MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2019.1595556

MEDICAL TEACHER

Taylor & Francis
Taylor & Francis Group

AMEE GUIDE

A practical guide to test blueprinting

Mark R. Raymond[a] and Joseph P. Grande[b]

[a]National Board of Medical Examiners, Philadelphia, PA, USA; [b]Mayo Clinic College of Medicine and Science, Rochester, MN, USA

ABSTRACT

A *test blueprint* describes the key elements of a test, including the content to be covered, the amount of emphasis allocated to each content area, and other important features. This article offers practical guidelines for developing test blueprints. We first discuss the role of learning outcomes and behavioral objectives in test blueprinting, and then describe a four-stage process for creating test blueprints. The steps include identifying the major knowledge and skill domains (i.e. competencies); delineating the specific assessment objectives; determining the method of assessment to address those objectives; and establishing the amount of emphasis to allocate to each knowledge or skill domain. The article refers to and provides examples of numerous test blueprints for a wide variety of knowledge and skill domains. We conclude by discussing the role of test blueprinting in test score validation, and by summarizing some of the other ways that test blueprints support instruction and assessment.

Introduction

Assessment plays a major role in the medical school curriculum by providing a way to monitor student progress toward the learning outcomes that we expect them to achieve (Shumway and Harden 2003). To ensure that assessments are consistent with course objectives and address truly important learning outcomes in a balanced manner, it is important that assessments be developed according to a well thought-out plan. This article describes a systematic approach to planning tests—an approach that documents what students should know and be able to demonstrate on each assessment. These planning documents are typically called test blueprints, although they also are known as test plans, tables of specifications, and test specifications. A test blueprint describes the key properties of a test. While any test blueprint should specify the content to be covered, many blueprints also describe properties such as the amount of emphasis allocated to each content area, the cognitive demand of the assessment tasks, the assessment format, and other important features (Millman and Greene 1989; Raymond 2016).

In the text that follows, we discuss the importance of learning outcomes in test blueprinting, introduce the notion of evidence-centered test design, and review two common taxonomies of learning in medical education. We then present a four-step process for developing test blueprints, providing several examples and key references along the way. The article concludes with a discussion of the broader contribution of test blueprints to teaching and assessment.

Practice points

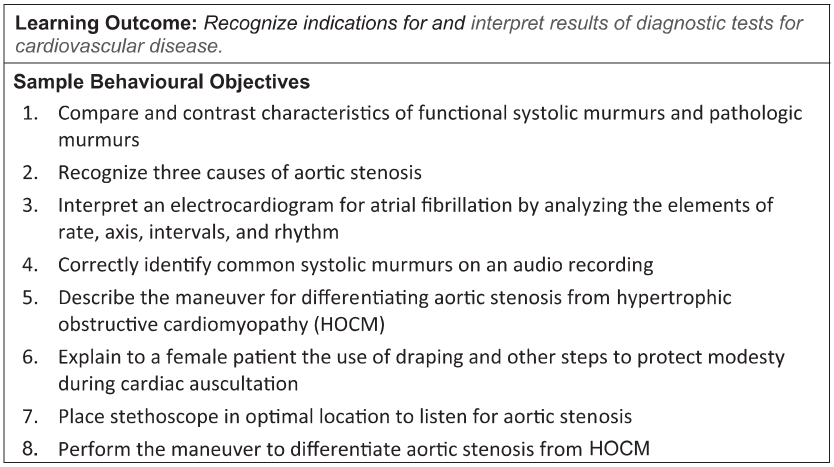
- Test blueprints describe the content to be covered by a test, along with other important features (e.g. emphasis given to each topic; the assessment format). A test blueprint is also known as a test plan, table of specifications, or test specifications.
- Sources of information for test blueprints include course outlines, lists of learning outcomes and behavioral objectives, lecture notes, textbooks, and other curricular materials.
- Developing a test blueprint consists of four stages: (1) identify the major knowledge and skill domains; (2) delineate the objectives or learning outcomes to be assessed within each domain; (3) determine the assessment formats; and (4) specify the weight to be given to each content category (i.e. knowledge and skill domain).
- The category weights of the test blueprint serve as a sampling plan; they indicate how much content (e.g. how many test questions) to sample from each knowledge or skill domain.
- Besides providing a link between instruction and assessment, test blueprints support medical education in other ways such as serving as a study guide for students, providing the basis for student feedback; and providing a framework for evaluating instruction.

Foundations of a test blueprint

Learning outcomes and claims about student competence

A test blueprint is a natural extension of the learning outcomes and course objectives that most instructors already have in place. We define *learning outcomes* as broad statements that describe the knowledge and skills that students should possess upon completing a course (Harden 2002). *Behavioral objectives* (or instructional objectives) are statements that describe in detail what students are expected to know and be able to do. Learning outcomes indicate

CONTACT Mark R. Raymond mraymond@nbme.org NBME, 3750 Market Street, Philadelphia, PA 19104, USA.



