14

Concepts in Assessment

Jared Danielson[1] *and Kent Hecker*[2]

[1] *College of Veterinary Medicine, Iowa State University, USA*
[2] *Faculty of Veterinary Medicine, Cumming School of Medicine, University of Calgary, Canada*

Box 14.1: Key messages

- Expect modest differences in terminology among the many communities that write about education and assessment.
- Assessment can occur at many levels within the educational endeavor, including at the student, classroom, and program levels.
- Assessment can serve both formative and summative purposes.
- Assessment can be norm or criterion referenced. Decisions regarding how or whether or not grades are assigned are independent of the quality of the assessment itself.
- Carefully characterizing what your students need to know or be able to do is fundamental to effective assessment. A taxonomy of learning outcomes and skillfully worded objectives can be helpful.
- High-quality assessment adheres to standards/principles of validity, propriety, utility, and feasibility.

Introduction

The language of education and educational assessment is not uniform, owing in large part to the many backgrounds of those engaged in it. For 2014, the Scimago Journal and Country Rank (Scimago Lab, 2015) listed 914 journals in the category of education. Among the top 50 were general education journals representing a variety of perspectives, including educational psychology, child development, learning sciences, educational technology, and higher education, as well as a variety of discipline-specific journals, including engineering, science, economics, sociology, second-language acquisition, mathematics, and medicine. The remaining 864 indexed education journals were similarly diverse. While critical concepts in assessment are common among these many perspectives, there is inevitable variability in context, emphasis, and language. Where possible, in this chapter we will employ terms that are common across domains, using specific examples from the terms used in health professional education where appropriate. However, do not be surprised if after reading

Veterinary Medical Education: A Practical Guide, First Edition. Edited by Jennifer L. Hodgson and Jacquelyn M. Pelzer.

this chapter, you encounter new and different language or terminology in subsequent reading.

To assess something is to examine it in order to characterize or quantify attributes of interest, usually for the purpose of making a decision. There are a number of labels that can be applied to assessment of students, such as *testing*, *evaluation*, or *measurement*. We consider the terms *evaluation* and *assessment* to be synonymous (Scriven, 1991). *Testing* is a specific assessment strategy that involves examinations (as opposed to other assessment strategies such as portfolios). We refer to *measurement* as the specific characteristics of assessment that are most closely tied to the validity of the claims that might be drawn from the assessment, as opposed to other considerations such as utility, feasibility, or propriety.

There are many potential approaches to devising a chapter on assessment. Some assessment texts focus primarily on the role of assessment in the context of an endeavor, such as higher education (Banta *et al.*, 1996; Huba and Freed, 2000). Others focus on the effective use of assessment information (Popham, 2006), or on the process of creating effective assessments (Hopkins, 1998; Secolsky and Denison, 2012). In this chapter we will define and explain key concepts in the context of practical problems that a veterinary educator might face, providing examples where possible. We will frame the chapter in terms of concepts in three general categories: *why* to assess (purposes of assessment), *what* to assess (defining outcomes), and *how* to assess (methodological and practical considerations). We hope that you can use this chapter to guide some specific assessment activities. For instance, we explain the characteristics of effective objectives, in hopes that you will be able to write better objectives after reading that section. However, the primary purpose of the chapter is to provide a conceptual overview for next steps in improving your own assessments, and to provide a foundation for additional reading.

Why: Purposes of Assessment

Levels of Assessment

Assessment can occur at numerous levels of the educational endeavor. At the most fundamental level, we seek to determine what individual students know and/or can do; this constitutes assessment at the *learner* level. Sometimes assessment at the learner level provides insufficient information to answer relevant questions about an educational outcome. Imagine, for instance, that learners in a clinical pharmacology course consistently master what they are taught, but are perennially unprepared to perform common dosage calculations in subsequent courses or clinical experiences. If this were the case, assessment would need to be conducted at the *course level* to determine why students who appeared to be learning the foundational knowledge were subsequently not competent in a key area. Assessment activities at the course level might explore questions such as whether the desired content is being taught at all, whether existing tests are adequately measuring the desired learning outcomes, and/or whether there are adequate opportunities for practice. Some important questions cannot be answered at the student or course level, however. Imagine, for instance, that veterinary students perform well in all of their courses, and that all of their courses prepare them adequately for subsequent courses, but that on graduation they are systematically deficient in a specific area. Such a scenario would reveal the need for assessment at the program level, answering questions such as whether or not knowledge and skills in the deficient area are taught or valued by the institution. One can conceive of important assessment questions at levels other than these three: for instance, it might be appropriate to conduct assessment at the department level, or much more broadly (for example, across a state, province, or country). However, most educational assessments in veterinary medicine can be characterized as being at the level of the learner, course, or program. In this chapter, we will provide concepts that can be applied at any of those levels, although the emphasis will mostly be at the learner level.

