from cognitive impairments. Assessments used to track rehabilitation following injuries that affect the brain, for example, could benefit from developmental and dynamic item designs that are cognitively-based.

5.5 Constructs Suited to Structured- or Constructed-Response Items

While not every aspect of all cognitive constructs will be best tapped with the item formats treated in this chapter, several might be. Before we examine item-writing mechanics, let's recap the common cognitive or proficiency-based constructs that we will likely to encounter when undertaking item design/selection efforts pertinent to this chapter.

Achievement and Learning. In classroom teaching-learning contexts, **achievement** is a commonly assessed "cognitive" construct, measurable with the items formats illustrated in this chapter**.** Achievement domains are comprised of targeted learning outcomes—the expected products of learning-- specifying the knowledge and skills that students are expected to master after exposure to a formal curriculum or course of study. Outside the students themselves, teachers and educators are the primary users of results, with the information is typically applied for making summative educational decisions.

Achievement, the more traditional notion, is now clearly differentiated from **learning** and **learning progressions** in the educational assessment literature. Learning, as we saw, deals with the active construction of knowledge in a domain. Learning progressions in given domains can be mapped developmentally between well-defined beginning and end points for any population. Also, learning occurs in all social or cultural contexts; as such, it could happen anywhere and at any time, whether in the classroom, home, workplace or community. Further, learning is distinct from growth in a domain that is attributable to biological maturation processes.

Through assessments, we will be able to capture *only* the observable, performance-based information on both learning- and achievement-based constructs. However, the inferences we draw from the results would be different in each case, assuming we are able to implement effective item design techniques.

The common inference drawn from assessment data on achievement is about the examinees' proficiency levels in a domain at the end of a unit of instruction. In inferential terms, achievement is defined as the performance dimension of desired learning outcomes expected to follow from formal instruction (after AERA, APA & NCME, 2014; see also Cizek, 1997; Good, 1973). In contrast, the common inferences to be drawn from assessment data on learning are about the examinees' learning needs, learning processes and learning progressions along a continuum.

Job Competence. In workplace assessment contexts, man**y job competency** domains are also specified in terms of demonstrable knowledge and skills pertinent to successful performance on the job. These are also measurable with structured or constructed response items dealt with in this chapter. Unlike achievement, job competency domains are not tied directly to an educational curriculum, but to valued aspects of job-related performances in given occupational or professional settings.

Job competency domains can be designed for point-in-time, summative decisions, as with achievement constructs. Similarly, we could also think of assessing domains of workplace learning along continua, as in cognitively-informed assessments for use in on-the-job coaching and professional development environments.

Aptitude. Aptitude, another cognitive construct measurable with item types discussed here, is defined as a person's *future potential* to succeed in a career or program of education based on current assessments of his/her knowledge and skills in a domain (Cohen & Swerdlik, 2010; AERA, APA, & NCME, 2014). As such, outside reflecting some learned body of knowledge and skills , results of **aptitude** assessments must also allow for prognostic inferences about examinees' future successes in some criterion domain. In Chapter 4, we saw that Graduate Record Examination (GRE) is a well-known aptitude test, tapping verbal, quantitative and analytic abilities of students at the end of their undergraduate training. Its main purpose, however, is for predicting their future success levels in graduate degree programs, the criterion.

Intelligence. Intelligence is another construct that is tapped with item types highlighted in this chapter. While definitions of, and theories on, the concept of **intelligence** are continually changing, an omnibus definition holds that it is a multi-faceted set of mental capacities that manifests itself in different ways across the life span of individuals. In a general sense, it refers to the capacity to gain and apply knowledge, to reason effectively, to think logically, to have sound judgement, to have facility with words, and to be a mentally alert, perceptive, problem-solver who can adapt to new situations (after Cohen & Swerdlik, 2010). Gardner (2008) proposed an alternate theory of multiple intelligences that is different, but that not everyone has accepted (Sternberg, 1984). Dimensions of human intelligence, once agreed upon, may be viewed on continua and assessed accordingly.

5.5 Typical Item Design Parameters

Regardless of the substantive differences in constructs, when tapping a cognitive or proficiency-based construct, there are two design parameters that would apply. The first deals with eliciting and obtaining evidence on the maximum or *best* possible performance of examinees. That is, the goal of the item- and task-designer should be to apprehend the maximum capabilities of individuals tied to a domain, rather than what may be typical or less than par performance (Hopkins, 1998). Therefore, during item design, we must orchestrate the conditions of assessment to attend to this parameter.

The second design parameter has to do with the medium of administration. Most cognitive or proficiency-based assessments are commonly written or **paper and pencil** tests. Today, such assessments can be delivered via computers or other technology based media with highly structured or less-structured tasks (See item examples from the Next Generation Science Assessment Project; http://nextgenscienceassessment.org/, retrieved 11/30. 2017).

Note that maximum performance is well-tapped with binary-scored items that are scored as correct (1) versus incorrect (0). This applies to selected-response items or fill-in-blanks with a word or brief phrase. True–false questions are another common example from this category.