Learning Outcome: *Recognize indications for and interpret results of diagnostic tests for cardiovascular disease.*

Sample Behavioural Objectives

1. Compare and contrast characteristics of functional systolic murmurs and pathologic murmurs
2. Recognize three causes of aortic stenosis
3. Interpret an electrocardiogram for atrial fibrillation by analyzing the elements of rate, axis, intervals, and rhythm
4. Correctly identify common systolic murmurs on an audio recording
5. Describe the maneuver for differentiating aortic stenosis from hypertrophic obstructive cardiomyopathy (HOCM)
6. Explain to a female patient the use of draping and other steps to protect modesty during cardiac auscultation
7. Place stethoscope in optimal location to listen for aortic stenosis
8. Perform the maneuver to differentiate aortic stenosis from HOCM

Figure 1. Learning outcome and sample behavioral objectives for a unit of instruction on the cardiovascular system.

the knowledge and skills that are expected of students, while behavioral objectives serve as a road map for getting there. We use the term *assessment objectives* to describe those learning outcomes or behavioral objectives that are specifically targeted for assessment.

A primary goal of assessment is to allow an instructor to make a claim or inference about what students know and are able to do. A test creates the opportunity to obtain evidence to support such claims (Mislevy and Riconscente 2006). To back up the claim that a student has "mastered the knowledge and skills to diagnose major conditions of the cardiovascular system", an instructor needs to identify assessment tasks to elicit the behaviors of interest, and then provide the opportunity for the student to demonstrate those behaviors. Evidence-centered design requires that faculty choose assessment tasks that provide the evidence to support the claims to be made about student competence. Such claims and evidence are necessary elements of Kane's (2016) argument-based approach to test score validation.

Consider a course that has an overall goal of ensuring that students can diagnose the most common diseases affecting each organ system. Figure 1 identifies a learning outcome and several objectives specific to the cardiovascular system. The learning outcome, "Recognize indications for and interpret results of diagnostic tests for cardiovascular disease" is fairly broad; being able to successfully demonstrate that outcome requires a substantial network of knowledge and skills. The eight behavioral objectives in Figure 1 are just a sample of the behaviors required to demonstrate mastery of that learning outcome. Although these behavioral objectives are intended to guide instruction, it is easy to see how they also can inform assessment. By preceding each objective with a clause like "*The student will be able to…*" these statements can be transformed into assessment objectives that support claims to be made about students. Although behavioral objectives are tedious to develop, their specificity can simplify the development of assessment tasks.

Types and levels of knowledge

While some of the objectives in Figure 1 require cognitive learning, others involve skill learning. It has long been recognized that these different types of learning require different methods of instruction and assessment. A student can learn the verbal description for mitral valve prolapse, but still not be able to identify it by sound until actually listening to the sound paired with the waveform from an ECG. Miller's (1990) pyramid is one popular framework for organizing types of learning. As depicted in Figure 2, the four levels of the pyramid are *knows*, *knows how*, *shows how*, and *does*. The base of the pyramid consists of conceptual knowledge, while the second level describes the ability to apply that knowledge in some context to solve a problem. The third level requires a student to explain how they would perform a task or to demonstrate that behavior in a simulated or practice situation, and the fourth refers to performance in routine practice.

The central portion of Figure 2 depicts the original pyramid, while the text on the left lists two sample behavioral objectives and where they fall in the hierarchy. Referring back to Figure 1, it can be seen that the first two objectives are at the bottom of the pyramid, while the remaining objectives would be located toward the apex. Meanwhile, the right portion of Figure 2 indicates the types of assessments suitable for different levels of the pyramid. It is evident that evaluating skills higher in the pyramid requires more clinically authentic assessment tasks. We return to Figure 2 later to show how the pyramid can help decide on an assessment format.

Bloom's (1956) taxonomy is another model that is useful for assessment. It classifies behaviors into three categories referred to as the *cognitive*, *affective*, and *psychomotor* domains (Anderson and Krathwohl 2001). One can see how the instructional objectives in Figure 1 can be classified into one or more of these three domains. For example, objective 7, "*Place stethoscope in optimal location to listen for aortic stenosis*", requires deciding where to place the stethoscope (cognitive), explaining its placement to the patient and asking permission (affective), and then positioning the stethoscope (psychomotor).

The cognitive domain has received the most attention in teaching and assessment. It consists of six levels: *knowledge*, *comprehension*, *application*, *analysis*, *synthesis*, and *evaluation*, with each level requiring a greater cognitive investment than the one preceding it. The level of a behavioral objective has implications for assessment. For example, while multiple-choice questions (MCQs) are