Formative versus Summative Assessment

Assessments at any of these levels can be *formative* or *summative* (Bloom, Hasting, and Madaus, 1971). *Formative assessment* is used principally for improvement. At the student level, a scored but ungraded test or other practice activity can provide students with the information they need to make improvements in knowledge or skills, and can help the instructor know what knowledge or skills to emphasize. At the course level, an instructor might ask students to respond to informal or unofficial course or faculty evaluations, or invite a colleague to observe and provide informal input. At the program level, a veterinary school department chair or dean might ask a group of colleagues to conduct a mock accreditation site visit. In all of these instances, the primary purpose of the assessment is to offer the entity being assessed a low- or no-risk opportunity to see how it is doing, and where improvements might be needed. *Summative assessment*, in contrast, is used to inform a decision regarding the future of the entity being assessed, and usually involves important stakes. At the student level, a graded examination is summative, because it results in a score or grade point average that might influence important decisions such as the student's ability to move forward in the curriculum or be placed in a residency program. At the classroom level, faculty and course assessments are summative whenever they are used to make decisions such as whether or not the instructor will continue to be invited to teach, obtain tenure, or receive a raise. Official accreditation site visits and reviews of annual reports to the accrediting body are summative assessments at the program or institution level, because they influence whether or not the institution will enjoy the privileges of accreditation.

Sometimes the terms formative and summative are used as if they were inherent characteristics of assessment instruments or tools. However, the key attribute that distinguishes a formative assessment from a summative one is the purpose for which it is used. The exact same exam is a formative assessment if used as a practice exam, and a summative assessment if used as a final exam. The term *feedback* is occasionally substituted in casual conversation for *formative assessment*, because the primary purpose of feedback is to provide information to the entity being assessed. However, both formative and summative assessments can provide valuable feedback to learners, teachers, and programs.

Grades, Normative versus Criterion-Referenced Assessment, and Rigor

Many faculty assess students, at least in part, so that they can assign a grade. Similarly, some students seem to engage in assessment activities primarily so that they can receive a grade (a good one, they hope). Grades in turn can be reported to external stakeholders, such as directors of residency programs, future graduate programs, or employers, who might use them as one source of evidence regarding what the student has learned or is likely to be able to achieve. Grades and grade point averages can be helpful because they provide a relatively concise mechanism for communicating what a student has learned or can do, relative to others who participated in a similar experience (see Box 14.2). Grades also communicate information about other kinds of issues, such as how good students are at anticipating what professors will ask on tests, how hard they work, how much extra credit work they complete, or whether or not they have unexcused absences. Assessment is not the same as grading. Assessment determines what students know or can do, and grading documents merit or rank in an educational setting. Instructors and institutions can use assessment information for a variety of important purposes, including deciding what to teach next/more/better/less, and when/how/to whom to provide remediation, as well as to assign grades.

Assessments are either *norm referenced* or *criterion referenced*, depending on whether they compare examinees to each other, or to an established standard (Glaser, 1963). Assessments

Box 14.2: Focus on assumptions about grades

Grades are not an inevitable characteristic of assessment, nor do they have to coincide with individual courses. A number of medical education programs use two-interval (pass/fail) grading (Spring *et al.*, 2011) rather than conventional letter grading. Furthermore, some medical programs employ progress testing (Finucane *et al.*, 2010; van der Vleuten, Verwijnen, and Wifnen, 1996), in which assessment occurs at the institution level and not the course level. Available studies suggest that such approaches improve student wellbeing without having an adverse impact on learning (Spring *et al.*, 2011). These strategies focus on ensuring that all students achieve an acceptable level of competence, without excessive concern for sorting or ranking. While students are not assigned a grade, they are, of course, assessed all the same. Determining what students know is essential, whether that knowledge is associated with a letter grade or not.

that are intended to compare examinees to each other are referred to as *norm-referenced* tests, because they reference a group *norm* or average. Some assessment purposes are best served by referencing norms. For example, placement exams like the Graduate Record Examination (GRE) are used by academic programs to select students when filling a limited number of seats. In such cases, programs will admit a full cohort whether those students are remarkably capable academically, or are barely able to meet the minimum standard. Decision-makers seek to admit the most qualified applicants from the available pool.