Constructed response tasks that allow examinees/respondents to reveal their depth and breadth of knowledge on a topic, or their capacity to think, apply what they know, problem-solve and reason in a domain, should also tap maximum performance. When checked or rated with scoring rubrics that are tied to the domain (see the example in Box 5.4), constructed response items help in differentiating among degrees of correctness/reasonableness in the answers. Essay examinations, referenced in the second *Peanuts* snippet, are another item format suited to these ends. As seen in the illustrations in Boxes 5.1-5.4, various combinations of short answer and structured response items are frequently useful for meeting assessment design purposes.

5.6 Guidelines for Writing Structured Response Items

This section now treats the standard guidelines for each of the structured item formats--multiple choice, true-false, completion, matching and complex interpretive exercises--in turn. Refer to Table 5.2, as we proceed in this section . For all the highly structured item formats discussed, five guidelines are common:

(1) Each item must be designed to the match the specifications of the selected outcomes or indicators in a domain, in terms of the content, cognitive level and conditions to be measured;

(2) Items must elicit the desired performance in as direct a manner as possible; hence, clear directions and presentation are critical;

(3) Each item should ideally deal with a single concept, principle or issue with one correct or best answer;

(4) Items or item sets should not provide clues, grammatical or otherwise, to the respondent; and lastly but importantly, (6) items should be carefully screened to be free of biases.

Because structured response questions have built-in format restrictions, they are best able to test one concept or principle at a time. Confusing respondents with too much extraneous information with “tricky” items are not desirable; hence item directions and presentation must be clear, succinct and on point. Likewise, there has to be consensus among knowledgeable experts that there is only one correct or best answer to each question. This may be verified by field-testing the items with appropriate reviewers, who could take the test themselves while performing the validation and critique.

Biases can be introduced through the subject matter or materials incorporated in an item or exercise. For example, bias towards particular subgroups could occur if there is a preponderance of item scenarios reflecting only one region of the world, disadvantaging segments of respondents in fully grasping what the item asks. Or, items may be presented via media or technologies to which only some examinees have access or exposure. Bias towards particular examinee groups can also be in the form of **inflammatory biases**, as in inadvertent use of language that undermines or offends a gender or religious group, affecting their performance. There can also be **opportunity to learn** biases, where the content of the items is new to significant segments of the examinee population as compared to others**.** For all these reasons, screening for potential biases should be a part and parcel of item design procedures**.**

Table 5.2 about here

5.6.1 Multiple Choice Items

The multiple choice format is the most widely used selected response item type. In its basic form, it consists of a “stem”, which presents the problem or question, followed by a number of response choices, of which only one is the best or correct answer. Incorrect response options are called “distracters” or “foils*”*. The number of answer options usually varies from three to five, with the probability of simply guessing the answer by chance decreasing from 33% to 20% with increased number of alternatives.

The basic multiple choice format uses an incomplete statement or a question in the stem, and can be enhanced with several modifications. Consider the examples in Boxes 5.5 and 5.6 that test a targeted learning outcome from a basic course in statistics**.**

Boxes 5.5 and 5.6 about here

The first is the standard format; the second item uses the analogy form. Another notable variation is a context-dependent item set, in which a reading passage, scenario, data table, and/or graph precedes a series of multiple choice questions. The examinee is required to use the information to determine the correct responses.

The two illustrative items in Boxes 5.5-5.6 tap recall and understanding levels. But, the popular perception (and one that is well-supported by common practice) that structured response items can measure only lower levels of cognition can be dispelled with skilled and creative use of the multiple choice format. Multiple choice items are, in fact, more versatile than the other formats in their ability to tap a wide variety of cognitive levels, ranging from simple recall to more in-depth interpretation, application, complex generalizations, or problem-solving skills. Particularly, context-dependent item sets lend themselves to assessing multiple cognitive levels.

The next illustrative item set in Box 5.7 shows this feature. Items here are excerpted from an examination designed to test resident physicians' abilities to extract and interpret statistical information from published research studies to improve patient care. The domain is called "evidence-based medicine" (EBM) (Wyer, 2008).

As evident from items in Box 5.7, a competent examinee must know the definitions of the concepts of “risk”, be able to calculate the risks versus “relative risk reduction” levels by comparing two treatment options, and then make extended generalizations from that information. Collectively, these indicators fall under the targeted outcome specified at the higher order thinking level. Physicians who are more able in the domain have the capacities to gauge the effectiveness of different medical therapies based on their reading of the scientific literature. They are then able to bring that information to bear when making clinical decisions concerning patients. Table 5.3, which will be discussed in a later section, presents the design specifications for compiling a multiple choice test covering a wide range of cognitive levels across content dimensions of the EBM domain (Wyer, 2008).

The disadvantages of multiple choice items are that even competent item writers may find it difficult to design good items with one indisputably correct or best answer. The format may be unsuitable for examinee populations who think more deeply, are able to find alternative but reasonable solutions that lie outside the answers provided, and where the subject matter is too complex to allow for singular “correct” answers.