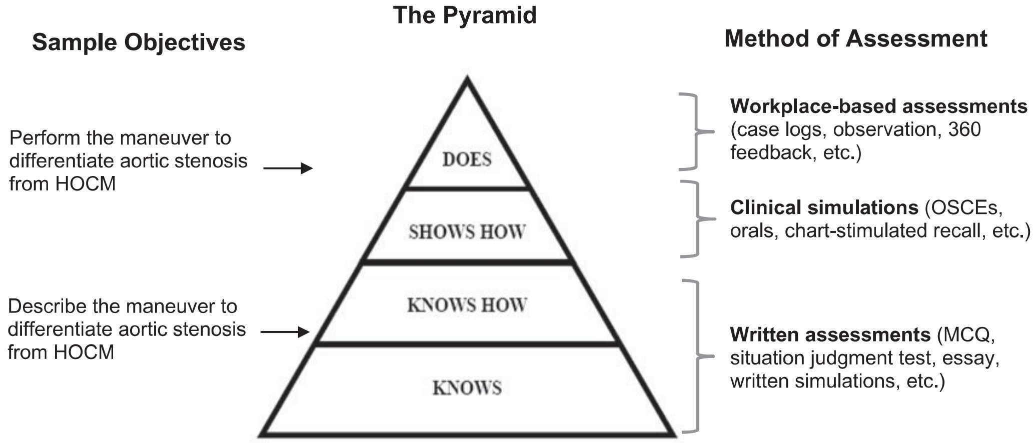

The Pyramid
Sample Objectives
Method of Assessment
Perform the maneuver to differentiate aortic stenosis from HOCM
Describe the maneuver to differentiate aortic stenosis from HOCM
DOES
SHOWS HOW
KNOWS HOW
KNOWS
Workplace-based assessments (case logs, observation, 360 feedback, etc.)
Clinical simulations (OSCEs, orals, chart-stimulated recall, etc.)
Written assessments (MCQ, situation judgment test, essay, written simulations, etc.)

Figure 2. Miller's pyramid with sample behavioral objectives and suitable methods of assessment.

effective for assessing knowledge and comprehension, they have limited utility for assessing synthesis and evaluation, and are ineffective for assessing many skills in the psychomotor and affective domains.

Although the scientific merit of Bloom's taxonomy and Miller's pyramid has been called into question over the years (Cizek et al. 1995; Al-Eraky and Marei 2016), these frameworks have endured because of their practical utility. Their use promotes skill acquisition beyond the memorization of simple facts. While such frameworks should not be taken too literally, they remind us that learning can be shallow or deep, and that it can reside in our heads or in the actions we take. In short, they encourage the development of tests that assess the higher-order thinking skills and clinical competencies that are important to professional practice.

Four stages to an effective test blueprint

Figure 3 and the text below describe four stages for developing a test blueprint. Although an effective blueprint will touch on all four stages at one point or another, consider the stages that follow as a general guide to assessment planning, not as a formal protocol that requires execution in strict sequential order.

Identify the major knowledge and skill domains

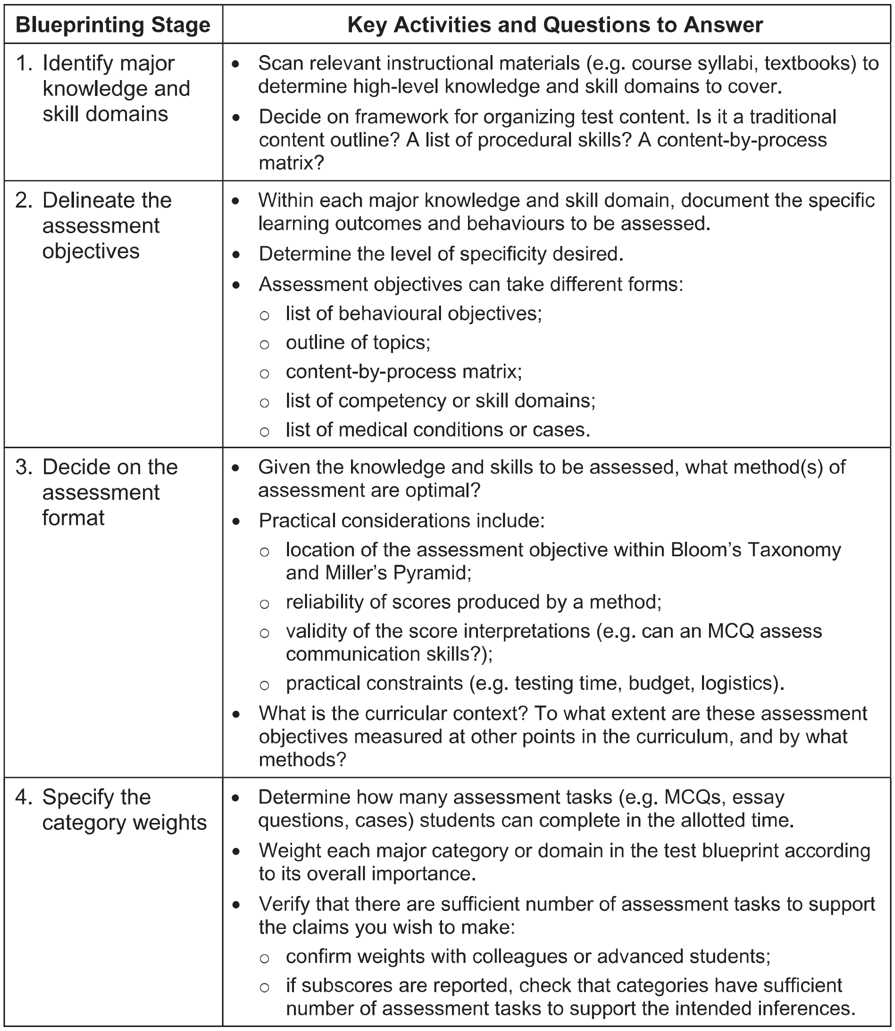
This first stage involves identifying the broad competency domains to be assessed. They should correspond to the high-level claims to be made about student behaviors. One way to approach this task is to ask, "If I were to partition my course into a few to several units, what would the labels be?" If the intent is to make claims about students' knowledge of immunology, then the framework should include major content categories that define the domain of immunology. If the intent is to make claims about a student's ability to interact with patients, then the framework will include major types of communication skills. Such documentation can be found in course outlines, learning outcomes, lecture notes, textbooks, and other instructional materials. In addition, curriculum surveys and job analysis reports have identified the skills that are important for curriculum design and assessment in medical education (Boulet et al. 2003; Patterson et al. 2008; Raymond et al. 2011; Gaffas et al. 2012; Zhao et al. 2012; Angus et al. 2014; Touchie and Streefkerk 2014; Baker et al. 2017).