In other cases, assessment serves to compare examinees to an established standard. Such assessments are *criterion referenced*, because the standard of success references one or more established criteria. Licensing exams are criterion referenced: test designers seek to establish the minimum knowledge and skills required of a competent practitioner in a specific discipline, and to determine whether or not each examinee is minimally competent. Hopefully, all competent examinees will pass the test, and all incompetent examinees will fail it. (For those of you who are suddenly getting flashbacks to your Public Health course, educational assessment is indeed similar to other kinds of assessment, and while educational assessment folks do not use those terms, this use of criterion-referenced assessments is similar to the use of diagnostic tests in veterinary medicine, and the concepts of sensitivity and specificity certainly apply.)

Bear in mind that it is the inference that is made from the test, and not the test itself, that is either norm or criterion referenced. However, as noted by W. James Popham:

> More often than not, … if an educational measuring instrument is built to provide norm-referenced inferences, it usually does *not* do a good job in providing criterion referenced interpretations. And, if an educational measuring instrument is originally built to provide criterion-referenced inferences, it usually does *not* do a good job in providing norm-referenced interpretations. (Popham, 2006, p. 35)

We suggest that in most educational settings, it makes more sense to design tests to support criterion-referenced rather than norm-referenced inferences. After all, interested stakeholders, including the students themselves and the public at large, have a greater interest in knowing how well each individual's performance matches a relevant standard than how well it matches the performance of other students.

Rigor is not a precise assessment term such as *validity* or *accuracy*, but is often used in lay language to reference how challenging or difficult an assessment is, with the assumption that a rigorous assessment is able to separate the truly prepared from the more casually prepared. Of course, designing an assessment that does clearly differentiate those who excel from those who do not is helpful. However, instructors are advised not simply to rely on statistical manipulation of test scores to create the appearance of rigor. Some instructors assume that an exam will automatically be rigorous if the grades are normally distributed, with scores being assigned automatically to grades based on where those scores fall in the distribution. Similarly, some instructors assume that the lower the average on a test, the more rigorous the assessment process. However, rigor is not a function of score distribution or grading scale. Rather, a "rigorous" assessment provides valid, useful, sufficiently comprehensive information about what students know, or are able to do with what they have learned. There is nothing inherently wrong with grades being normally distributed, but faculty who are primarily interested in a normal distribution of scores might just as well assess students based on their height, weight, or body temperature. As noted by Bloom, Hastings, and Madaus (1971, p. 45), "There is nothing sacred about the normal curve. It is the distribution most appropriate to chance and random activity. Education is a purposeful activity, and we seek to have the students learn what we have to teach."

Describing What to Measure

If it is important to establish the criteria by which learners will be assessed, it is first necessary to characterize the outcomes of learning accurately and precisely. Taxonomies of learning facilitate this purpose. Just as veterinarians and animal scientists use taxonomies of organisms to learn rules that apply to certain orders, families, or species of animals, and not to others, educational practitioners learn rules that apply to certain types of learning outcomes, and not to others. Most educators are familiar, conceptually, with broad types of learning outcomes; they often hear messages to teach "higher-order" rather than "lower-order" thinking, or to emphasize "critical thinking" rather than "rote memorization." Such ideas can be useful reminders that we would like our students to be able to do things that are important or meaningful, but they are not sufficiently precise to guide specific assessment or teaching decisions. Taxonomies of learning outcomes are intended to help provide more precision. Unfortunately for educators, there is no universally accepted taxonomy of learning outcomes, so it is not uncommon to become familiar with one taxonomy, only to find that colleagues are familiar with another (Alexander, Schallert, and Hare, 1991; de Jong and Ferguson-Hessler, 1996). We will introduce two common taxonomies in hopes that you can apply the principles discussed regardless of the taxonomy you happen to be employing at any particular time.