Item-writing guidelines can help. In Table 5.2, Section B provides a list of criteria that could collectively assist in reviewing and improving multiple choice items as a part of the content validation phase when designing assessments (Figure 1.6, Chapter 1).

Reflection Break

Select 2-10 multiple choice item examples presented in the Boxes 5.1-5.7.

- Evaluate the quality of the items using the general criteria in Table 5.2, Section A. What issues did you identify? Which would you revise and how? Explain your reasons.
- Now evaluate the quality of the items using the criteria using Table 5.2, Section B. Which would you revise and how? Explain your reasons.

Critically Reviewing Multiple Choice Items. Table 5.2 offers criteria to distinguish between good and poor item-writing. A criterion for claiming content-based validity is a match of any item with the content, conditions, and the cognitive level specified in the targeted outcome and embedded indicators. For example, in Table 5.7, the items would fail to meet the *condition* criterion if questions on relative risk were presented without a data table. If the questions covered statistical concepts that fell outside risk assessments, the *content* criterion would not be met. If the *cognitive demands* of the items were merely recall and understanding, the items would fail to meet the taxonomic levels in the targeted outcome.

Implausible options are usually answers unrelated to the topic in the outcome measured, or that are clearly unreasonable answers. These provide opportunities for the test-wise student to use processes of elimination to discern the correct answer. A plausible response, in contrast, is one that will look like the right answer to examinees with partial knowledge of the material, or to those that may have some conceptual misunderstandings.

Consider the following item on mathematical patterns. Note that the *verbal load is in the stem*, a desirable characteristic that helps the designer frame the problem better for the examinee to respond.

> Look at the increasing series of numbers below. Figure out the rule for the number pattern and fill in the next three blanks in the series.
> 12, 60, 300, ______, _______, ________. The next 3 numbers are:
>
> A. 12, 60, 300

B. 348, 396, 444
C. 1500, 7500, 35500
D. 1500, 7500, 37500
E. All of the Above

There are other desirable item properties here. The correct response (option D) is balanced by plausible distracters that reflect common errors made by elementary grade students. For example, not being able to distinguish between repeating and increasing patterns, (A), or choosing the wrong operations to solve the problem (subtracting 12 from 60, to get 48 and adding it to generate next numbers in B, or making computation errors, as in C) are plausible.

But, an implausible response option is E, *All of the above* as it cannot be correct. Along with *None of the Above*, this answer option should be used with thoughtful judgment to improve measurement, rather than simply because item-writers cannot think of a suitable distracter. Notice also that the distracters are on similar levels of difficulty, with parallel form, length and construction to avoid inadvertent clues that help examinees rule out wrong answers. A common item-writing error is where the correct answer is always the longest one! A poor set of distracters for the same item, follows.

Look at the <u>increasing series</u> of numbers below. Figure out the rule for the number pattern and fill in the next three blanks in the series.
12, 60, 300, _____, ______, _______. The next numbers in the series are:

A. 7500
B. 5
C. 348
D. 1500, 7500, 37500
E. All of the Above

The distracters above could give away the answer because (a) they are logically or semantically inconsistent with the stem, which states there should be *three* numbers; (b) are not in a sensible order—ascending, descending or alphabetical, as may be applicable; (c) use mixed digits,5, 348, 37500; or (d) are simply unreasonable in other ways.

The longest option, and only one with three numbers, happens to be the correct answer. All of these features offer possible clues to examinees, making the item a poor one.

On the other hand, an examinee who chose 5 (option B) might show partial understanding towards mastery of patterns problem-solving, as this choice would suggest an ability to identify the correct multiplier to identify the pattern. However, that option might be underused by respondents due to the obvious clue to the correct answer, with a loss of valuable information for test users. All of this calls for a better designed item.

Negative language should be underscored or highlighted in all items, and double negatives avoided as these often create ambiguity. See the item example that follows, where the correct answer in bold is C. Option A shows a repeating pattern; B shows an increasing pattern by a factor of 4; C is not a discernable pattern, and hence wrong. Note that D. *All of the above* is a plausible response here, and it may help in differentiating learners who have grasped the concepts to greater or lesser degrees.

Which of the following is NOT a number pattern?
A. 1,2,3, 1,2,3
B. 5, 20, 80, 320
C. 1500, 7677, 9599
D. All of the above are number patterns. Explain:

Finally, multiple choice items should present options vertically to improve clarity in presentation, readability and avoid confusions in examinees. In a compiled set of items, the correct answer position should be varied in some random order that is not predictable. Such patterns become clues.

Clues can be introduced through various sources and should be avoided. Providing an article like “an” in the stem with a correct answer like, “apple” among the options, for example, would be a grammatical clue that one might find on a reading or science test item for young children.

Last, but not least, any inflammatory or other biases towards different cultural, gender, religious, disabled or other groups can enter an item without intent. These should be removed through careful reviews of the language, presentation format, graphics and overall content of items.