Test blueprints typically are cast in the form of an organized list, outline, or table. A test blueprint's organizational framework is important because the categories often parallel the claims to be made and the feedback provided to students. As described next, test blueprints can be organized according to the content to be tested or around the behavioral processes required of the assessment tasks (Millman and Greene 1989; Raymond 2016).

Content-oriented blueprints describe tests in terms of the topics or subject matter covered. They usually slice up the test material according to traditional academic disciplines. For example, a comprehensive test covering a preclinical year of medical school might include categories such as physiology, pharmacology, biostatistics, and so on. At a more specific level, a test for a course in biostatistics would include categories such as descriptive statistics, multiple regression, and related topics. Note that blueprints for problem-based curricula take on a slightly different structure; they are best organized according to the types of cases encountered, based on the diagnosis or presenting complaint of each case (Boulet et al. 2003).

Process-oriented test blueprints delineate the procedural skills students are expected to demonstrate. Many process-oriented test blueprints include skills from the cognitive domain of Bloom's taxonomy. An instructor might use Bloom's taxonomy to document that 40% of a test requires comprehension; 40% requires application of knowledge to solve clinical problems; and 20% requires analysis of an experiment. Miller's pyramid can be used in a similar way to decide, for example, that 70% of a statistics test will include tasks at the *knows* and *knows how* levels, while 30% will require students to *show how*. Process-oriented frameworks are particularly useful for clinical training where the emphasis is on procedural skills and the affective domain. Two process models with particular relevance to medical education are the CanMEDS framework (Frank et al. 2015) and the ACGME competencies (Batalden et al. 2002); different parts of these frameworks can be very useful for developing classroom assessments.

The preceding text suggests that test blueprints are either content-oriented *or* process-oriented outlines. In fact, many blueprints integrate these two dimensions into a single framework called the content-by-process matrix. The



Blueprinting Stage	Key Activities and Questions to Answer
1. Identify major knowledge and skill domains	• Scan relevant instructional materials (e.g. course syllabi, textbooks) to determine high-level knowledge and skill domains to cover. • Decide on framework for organizing test content. Is it a traditional content outline? A list of procedural skills? A content-by-process matrix?
2. Delineate the assessment objectives	• Within each major knowledge and skill domain, document the specific learning outcomes and behaviours to be assessed. • Determine the level of specificity desired. • Assessment objectives can take different forms: ○ list of behavioural objectives; ○ outline of topics; ○ content-by-process matrix; ○ list of competency or skill domains; ○ list of medical conditions or cases.
3. Decide on the assessment format	• Given the knowledge and skills to be assessed, what method(s) of assessment are optimal? • Practical considerations include: ○ location of the assessment objective within Bloom's Taxonomy and Miller's Pyramid; ○ reliability of scores produced by a method; ○ validity of the score interpretations (e.g. can an MCQ assess communication skills?); ○ practical constraints (e.g. testing time, budget, logistics). • What is the curricular context? To what extent are these assessment objectives measured at other points in the curriculum, and by what methods?
4. Specify the category weights	• Determine how many assessment tasks (e.g. MCQs, essay questions, cases) students can complete in the allotted time. • Weight each major category or domain in the test blueprint according to its overall importance. • Verify that there are sufficient number of assessment tasks to support the claims you wish to make: ○ confirm weights with colleagues or advanced students; ○ if subscores are reported, check that categories have sufficient number of assessment tasks to support the intended inferences.

Figure 3. Summary of key activities for developing test blueprints. See text for details.

past two decades have witnessed the widespread adoption of problem-based curricula or integrated curricula (Brauer and Ferguson 2015). A content-by-process matrix dovetails nicely with either of these approaches to curriculum design. Figure 4 presents a blueprint where the rows of the matrix identify patient conditions, while the columns list major immunology concepts. The check marks make explicit what condition-concept pairings will be covered on the assessment. The numerals in the bottom row and last column indicate the number of test items for each topic and for each patient condition. A content-by-process matrix like that in Figure 4 offers considerable flexibility for test design and can be used across the curriculum. A common variation of this design is to replace either the rows or columns with the cognitive levels from Bloom's taxonomy (e.g. knowledge, application, analysis).

Delineate the assessment objectives

The outcome of the previous stage is a list of the major content and/or process categories; this second stage introduces detail to the documentation. Test blueprints should describe what is expected of students by listing specific, low-inference behaviors (Mookherjee et al. 2013). Low-inference behaviors are sufficiently observable to objectively determine whether the student demonstrated the behaviors of interest. Sometimes the required detail will already exist as part of the course objectives. A blueprint for clinical skills assessment developed by Mookherjee et al. (2013) relied on milestones from the medical school's existing clinical skills curriculum. They produced their blueprint by mapping each milestone to the appropriate category in the ACGME competency framework. The CanMEDS professional roles also can be used as the basis for classroom test blueprints, as can the entrustable professional activities (EPAs) developed by medical schools and specialty societies. Some certifying agencies also publish test blueprints that contain detailed objectives which can be adapted for local assessment purposes (Gaffas et al. 2012).