Bloom's taxonomy, perhaps the best-known classification scheme for learning outcomes, defines three domains of knowledge: *cognitive* (having to do with intellectual or thinking/reasoning skills), *affective* (having to do with attitudes, beliefs, and motivation), and *psychomotor* (having to do with the ability to manipulate tools, objects, or one's own body; Bloom *et al*., 1956). The cognitive domain, as found in the 2001 revision of Bloom's taxonomy (Anderson *et al*., 2001), conceptualizes learning outcomes across two dimensions: the cognitive process dimension, which includes *remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create*; and the knowledge dimension, which includes *factual knowledge*, *conceptual knowledge*, *procedural knowledge*, and *metacognitive knowledge*. In this framework, instructional outcomes are classified in terms of both the cognitive process and the knowledge dimension being targeted. This approach is flexible and powerful, but requires more than a casual familiarity with the specific terms and concepts of the taxonomy. Therefore, for veterinary educators with an interest in clearly and effectively classifying educational outcomes, Bloom's taxonomy can provide a valuable approach, but a thorough discussion of it is beyond the scope of this chapter.

Other categorizations of learning outcomes have also been created to facilitate instructional design and evaluation. Among those, Gagné, Briggs, and Wager's (1992) taxonomy has been particularly influential among practitioners who design/evaluate instruction. Smith and Ragan's (2005) adaptation of Gagné's framework is often used to train beginning instructional designers, and is useful because it is relatively concise and involves all three knowledge domains (cognitive, psychomotor, and affective). In Table 14.1, the left-hand columns provide Smith and Ragan's categorization of learned capabilities; the middle column offers a brief definition with examples from a veterinary context; and the right-hand column contains appropriate assessment approaches for the capability. The latter are not meant to be limiting, but to provide the reader with the approaches that most commonly and best match the learned capability. Furthermore, assessment of lower-order knowledge/skills can often be embedded into assessment of higher-order knowledge/skills.

There are at least two important problems that can arise from a fundamental misunderstanding of, or inattention to, kinds of learning outcomes:

- *Outcomes snobbery, and a resulting sloppiness in measurement.* It can be tempting to want to appear to "keep up with the Joneses" when it comes to learning outcomes. Instructors want to be seen as teaching important things like problem-solving or "critical thinking," as opposed to "lesser" outcomes like facts or terms. However, this impulse can be counterproductive. The purpose of using a taxonomy of learning outcomes is not to identify and favor certain learning outcomes while avoiding others, but accurately to characterize what needs to be learned in any given context. "Lower-order" outcomes are often fundamental to more advanced ones. Facts and terms are important building blocks of a discipline; it is not possible to learn advanced concepts, principles, and so forth without them.
- *Mismatching desired outcomes and assessments.* It is remarkably common for educators to create assessments that seem relatively unrelated to their desired learning outcomes. For instance, an instructor might claim to be testing problem-solving (or "critical thinking") when the test questions themselves primarily measure students' ability to remember definitions of terms. Often such misalignment goes quite unnoticed by the instructor, for lack of having carefully defined the desired learning outcomes in the first place. Considering where the desired learning outcome fits in the context of a taxonomy of learning outcomes can help to prevent this problem.

In addition to general taxonomies of learning outcomes, other frameworks have been suggested for identifying important types of learning outcomes in specific fields. One framework that has gained popularity in medical education is Miller's Framework for Clinical Assessment (Miller, 1990). Miller conceived of this framework in response to the need to document clinical proficiency and the knowledge related to it as it develops during the transformation of novice into independent practitioner. He defined ability at the lowest ("Knows") level of his framework as knowing "what is required in order to carry out … professional functions effectively." The second level ("Knows How") involves "the skill of acquiring information from a variety of human and laboratory sources, to analyze and interpret these data, and finally to translate such findings into a rational diagnostic or management plan." The "Shows How" level involves a student actually demonstrating the ability "when faced with a patient." The "Does" level aims to "predict what a graduate does when functioning independently in a clinical practice" (p. S63). Not intended to be a complete taxonomy of learning outcomes, Miller's framework provides a useful approach for conceptualizing outcomes specific to the task of becoming a medical practitioner. Ultimately,

Table 14.1 Smith and Ragan's adaptation of Gagné's learned capabilities.