5.6.2 True-False Items

A typical true/false item consists of a propositional statement that the examinee must affirm or negate with a *yes* or *no*, a *true* or *false*, or a *right* or *wrong* response. The statement can be posed in question form or as a direct statement. Typically, one finds a series of 5 to 15 true/false propositions clustered on a test--a design strategy that helps in providing common directions for the entire item set.

True/false (T/F) items can be modified to include an explanation or justification component, where the examinees must defend the response they choose. Such additions serve to raise the cognitive demands of the items. The item writers' competence in designing sound true/false items is contingent on their ability to identify useful propositional material in the content domain or in the scenario, if a context dependent item is created. The true/false format has been used in original form or slightly adapted by skilled item writers to test performances requiring generalizations, comparisons, causal relationships, evaluations, as well as computational skills. For an example with a variation, see Item 3 in Box 5.7.

Critically Reviewing True False Items. Box 5.8 shows some T/F item examples tied to a beginner’s course in statistics. Review the items alongside the criteria for writing sound True/False items in Table 5.2.

The first rule for True/false (T/F) item format speaks to improving the content-based validity of the items—that is, verifying a match of the item with the content, condition, and taxonomic level specified in the selected learning outcome(s). Several of the general criteria apply here for the same reasons discussed with multiple choice items, including the avoidance of "double-barreled" items that test more than one concept or principle.

Consider the T/F item below

Indicate whether True or False
Albany is the capital city with the largest population in New York state. ____

(Poor item!)

Here, the first half of the item is true, but the latter half is false. Hence, this item not only tests two different propositions, but one is false while the other is true! This makes it a poorly designed T/F item.

It is important for a T/F proposition to be indisputably true or false. In an opinion-based item where opinions can vary, or when there is a need to test knowledge of different sources that provide different data or facts, the source should be clearly identified. Consider the next four examples.

Indicate whether True or False

1. Albany, NY has a population of 80,000. (*Vague. Depends on year of census, without which it is unclear as to whether the statement is True or False*)
2. The 2016 state database states that Albany, NY had a population close to 100,000. (*True, as the population is reportedly ~97,000, Acceptable item!*)
3. According to the 2016-17 state database, the population of Albany, NY was between 90,000-100,000. (*True. Most clear and fact-based statement*)

Negatively oriented words, as with multiple choice items, should be used sparingly and highlighted, to avoid ambiguous communications. Double negatives should always be avoided, as these create comprehension barriers unrelated to the outcome/indicator, outside making a statement neither true nor false.

Indicate whether True or False

1. According to the 2016 state data base, Albany, NY did **NOT** have a population of 80,000. (*Acceptable!*)

2. **Nobody** knows whether Albany, NY **does not** have a population of 80,000. (*Poor!*)

As before, clues should be avoided by keeping true and false statements of approximately equal length, distributing T/F answers randomly in a list of items, and with approximately equal numbers of true or false answers on a test. One type of clue applicable to all structured response items is a "clang" association. Here, the correct answer is made obvious because of repeated use of the same word(s) or language in an item or response option.

Indicate whether True or False

1. **Content-based** validity has to do with matching the **content** of a test to the **content** of the items. (True. Poor item as it suggests the answer by clang association!)

For all items, the answers should not be obvious to uninstructed groups. We should therefore avoid measuring trivial topics or ideas that would be common sense, as this makes the assessment process inefficient. For example:

Indicate whether True or False

1. Teachers should use tests. (*Poor!* The statement is relatively obvious and trivial. Also, it is an unsupported opinion, not suitable for T/F.)
2. According to the 2014 NRC report on science assessments, teachers should use tests to support students' learning progressions. *(Better!)*

Finally, absolutes such as "always" or "never" are usually found in statements where the item writer attempted to make the statement completely true or completely false. tem writers tend to use catchall terms such as "usually" or "generally" for the same purposes. Such “specific determiners” serve as semantic clues to verbally adept examinees, giving reason to avoid their use when constructing items. These item flaws are easily detected by the test-wise!

5.6.3 Matching Exercises

In its typical form, a matching exercise is comprised of two adjacent columns of information that must be matched by the examinee through simple association. Column A presents the item prompts or *premises*, while Column B lists the *response* options.

Review the item examples in Box 5.9 and 5.10. The topic again deals with terms and definitions of descriptive statistics from a basic statistics curriculum. The next box illustrates another matching exercise from an elementary school science curriculum. In the first example, the examinee must associate the terms with their correct definitions. Items are numbered. Answer options are identified with letters. In the second, the test taker matches the parts of a seed bearing plants with functions.

Common applications of matching item exercises utilize the following types of lists for Columns A (in bold) and B.

A	B
Inventors	Inventions
Authors	Quotations
Problems	Solutions
Scientific Theories	Applications

A	B
Principles	Illustrations
Parts	Functions
Rules	Examples
Terms	Definitions

Although the taxonomic level most typically measured is concept recall and understanding, not all matching exercises need be at rote levels or lower cognitive levels requiring students to make associations. For example, a matching exercise on Problems versus Solutions in mathematics could be devised to assess application and complex procedural skills. Similarly, Scientific Theories versus Applications could be used to test higher order thinking. As with other structured formats, a matching exercise could be combined with open-ended explanations to increase cognitive demands of the task.