If detailed documents such as milestones and learning outcomes are not readily available, then it may be necessary to write the assessment objectives. Well-written assessment objectives are similar to the behavioral objectives in Figure 1. They specify the content to be mastered and the type of knowledge or skill that the student is expected to demonstrate. Another approach is to rely on a matrix blueprint to specify the assessment objectives. Figure 5 presents the skeleton of a content-by-process blueprint for a clerkship exam on cardiovascular medicine. The rows indicate the content to be mastered, while the columns indicate the skills. The cells of the matrix represent the integration of the two, and each cell can be interpreted as

Primary Framework: Patient Condition	Secondary Framework: Immunology Topic							# Questions
	Structure of the immune system	The innate immune system	Humeral immunity	Cell mediated immunity	Hypersensitivity reactions	Transplantation	Systemic disorders affecting immune function	
x-linked agammaglobulinemia	✓		✓					2
adenosine deaminase deficiency	✓		✓	✓				3
DiGeorge syndrome	✓			✓				2
acquired immune deficiency syndrome	✓						✓	2
Chediak-Higashi syndrome		✓						1
leukocyte adhesion deficiency		✓						1
acute cellular rejection						✓		1
asthma					✓			1
Goodpasture's syndrome					✓		✓	2
post-infectious glomerulonephritis					✓			1
poison ivy					✓			1
asplenia	✓	✓	✓					3
hereditary angioedema	✓	✓						2
complement c8 deficiency	✓	✓						2
systemic lupus erythematosus							✓	1
# Questions	7	5	3	2	4	1	3	**25**

Figure 4. Sample problem-based blueprint for a test in immunology illustrating a variation on the content-by-process matrix.

Patient Condition	Competency Domain							# Test Items
	Medical Knowledge		Patient Care				Communication Skills	
	A,P&P	Other Basic Science	Data Gathering	Diagnosis	Medical Mgmt	Surgical Mgmt		
Routine office visit	2-4	1-3	3-5	2-4	0	0	2-4	14-16
Cardiomyopathies *dilated* *hypertrophic* *arrhythmogenic*								9-11
Conduction disorders	C.8			C.8				
Congenital Heart Disease								
Heart Failure								
Hypertensive Disease								
Ischemic Heart Disease								
Vascular Disease								
Valvular Disease								
Other Conditions								3-6
# Test Items	9-11	9-11	14-16	23-27	18-22	4-6	15	100

It is desirable to list specific conditions under each category, as done with Cardiomyopathies .

Cell values sometimes specify the number of test items. Alternatively, the cells can:

- Be left blank (ie, the number of items are specified only in the bottom row and last column).
- Specify the assessment format (MCQ, essay, OSCE).
- Refer to specific assessment objectives contained in another document. An objective can link to multiple cells as with objective C.8:

  Explain the normal sequence in the depolarization and repolarization of the heart and identify their representation on an ECG lead.

Figure 5. Content-by-process matrix for a test on the cardiovascular system. The content dimension consists of classes of cardiovascular conditions, while the process dimension refers to a subset of ACGME competencies.

an assessment objective (e.g. diagnose dilated cardiomyopathies).

Figure 5 includes other features worthy of comment. First, note the level of detail under cardiomyopathies. Ideally, each class of cardiac disorders would include specific conditions as appropriate. Second, the values in the right column and the bottom row indicate the number of questions allocated to each category. Having a range of questions provides some flexibility when implementing the blueprint. Third, this particular example specifies the number of test items for each cell of the matrix, which is quite common. However, as the Figure indicates, the cells can be used in other ways as well (e.g. linked to course objectives). Finally, the level of granularity is important to consider. Too much detail and one can spend hours developing the test blueprint, with the benefit that it will be easier to produce assessment tasks and to assemble a balanced test. Too little detail and the test may be unfocused, and students will not know what to expect. Most authors lean toward greater specificity in test blueprints (e.g. Coderre et al. 2009; Fives and DiDonato-Barnes 2013; Mookherjee et al. 2013).

Decide on the assessment format

Choosing an assessment format is a matter of matching the method of assessment with the claims to be made about what students know and can do. Assessment methods fall into three general classes: written assessments, simulations, and workplace-based assessment. Several articles and reference books summarize the benefits and limitations of over a dozen assessment formats (e.g. Shumway and Harden 2003; Schuwirth and van der Vleuten 2004; Epstein 2007; Downing and Yudkowsky 2009; Lane et al. 2016). The choice of an assessment format will be influenced by validity and reliability concerns, by resources and logistics, and by the context of any particular assessment relative to other assessment opportunities.