Learned capability		Definition with veterinary example	Common assessment approach
Intellectual skills	Problem-solving	Applying known principles or other knowledge/skills to address previously unencountered problems. In veterinary medicine, we often think of diagnostic and clinical decision making as examples of problem-solving.	Examinee solves an authentic problem. Responses can vary from selecting an option in an extended matching or script concordance test, to providing descriptions, differential diagnosis lists, problem lists, treatment protocols, etc. Testing environments can include classroom, clinical, and computer-based settings.
	Procedure	Learning and performing the appropriate ordered steps in a task. Drawing blood, making a blood smear, performing a physical exam, conducting a medical interview, and countless other veterinary tasks all involve procedures.	Examinee performs the procedure while examiner uses an observation protocol to assess the performance.
	Principle	Principles are also called rules, and describe the relationships among concepts or ideas. Understanding mechanisms of disease or health, the action of drugs, and so forth involves principles.	Examinee explains or applies principles in response to open-ended or carefully written multiple-choice items.
	Defined concept	The ability to categorize accurately based on defined or theoretical attributes. Examples of defined concepts are innumerable, and include things such as specific diseases, conditions such as stress or anxiety, and so forth.	Examinees are shown examples and close nonexamples of the target concept(s) and asked to identify the accurate examples, through either multiple-choice or short-answer questions.
	Concrete concept	The ability to categorize accurately based on concrete attributes such as size, shape, and color. Examples are endless, such as surgical instruments, species of animals, and body condition scores (by appearance).	
	Discrimination	Ability to perceive that two things match or do not. For instance, students learning cytology must be able to distinguish subtle differences between the appearance of normal and abnormal cells; students learning abdominal palpation must be able to detect subtle differences in size, firmness, or location of the structures being palpated.	Similar to concepts, examinees choose between examples and nonexamples.

(Continued)

Table 14.1 (Continued)

Learned capability	Definition with veterinary example	Common assessment approach
Declarative knowledge	Knowing "that" something is. Being able to recite information from memory such as facts or labels. There is no assumption of understanding embedded meaning.	Rote production (written or spoken) is necessary to ensure mastery, though multiple-choice-type items are sufficient to ensure recognition.
Cognitive strategies	Strategies that learners/problem solvers employ to manage their own learning/thinking. Rehearsing what one has learned, tying new knowledge to prior knowledge, and so forth.	Not typically assessed in higher education environments; assumed to translate into mastery of other learned outcomes.
Attitudes	Attitudes are affective learning outcomes. An attitude is a mental state that influences what we choose to do. Attitudes have cognitive and behavioral components. When we are tasked with measuring things like "ethical behavior" or "professionalism," we are often trying to measure attitudes.	Attitudinal surveys or observation of behaviors that demonstrate the desired outcome in the absence of a threat for noncompliance.
Psychomotor skills	"Coordinated muscular movements that are typified by smoothness and precise timing" (Smith and Ragan, 2005, p. 82). Many veterinary tasks have psychomotor components, including countless surgical, clinical, or diagnostic techniques such as knot tying, making a blood smear, venipuncture, instrument handling during surgery, and so forth.	Examinee performs the procedure while examiner uses an observation protocol to assess the performance.

that is the purpose of any taxonomy of learning outcomes: to allow the educator to conceptualize specific learning outcomes in ways that facilitate learning and assessment.

Objectives

Critical to defining learning outcomes is creating useful outcomes statements, most commonly and broadly referred to as *objectives* (Anderson *et al.*, 2001). You have likely heard a variety of labels substituted for objectives, some of which include goal, outcome, proficiency, competency, entrustable professional activity, content standard, curricular aim, performance standard, and academic achievement standard (Anderson *et al.*, 2001; Flynn *et al.*, 2014; Popham, 2006). The term objective itself is sometimes qualified with terms such as *instructional*, *educational*, *global*, or *program* (Anderson *et al.*, 2001). The Council on Education (COE) of the American Veterinary Medical Association (AVMA) refers to its nine clinical proficiency outcomes areas as *competencies* (AVMA COE, 2014); similarly, the Royal College of Veterinary Surgeons (RCVS) refers to its required outcomes of veterinary programs as *Day One Competencies* (RCVS, 2014). The Association of American Medical Colleges (AAMC) recently abandoned the term competency for the required outcomes to enter residencies in favor of *core entrustable professional activity (EPA)*, arguing, among other things, that doing so allowed for framing those