Critically Reviewing Matching Exercises. Table 5.2 presents some of the main guidelines for this item type. We should begin by evaluating the exercise vis-à-vis the general criteria for all items, including alignment of the items with the targeted outcomes and indicators and screening the exercise for potential biases. Then, examine the remaining format-specific rules.

Heterogeneous response options make a matching exercise too easy, as the distracters that do not belong may look obviously wrong to examinees with minimal knowledge of the domain tested. Consider the following list of heterogeneous answer options (Column B) for the exercise shown in Box 5.9, tapping into knowledge of statistical terms.

Column B

a. Mean
b. Instruction
c. Median
d. Standard deviation
e. Lesson plans
f. Range
g. Quartile

Here, Instruction and Lesson plans are from a pedagogical curriculum for teachers, and lie outside the outcomes of the statistics course. They are poor distracters, as well, and will be easy for a canny examinee to rule out. Instead, distracters such as Frequency and Percentile (Box

5.9) are more homogeneous with the item content specifications, plausible, and a good fit for the domain. Similarly, having more response options than prompts increases the difficulty level of the exercise.

Again, the responses should ideally be listed in some logical order to prevent clues. In Box 5.9, they are organized alphabetically. Contrast that with the example in Box 5.10, where the last response option, Pollen, is not in alphabetical order and also a wrong answer. This option may be easier to strike out because of its location, rather than a grasp of the material tested.

Other standard guidelines pertain to the clarity of presentation and format of matching exercises. Directions should be clear, with key terms bolded to draw examinees' attention to relevant information, thereby lowering chances for random, unreliable responses.

As in multiple choice items, the reading load should be in the premises presented in Column A of matching exercises. Longer statements in premises should be balanced against shorter response options, to help item-writers pose question clearly and completely to examinees. Also, the exercise should fit one page, to prevent interrupting the flow of thinking as examinees respond.

5.6.3 Completion

The completion or fill-in-the-blank item is another item format that is relatively easy to build. It is different from the selected-response variety in that examinees must supply an answer on their own. The response is usually required in a single word, short phrase, a number, symbol, or formula. The common presentation formats are (a) the incomplete sentence, (b) a direct question, or (c) a statement of the problem in a complete sentence, each followed by the blank(s).

The supply format opens up the possibility of examinees coming up with a range of answers that might be partially correct or nearly correct, but not a perfect match with the response in the answer key. This introduces potential ambiguity and possible errors in objective scoring of the completion items. It is therefore important that item writers specify the *exact range of answers that they consider to be fully correct,* as opposed to partially so or incorrect, when a completion item is conceived. Such information should ideally be included in the design specifications for the assessment. Rubrics, discussed later, will also help.

Completion item formats have been used effectively for tapping basic concept recall and understanding, as well as application levels. Most such usages, however, involve only a few steps of calculations or interpretations. The last item in Box 5.11 illustrates how this can be achieved.

Critically Reviewing Completion Items. Table 5.2 summarizes the main rules for writing completion items. To start, the general criteria would apply to this item format as well, beginning with an evaluation of how well the items match the content, conditions, and taxonomic level of the targeted outcomes and indicators. Ensuring item clarity and absence of biases follows next.

As shown in Box 5.11, well-written completion items present the problem clearly, whether in question form, or as incomplete statements. The blank should test *important* concepts, key words or principles that are directly relevant to the content or skill specifications in the learning outcome. Verbatim lifting of text from books or required readings leads to testing of rote learning and memorization, rather than deeper understandings—therefore, we should avoid it for all item formats.

Multiple blanks in a completion item are unacceptable; these add much ambiguity to the question and could lead to an array of responses that are difficult to score systematically. Consider the four versions of the same item tied to the outcome in Box 5.11 next.

1. In a data distribution, the score_____ the highest frequency is called the mode.
   Answer: with. *Poor item! The preposition, "with" is trivial and not a key concept worth testing*.

2. In a ____ distribution, the _____ _____ the ____ frequency is ___ the ____.
   Answers: data, score, with, highest, called, mode. *Still poor! Problem is unclear and incomplete, as presented*. *Too many blanks in the item, with trivial, contextual and key words missing, hampers communication.*

3. In a data _______, the score with the highest frequency is called the mode.
   Answer: distribution. *Still weak, as "distribution" is a contextual detail, not the key concept*.

4. In a data distribution, the score with the highest frequency is the _____.
   Answer: mode. *Much better! Key concept tested*, *tied to outcome*, *clear question with single blank placed at the end*.

A "Clozing" Note. While the better designed completion item has one blank, placed at the end, the Cloze exercise is a variant of the completion item format, where multiple blanks may be applied. Chapter 3, Box 3.4 provides an example.

When invented, *Cloze* exercises were intended to assess basic reading comprehension (Taylor, 1953). In their original form, words from a body of running text would be deleted at fixed intervals by the designer. For instance, every fifth word in a sentence would be a blank, irrespective of its content, length, or relevance to the main theme of the text. Respondents filled in the missing words based on context cues in the larger passage, thereby demonstrating ability to comprehend the overall gist. Cloze tests have been effectively employed in foreign language assessments. Cloze exercises can be made more effective by deleting key words rather than words at particular intervals.