Validity is certainly an important factor to consider when deciding on a format. The two types of validity evidence most relevant to the development of test blueprints are content validity and response process (Tavakol and Dennick 2017). It is widely recognized that a test blueprint serves as a primary source of content-related evidence (Kane 2016; Raymond 2016). This is because a thoughtfully developed test blueprint can help ensure that the assessment aligns with content covered during instruction (Notar et al. 2004; McLaughlin et al. 2005; Fives and DiDonato-Barnes 2013). Response process validity refers to the extent to which the cognitive, psychomotor, and affective processes elicited by the assessment tasks are similar to the processes implied by the claims to be made about student behaviors. For example, an MCQ could support the claim that “the student is able to interpret physical examination and chest X-ray to determine the need for thoracentesis”. However, it would take a very clever MCQ to support the claim that a student can actually perform thoracentesis. This latter objective would require a format involving direct observation of the student interacting with a real or simulated patient. Miller’s pyramid can be very useful at this stage for matching an assessment format to the behaviors of interest (Figure 2). Reliability also will influence the choice of an assessment format, with MCQs generally producing more reliable scores than simulations or workplace-based assessments. Cost, logistics, and other practical constraints will most certainly influence the choice of an assessment format. Multi-station simulations may enhance validity but may exceed a medical school’s capacity in terms of staff support or physical space.

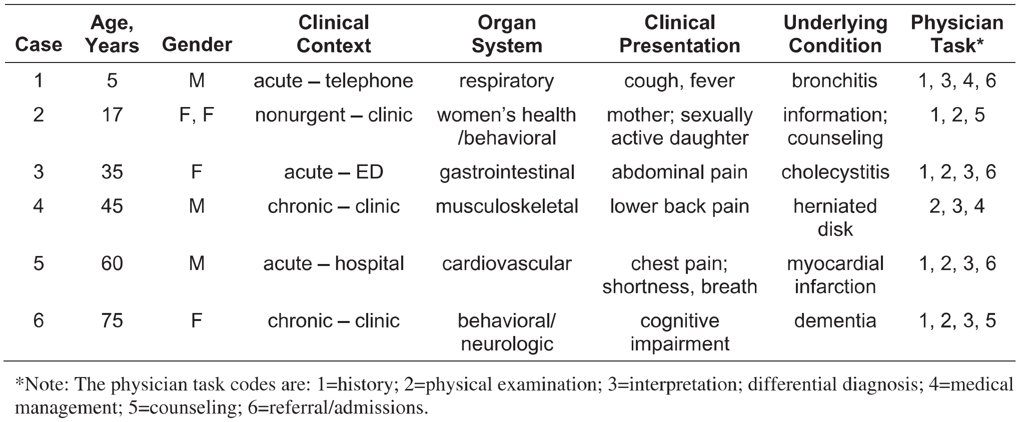
Without question, the assessment landscape in medical education has evolved over the past few decades. Clinical simulations and workplace-based assessments are becoming more common, while written assessments are being limited to those domains for which they are most effective. Although it is common to associate test blueprints with written tests, they also have a role in simulations and workplace-based assessments. Indeed, test blueprints can be even more important for clerkships and residencies where the curriculum is dependent on the idiosyncrasies of the clinical setting. In such instances, test blueprints can help document the curriculum. Excellent examples of clerkship blueprints can be found elsewhere (McLaughlin et al. 2005; Coderre et al. 2009; Mookherjee et al. 2013).

Structured clinical simulations, such as OSCEs, afford greater opportunity than workplace-based assessments to control test content. To ensure that OSCEs are relevant and remain balanced across different student cohorts, test blueprints should specify those case characteristics most likely to affect student performance. In theory, an OSCE blueprint could consist of multiple factors, such as patient age, gender, medical condition, and type of patient management; this would produce hundreds of cells in a multidimensional matrix. In practice, a simple table may suffice. Figure 6 illustrates a blueprint for an OSCE. This blueprint describes the cases and physician tasks for a single test form. To generalize to additional test forms, it would be desirable to document additional content constraints for each column. For example, the constraints for patient age could indicate that each OSCE includes one infant, one adolescent, and three adults, while the constraints for respiratory cases might specify that the OSCE be limited to asthma, bronchitis, or pneumonia.

Specify the category weights

Testing time is limited. As a practical matter, it is necessary to allocate time and space to the different assessment objectives through the use of content weights or category weights. For written assessments, the weights correspond to the number or percent of test items for each category. For simulations and workplace assessments, the weights more likely translate to the amount of testing time. One challenge when assigning category weights is that the number of assessment objectives usually outweighs the available testing time. The domain sampling model speaks to this challenge; it is based on the principle that any test represents a sample of behaviors from the larger knowledge and skill domains of interest (Tavakol and Dennick 2017). The test blueprint is the primary tool for executing that sampling plan by indicating how much content to sample from each domain.

Category weights reflect the importance of the topics within a domain (Millman and Greene 1989). Importance might correspond to the instructional time devoted to a



Case	Age, Years	Gender	Clinical Context	Organ System	Clinical Presentation	Underlying Condition	Physician Task*
1	5	M	acute – telephone	respiratory	cough, fever	bronchitis	1, 3, 4, 6
2	17	F, F	nonurgent – clinic	women's health /behavioral	mother; sexually active daughter	information; counseling	1, 2, 5
3	35	F	acute – ED	gastrointestinal	abdominal pain	cholecystitis	1, 2, 3, 6
4	45	M	chronic – clinic	musculoskeletal	lower back pain	herniated disk	2, 3, 4
5	60	M	acute – hospital	cardiovascular	chest pain; shortness, breath	myocardial infarction	1, 2, 3, 6
6	75	F	chronic – clinic	behavioral/ neurologic	cognitive impairment	dementia	1, 2, 3, 5

*Note: The physician task codes are: 1=history; 2=physical examination; 3=interpretation; differential diagnosis; 4=medical management; 5=counseling; 6=referral/admissions.

Figure 6. Condensed blueprint for a specific test form of an objective structured clinical examination (OSCE). To be useful for assembling alternate test forms, the blueprint would be supplemented by background documents specifying constraints for each column (e.g. proportion of male versus female cases, allowable conditions).

topic; how often it is applied in practice; or the criticality of a topic for subsequent learning. Category weights can be derived from national data reporting the incidence of various medical conditions and procedures (Boulet et al. 2003; Baker et al. 2017). Alternatively, one can survey colleagues such as faculty, residents, or students to determine topic importance. For example, some studies have empirically defined importance as a joint function of case frequency and its urgency or clinical impact, and relied on colleagues to provide judgments of frequency and impact (McLaughlin et al. 2005; Coderre at al. 2009; Patil et al. 2015).

Although these rigorous empirical approaches to deriving weights are admirable, less demanding methods also are suitable for classroom tests. Two effective strategies are the top-down and the bottom-up methods (Raymond 2016), both of which can be applied by a single instructor, or by including colleagues and advanced medical students. The top-down method involves the assignment of percentages to each major category in the blueprint such that the percentages sum to 100%. Weights can be obtained in a similar fashion for subcategories, if desirable. The bottom-up method requires specifying numbers of items, rather than percentages; and that numbers first be assigned at a lower level of the test blueprint (e.g. at the subcategory or specific objective). One challenge with the bottom-up approach is that the total number of items may exceed the maximum feasible test length; consequently, some adjustment to the initial weights is often necessary.

Summary and concluding comments

A test blueprint articulates the knowledge and skill domains to be covered by a test. It also describes other features, such as the emphasis allocated to each content category; the demands of the assessment tasks in terms of cognitive, affective and psychomotor processes; and the assessment format. There is no best way to construct a test blueprint. They vary in terms of content, organization, format, and granularity; and they can be adapted to meet various assessment needs.

Test blueprints assure that the content of a test aligns with the curriculum (Notar et al. 2004); this is a critical aspect of validity (Tavakol and Dennick 2017). Test blueprints support content validation in other ways, by helping to ensure that scores on a specific test generalize to the larger domain of interest (Kane 2016). Test blueprints also provide a framework for evaluating the validity of response processes. For example, performing a thoracentesis is a procedural skill that is located toward the apex of Miller's pyramid and falls into the psychomotor domain of Bloom's taxonomy. An assessment that requires a student to respond to an MCQ probably will not support the claim that the student is competent at thoracentesis because MCQs do not elicit responses that require psychomotor processes. Well-developed test blueprints make explicit the student processes targeted by an assessment, and those targeted processes will inevitably point away from some assessment formats while pointing toward others.

Test blueprints have many practical uses beyond their service to validity, including the following:

- They indicate what instructors value and expect of their students, and can be used as a study guide. Although there are pros and cons to sharing blueprint with students, studies support this practice (McLaughlin et al. 2005; Patil et al. 2015).
- The content categories and competency domains included on test blueprints provide the basis for feedback to students.
- They facilitate the development of assessment-related materials. For example, it is straightforward to transform a test blueprint into scoring rubrics and feedback reports for simulations and workplace-based assessments (Mookherjee et al. 2013).
- Test blueprints are essential for organizing departmental item-writing and review efforts because they succinctly communicate item-writing assignments.
- They provide metadata for managing test materials. Once test items have been coded according to a test blueprint, it is straightforward to retrieve them from a larger pool of items and to assemble them into test forms for different purposes.
- Test blueprints contribute to educational quality improvement. The categories used for student feedback also provide faculty with measures of instructional effectiveness. Test blueprints clarify the connections between planning, instruction, and assessment, which can inspire faculty self-reflection (McLaughlin et al. 2005).

Instruction and assessment have evolved to the point that medical educators now wish to make claims about student competence that extend far beyond the reach of MCQs and the domain of medical knowledge; it is therefore vital to employ assessment methods capable of eliciting the behaviors specified by these new claims. Deciding what to assess and how depends not only on what is important for a particular course, but also on how that course and other assessments fit into the entire curriculum (van der Vleuten and Schuwirth 2005). By documenting the knowledge and skills addressed by each assessment, test blueprints also serve as a tool to facilitate sound curriculum design.

Acknowledgments

The authors wish to thank the National Board of Medical Examiners and the Mayo Clinic College of Medicine and Science for their support of this work. The viewpoints expressed in this article are those of the authors and do not reflect the opinions of their employers.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Notes on Contributors

Mark R. Raymond, Ph.D., is a Research Director and Principal Assessment Scientist at the National Board of Medical Examiners in Philadelphia, PA, USA. For 30 years Dr. Raymond has worked for and consulted with licensing agencies, professional associations, and universities on assessment activities ranging from item-writing workshops to standard-setting studies.

Joseph P. Grande, M.D., Ph.D., is a Professor of Laboratory Medicine and Pathology at the Mayo Clinic College of Medicine and Science in Rochester, Minnesota. He has worked as a renal pathologist and medical educator for nearly 30 years, and served as Associate Dean for Academic Affairs at Mayo from 2007–2013.