5.6.4 Context-dependent Item Series

Context–dependent item series are more useful for tapping higher levels of cognition than any structured response item format alone, as the example in Box 5.7 demonstrated. Another example of a context-dependent item set is shown in Box 5.12. This example capitalizes on a variation of the T/F item format to test young students' capacities to problem-solve by sorting through relevant and irrelevant information in an age-appropriate scenario. The open-ended component in Box 5.12 requires justification of the answers, which increases the cognitive demands on examinees. Appropriate rubrics are necessary to score such combinations of responses with validity, fairness and reliability—a topic we undertake in-depth in the sections on Constructed Response Tasks.

Context-dependent item series have been successfully applied to assess the following types of outcomes and indicators (Linn & Gronlund, 2000; Miller, Linn & Gronlund, 2009):

- Discerning the relevance of information for a purpose, as illustrated in Box 5.12;
- Evaluating arguments presented in contextual material;
- Recognizing, or distinguishing between, valid and invalid conclusions from contextual material provided in tables, graphs, pictures or text passages, illustrated in Box 5.7; or
- Interpreting patterns or relationships in data tables or informational text passages provided.

Critically Reviewing Context-dependent Exercises. All the rules for writing specific structured response items would still apply when we design context dependent exercises. However, a question to settle beforehand is whether such items should be **content free**, where there are no expectations about examinees' having any particular subject matter expertise, versus **content-dependent**, where examinees are expected to draw on some subject matter knowledge when problem-solving. Whether or not an item series can be created to be completely free of content, is still unresolved. However, in content–heavy item series that also test higher-level thinking, item-writers should pre-specify the parameters of subject matter knowledge required.

Other guidelines are the following. Contextual material should preferably be new to examinees but appropriate for measuring the outcomes and indicators for the targeted level. The contextual material should be clearly presented, uncluttered and easy to read. For efficiency, questions should target higher levels of thinking, analysis or evaluation that cannot be captured well with individual structured response items. All items should be tied to the contextual material. Finally, the list of questions and the contextual material should be balanced in terms of space. The complete exercise should be short enough to fit on one-half to one page, unless examinees are older.

Application Break

- Create two (2) items in each of the following formats to tap any cognitive or proficiency-based domain of your choice. You may use the Long Division domain in Table 5.1, as well.
    - Multiple choice, True/False, Completion, Matching, and Context-dependent Item sets.
    - Evaluate the quality of your items using the criteria in Table 5.2. Which would you revise and how? Explain your reasons.

- Clues threaten sound measurement with structured response items and context-dependent item series. List the different types of clues that we should avoid. Give item examples to illustrate two of the most common clues.

5.7 Guidelines for Designing Constructed Response Tasks and Essays

The constructed response item format includes **short response tasks** or **essays.** Such items require a written prompt and assume that respondents are capable of written communication. The examinees generally supply a response on an answer booklet or a computer. Both formats are well-suited for measuring higher levels of thinking and problem-solving, and the length of the response can be delimited with appropriate directions. While some degree of open-endedness in item structures may work for very young children or special needs populations in interview-based presentations, tasks that call for lengthy responses or independent writing are not suitable for non-writers.

Constructed response items should involve thought-provoking tasks, tapping learning outcomes and indicators are note well measured with other item formats. Some are:

- Broad domains with multiple indicators that need to be measured as a whole, such as the writing of a letter, construction and interpretation of a graph using data tables, or solving a multi-step problem in math.

- Outcomes that involve skills in reasoning, argumentation, defending, justifying, or explaining a problem, solution or issue, such as, presenting a persuasive argument on global climate change.

- Process indicators, involving complex procedural skills. An example might be application of a creative writing process to generate a publishable poem, report or a story. A writing process, taught in most Language Arts curricula at the secondary and post-secondary levels, commonly involves a multi-step procedure, including brainstorming ideas, drafting, editing, revising, and generating a finished article.

Clearly, the constructed response task format is a highly desirable format for some assessment purposes and domains; but it is also vulnerable to several sources of error. A first challenge is scoring the open-ended responses with validity and reliability. Very cumbersome or loosely defined criteria can also affect the utility of scoring rubrics in different ways. See Figures 5.2 -5.3 for two design options.

Figures 5.2-5.3 about here

5.7.1 Devising Scoring Criteria

Figure 5.2 provides an example of a set of design specifications for assembling a developmental assessment on number patterns utilizing constructed response tasks combined with completion items. These assessments were intended for assessing students' learning progressions for classroom decision-making purposes and for end-of-year summative decisions at school. As indicated, greater standardization and rater training are necessary when high stakes decisions are made with rubric-scored results.