References

Al-Eraky M, Marei H. 2016. A fresh look at Miller's pyramid: assessment at the 'Is' and 'Do' levels. Med Educ. 50:1253–1257.

Anderson L, Krathwohl DR. 2001. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York (NY): Longman.

Angus S, Vu TR, Halvorsen AJ, Aiyer M, McKown K, Chmielewski AF, McDonald FS. 2014. What skills should new internal medicine interns have in July? A national survey of internal medicine residency program directors. Acad Med. 89:432–435.

Baker AJ, Raymond MR, Haist SA, Boulet JR. 2017. Using national health care databases and problem-based practice analysis to inform integrated curriculum development. Acad Med. 92:448–454.

Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. 2002. General competencies and accreditation in graduate medical education. Health Aff (Millwood). 21:103–111.

Bloom BS, editor. 1956. Taxonomy of educational objectives, handbook 1: cognitive domain. New York (NY): McKay.

Boulet JR, Gimpel JR, Errichetti AM, Meoli FG. 2003. Using national medical care survey data to validate examination content on a performance-based clinical skills assessment for osteopathic physicians. J Am Osteopath Assoc. 103:225–231.

Brauer DG, Ferguson KJ. 2015. The integrated curriculum in medical education: AMEE Guide No. 96. Med Teach. 37:312–322.

Downing SM, Yudkowsky R. 2009. Assessment in health professions education. New York (NY): Routledge.

Cizek GJ, Webb LC, Kalohn JC. 1995. The use of cognitive taxonomies in licensure and certification test development: reasonable or customary? Eval Health Prof. 18:77–91.

Coderre S, Woloschuk W, McLaughlin K. 2009. Twelve tips for blueprinting. Med Teach. 31:322–324.

Epstein RM. 2007. Assessment in medical education. N Engl J Med. 356:387–396.

Fives H, DiDonato-Barnes N. 2013. Classroom test construction: the power of a table of specifications. Pract Assess Res Eval. 18:1–7.

Frank JR, Snell L, Sherbino J, editors. 2015. CanMEDS 2015 physician competency framework. Ottawa (ON): Royal College of Physicians and Surgeons of Canada.

Gaffas EM, Sequeira RP, Namla RA, Al-Harbi KS. 2012. Test blueprints for psychiatry residency in-training written examinations in Riyadh, Saudi Arabia. Adv Med Educ Pract. 24:31–46.

Harden RM. 2002. Learning outcomes and instructional objectives: is there a difference? Med Teach. 24:151–155.

Kane MT. 2016. Validation strategies: delineating and validating proposed interpretations and uses of test scores. J Educ Measure. 50: 1–73.

Lane S, Raymond MR, Haladyna TM, editors. 2016. Handbook of test development. 2nd ed. New York (NY): Routledge.

McLaughlin K, Lemaire J, Coderre S. 2005. Creating a reliable and valid blueprint for the internal medicine clerkship evaluation. Med Teach. 27:544–547.

Miller G. 1990. The assessment of clinical skills/competence/performance. Acad Med. 65:63–67.

Millman J, Greene J. 1989. The specification and development of tests of achievement and ability. Linn RL, editor. Educational measurement. 3rd ed. New York (NY): Macmillan; p. 13–103.

Mislevy RJ, Riconscente MM. 2006. Evidence-centered assessment design. Downing SM, Haladyna TM, editors. Handbook of test development. Mahwah (NJ): Erlbaum; p. 8–11.

Mookherjee S, Chang A, Boscardin CK, Hauer KE. 2013. How to develop a competency-based examination blueprint for longitudinal standardized patient clinical skills assessments. Med Teach. 35:883–890.

Notar CE, Zuelke DC, Wilson JD, Yunker BD. 2004. The table of specifications: insuring accountability in teacher made tests. J Instruct Psychol. 31:115–129.

Patil SY, Gosavi M, Bannur HB, Ratnakar A. 2015. Blueprinting in assessment: a tool to increase the validity of undergraduate written examinations in pathology. Int J App Basic Med Res. 5:76–79.

Patterson F, Ferguson E, Thomas S. 2008. Using job analysis to identify core and specific competencies: implications for selection and recruitment. Med Educ. 42:1195–1204.

Raymond MR. 2016. Job analysis, practice analysis, and the content of credentialing examinations. Lane S, Raymond MR, Haladyna TM, editors. Handbook of test development. 2nd ed. New York (NY): Routledge.

Raymond MR, Mee J, King A, Haist SA, Winward ML. 2011. What new residents do during their initial months of training. Acad Med. 86: 59–62.

Schuwirth LW, van der Vleuten CP. 2004. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 38:974–979.

Shumway JM, Harden RM. 2003. AMEE Guide No. 25: the assessment of learning outcomes for the competent and reflective physician. Med Teach. 25:569–584.

Tavakol M, Dennick R. 2017. The foundations of measurement and assessment in medical education. Med Teach. 39:1010–1015.

Touchie C, Streefkerk C. 2014. Blueprint project: qualifying examinations blueprint and content specifications. Ottawa (ON): Medical Council of Canada.

van der Vleuten C, Schuwirth LW. 2005. Assessing professional competence: from methods to programmes. Med Educ. 39:309–317.

Zhao X, Dowd K, Searcy C. 2012. Assessing statistics and research methodology in the MCAT Exam. Chance. 25:11–17.