On the right hand side of Figure 5.2, we see the domain with outcomes and progressively ordered indicators. The latter calls for examinees to identify, continue, and explain patterns with mathematical content at five levels. On the left hand side, we see a sample, 4-part task at Level 3 to match the indicator, accompanied with a scoring rubric. The rubric has a rating scale at 4 levels, ranging from lowest (0) to highest (3). Each performance level tied to descriptors.

This is a holistic rubric. As seen, it yields a summary score denoting the overall proficiency level evidenced in a 4-part response. Holistic rubrics help rank order performances by quality, and are suitable for making summative decisions.

To devise a holistic rubric like the one shown, one starts with the domain and then draws on sample answers of typical examinees. **Anchor papers** are actual answers that represent performances on a given set of tasks, organized at different levels. Once tasks are tried out, the real examinee responses can be critically reviewed and placed in separate piles by expert raters, distinguishing among those that qualify for a score of 1 versus a 2 or 3. Anchor papers are not only useful for writing the descriptors that define points of the rating scale in a holistic rubric, but also during scoring when they serve as concrete references for scorers. Anchors can significantly cut down both systematic and random rater variabilities, improving validity and reliability of the results. Ideally, we should plan to use real response data as a part and parcel of designing constructed response tasks.

Contrast the holistic rubric to the **analytic rubric** presented in Figure 5.3 for the same domain and assessment tasks. In an analytic scoring system, the response is broken down into relevant parts, and each part is assigned a separate score. Indicators might be weighted differently, based on their relative importance to designers and users.

Often, a student or examinee may show a strong performance on some indicators, but much weaker performance on others. Hence, analytic rubrics offer a diagnostic profile useful for formative decision-making. All rubrics must align with relevant indicators specified in the domain, but as with holistic rubrics, analytic rating scales and criteria can be significantly improved by drawing on real responses. A detailed analysis of common errors that typical students/examinees make is often the best data source for refining analytic rubrics.

In the example of the analytic rating scale shown (Figure 5.3), if different students vary in their competence on the criteria, such as, at identifying the number pattern, effective strategy use (like trial and error), performing computations, using mathematical terms, or even understanding task directions, well-designed analytic rubrics will pick these differences up. While it is possible to sum ratings across all the indicators to obtain a summative total score, it is the property of describing differential performances on indicators that gives the rubric its analytic character and diagnostic utility.

5.7.3 Critically Reviewing Essay and Constructed Response Tasks

Table 5.4 summarizes the essential criteria to guide item design and task reviews. The first set (A) apply to all assessment tasks.

In section B, we see specific criteria for refining or evaluating this class of items. Because open-ended tasks are costly in terms of time and effort needed for scoring, they should be reserved for only those outcomes/indicators that cannot be tapped otherwise. Connected groups of indicators, capturing the culminating performances are well-tapped by these items.

Delimiting open ended tasks with clear parameters is key to assuring measurement quality. Vague essay prompts, and associated directions, do a poor job of focusing the respondents on the expected performance and are likely to generate widely variable responses that are difficult to grade with any kind of meaningfulness or consistency. Consider the following prompt, which led secondary students to provide the following range of responses.

Prompt: Explain Climate Change.

Answers obtained:

A list of major facts on global warming and climate change
A description of the event with some key scientific facts
An explanation of the major facts with supporting details

A brief analysis of possible causes and effects
A well-argued case for environmental policy changes
A combinations of the above, plus/minus factual errors

To correct the issues above, item designers could alter task-specific directions as well as the prompt. For example:

Directions: Write a 2-to 4-paragraph essay on the following science topic. You will be graded based on your ability to link the causes with effects, and on how well you can justify your answer with scientific facts and sources. Points will be awarded for accuracy of scientific evidence, principles, events, and any names of people and organizations you cite. No points will be deducted for English language, but please try your best to follow formal writing conventions discussed in class.

You will have 90 minutes to complete the essay. Please type your answer on the computer. This is a closed book test.

Essay Prompt: Describe the phenomenon of **global warming** based on the most current scientific evidence and facts. What steps are recommended by the scientific community to curb the phenomenon? Do you agree with the hypothesized causes and effects of **global warming** on society? Justify your answer.

To improve the design of the prompt, we must ask: What were we looking for as evidence of high performance? Referring back to targeted learning outcomes will help us delimit the exercise. Notice that the length of the expected answer, the scoring criteria, the expected contents of the essay, and conditions (e.g., time limits, closed book) under which the student must respond, are all defined for the examinee in the directions. Inclusion of the actual scoring rubric could further focus the examinee on the performances we expect. Structuring of the exercise in the manner shown will control for likely errors from scorers, as well.

Bluffing can be built into poorly conceived essay or short answer question prompts. Effective bluffing will camouflage what examinees do *not* know or *cannot* do!
Without being held accountable for explicit kinds of knowledge or skills tied to a topic or domain, some respondents may be tempted to cleverly "fake" their way through an answer with good writing skills or other distracting information having little to do with curriculum objectives or indicators being tested.

Well-prepared scoring rubrics must accompany all constructed response items, and preferably be shared with examinees prior to assessment. Validity in rubrics is determined by the extent to which listed criteria match the targeted outcomes/indicators of the construct domain. To facilitate consistent scoring, rubrics should be clear, unambiguous, and preferably supported with sample answers and data on typical errors. All these will enhance the ease with which rubrics can be used by raters.

5.8 Putting it all Together: Applying the Process Model

Figure 5.1 highlighted that Phases II-III deal with specifying the operations and designing/selecting items to measure the targeted construct, and assembling the instrument. Item design procedures should be situated within the complete Process Model given in Figure 1.6. We end this chapter, with a quick review of how the process would apply for both assessment designers and users.

A first step in operationalizing a construct is developing the assessment design specifications and compiling a set of tasks or items to match. Figure 5.2 presented an example of assessment design specifications for compiling mathematical patterns assessments to support K-12 school-based inferential needs and uses. As clear, the assessment operations go well past just the items, tasks and scoring rubrics.

Well-thought out assessment design specifications must lay out the assessment purposes, the domain, and the population (Phase I). They also specify the assessment conditions, item features and scoring methods, and materials in adequate detail to help generate a tightly designed item pool (Phase II-III). Once the complete design is specified in detail, it is easier to compile an assessment to match the requirements. Who puts together the specifications? It should ideally be both designers and users. In the project through which the patterns tasks were produced, the assessment specialists worked in close concert with teachers and school leaders of the school system (Banerji, 1999; Banerji & Ferron, 1998).

Another example of assessment design specifications is in Table 5.4. This medical education domain is the foundation for assessment items like the ones in Box 5.7 (the same domain was illustrated in Chapter 4). Horizontally, we see the overall content dimensions to be tapped by the items. These refer to four sub-domains of clinical action expected of physicians, dealing with Therapy, Diagnosis, Prognosis, and Harm--the latter deals with assessing harmful side effects of given treatments. The vertical dimension lays out the cognitive levels to be tapped by the items comprising the test. The cells show the number of items in each category, indicating how the total score generated by the test would be weighted across the content and cognitive level dimensions. This table served as the blueprint for producing parallel test forms for a research project examining resident physicians' EBM practices, each with distributions of items shown (Wyer, 2008; Wyer & Chatterji, 2013).

Phase IV, content and empirical validation, should follow all item design efforts, whether in research or applied contexts. As we saw, before items can be finalized, some level of content validation and critical review enhances the quality of tasks and items, and is necessary even in informal assessment settings.

Instrument assembly following item design calls for attention to several added details. Some were suggested throughout. The main needs now summarized below. A finished assessment should ideally include:

- A succinct title that clarifies the assessment's purpose and main topic;
- Systematic item organization and layout design, whether by sub-domain or in random order;
- A medium for item presentation that is suitable for the population, whether paper and pencil, computer or technology-based medium;
- A clear set of directions for items and overall assessment, as needed, for test-takers, scorers, and others involved in assessment administration, especially if it is a standardized assessment.
- A validated scoring key and rubrics, as applicable, with supporting materials for scorers.

5.9 Summary

What better way to close the discussion, than to look at another instructive message from a *Peanut*s cartoon strip? (United Features Syndicate, Inc., 1970, as excerpted in Mehrens & Leham, 1984, p. 94).

Patty (Reads an examination in school.):

> .....*"An essay test. I'm doomed".*
> ... "*Why couldn't she have given us a multiple choice test?."*
> ... *" Or a true false test?."*
> ... "*I hate it when you have to know what you're talking about...."*
> ....

Chapter 5 dealt with assessing cognitive and proficiency-based constructs. The chapter began by reviewing key findings of the cognitive science literature and their implications for rethinking item design approaches. Good item designs aim to bring out what examinees “know” and can do in given domains and at particular points in time. Current research and understandings about the human brain suggest that our mental capacities are not permanent, but develop and change throughout lifetimes.

Traditional versus cognitively-informed item designs are grounded in different research bases and guided by contrasting philosophies of assessment design. Assessment designs that are cognitively-informed attempt to bring to light a given person’s knowledge structures in a domain on a developmental continuum. They are well suited for diagnostic and formative assessment purposes. Item design approaches to facilitate classroom instruction and assessment, for example, yield information to help teachers in providing feedback, practice and specific interventions to help students alter their thinking and reach learning targets. Such designs have utility in health and rehabilitative assessment applications, as well.

Traditional assessment design approaches focus more on products of learning. These are best suited for point in time, summative measurements of a person’s status in cognitive domains that could be used for predictions and other applications in education, the workplace or sundry societal contexts.

The intended functions for an assessment should guide item designs we choose. The chapter provided guidelines for writing and critically reviewing the quality of several item types: multiple-choice, true- false, completion, matching, short answer and essay. Several item illustrations demonstrated common pitfalls and strategies to circumvent them.

Common item construction errors arise from ill-chosen or inappropriate use of language; mismatched reading levels for the targeted population; inadvertent and built-in clues; erroneous keys or badly crafted scoring rubrics; exercises that overlook important construct indicators or are misaligned with critical indicators of the construct domain; confusing item directions; biases; and complicated task formats.