Validity, Reliability, and Defensibility of Assessments in Veterinary Education

Kent Hecker ■ Claudio Violato

ABSTRACT
In this article, we provide an introduction to and overview of issues of validity, reliability, and defensibility related to measurement of student performance in veterinary medical education. Validity has to do with the extent to which the instrument measures whatever it is supposed to measure, reliability has to do with the consistency of measurement, and defensibility deals with the appropriate standards of the assessment procedure. An explanation of the methods that can be used to determine reliability and validity are given and examples of how they have been used in recent research findings are provided. Veterinary educators should have an understanding of each of these concepts because evaluation of veterinary students and performance measures provides evidence of utility of assessment tools, informs teaching practices, and can guide curriculum development and revision.

Key words: reliability, validity, defensibility, assessment, veterinary education

INTRODUCTION
Students in veterinary schools are extensively tested and assessed. From admissions to licensure examinations, assessments include interviewer scores of veterinary school applicants; multiple-choice and essay examinations for demonstration of knowledge and application of knowledge; objective structured clinical examinations (OSCEs) for demonstration of clinical and professional skills; and in-training evaluation reports and clinical assessment exercises (e.g., mini-clinical evaluation exercise) for demonstration of workplace skills during clinical rotations. The measurement properties—specifically reliability and validity—of these assessment methods are seldom reported in the veterinary education literature. Rhind et al.'s systematic review on assessing competence in veterinary medical education, for example, reported only five papers that provided evidence of evaluation of the assessment method that was used.1

The goals of assessment in health professional education include choosing applicants for selected positions, enhancing student learning, guiding faculty teaching efforts, informing curricular development, and contributing to educational research.2 To achieve these goals, student assessment should not be thought of as separate to the curriculum, but embedded within and an integral part of the curriculum. When reviewing existing or developing new assessment methods for courses or programs, the following questions should be taken into consideration:

1. Where in the curriculum does this assessment fit and what are we trying to assess—knowledge, skills, or attitudes? Curriculum developers and course coordinators should envision each assessment method as part of an overall assessment program that is designed to build on what was taught previously to achieve the goals and objectives of the program.3

2. Should the assessment be formative or summative? The purpose of formative assessments is to deliver feedback for the improvement of the learner. They are used to identify strengths and weaknesses for both the student and instructor, which may alter teaching. Summative assessments are for decisions and judgment. The evaluator is typically the instructor and summative assessments are used to assign grades, judge competence, and determine promotion.

3. Should the assessments be criterion or norm referenced? Criterion-referenced assessments involve determining students’ performances by comparing their abilities with clearly stated learning outcome criteria and standards for levels of performance that are set before the assessment. There is no predetermined grade distribution when using criterion-referenced assessments; theoretically, everyone could pass or everyone could fail depending on where the standards have been set. Norm-referenced assessments provide grades based on student ranking within a particular cohort. This method involves fitting a ranked list of student scores to a predetermined distribution for awarding grades.4

Underpinning these questions and the use of assessment methods are issues of reliability, validity, and defensibility. To assist veterinary educators in determining the reliability, defensibility, and validity of the tools that they have created or are using, we provide an introduction to each of these areas as well as some practical examples of what analyses to run given the questions that are asked.

IMPORTANT CHARACTERISTICS OF MEASUREMENT INSTRUMENTS
There are three critical elements of any assessment procedure or measurement instrument: validity, reliability, and defensibility. Validity has to do with the extent to which the instrument measures whatever it is supposed to measure; that is, validity focuses on the question of how
well an assessment carries out its intended function. Reliability has to do with the consistency of measurement, while defensibility deals with the appropriate standards of the assessment procedure.

Other characteristics that can be taken into consideration when evaluating an assessment procedure or assessment program include utility, acceptability, educational impact, and cost.\textsuperscript{3, 5} These criteria are more subjective as well as resource- and institution-specific. As van der Vleuten and Schuwirth have stated, including these factors with reliability and validity is intended to convey "that choosing an assessment method inevitably entails compromises and that the type of compromises varies for each specific assessment context."\textsuperscript{3} We will focus specifically on validity, reliability, and defensibility. Readers interested in a more detailed discussion of these other concepts are referred to van der Vleuten.\textsuperscript{3, 5}

### THE NATURE OF VALIDITY

When examining the validity of an instrument, there are four levels of analysis: 1) face validity, 2) content validity, 3) criterion-related validity (predictive or concurrent), and 4) construct validity. Face and content validity can be established without empirical evidence, while criterion-related and construct validity require data for their demonstration (see Table 1).

#### Face Validity

Face validity has to do with appearance—does the test appear to measure whatever it is supposed to measure? From the respondent’s perspective, does the instrument seem to measure the relevant domain? Face validity can be crucial in establishing rapport and motivation, and in determining how seriously the assessment will be taken. Face validity, however, plays a secondary role to content validity.

#### Content Validity

Content validity concerns the extent to which an assessment adequately samples the domain of measurement, the content. A client questionnaire that requires the client to evaluate the veterinarian’s clinical skills and cognitive knowledge would lack content validity because these are not areas in which a client has sufficient expertise.

#### Criterion Validity

Criterion validity involves sampling or selecting. The domain of measurement, therefore, must be clearly defined and detailed. Not only must the content areas or subject matter be identified, but also the cognitive processes involved. The cognitive processes refer to the levels of Bloom’s taxonomy\textsuperscript{6} (knows, comprehends, applies, analyzes, synthesizes, evaluates). Content validity for an examination in a first-year anatomy course would be ensured by a table of specifications where the course content and extent of understanding are identified and the number of questions chosen reflects the emphasis placed on the topics and the overall course objectives.\textsuperscript{5}

#### Criterion-related Validity

When we are interested in how performance on an assessment correlates with performance on some other criterion, we are concerned about criterion-related validity. The criterion may be any performance on some other assessment. There are two subcategories of criterion-related validity: predictive and concurrent. Predictive validity refers to how current performance correlates with some future performance on a criterion, and thus involves the problem of prediction. Concurrent validity refers to how performance correlates concurrently (at the same time) with some criterion. The several assessments employed in the pre-clinical years to measure knowledge should correlate with each other—a measure of concurrent validity. If we assess the students’ performance in the future, we may be interested in whether the pre-clinical measures correlate with future performance during clinical rotations, thus indicating predictive validity.

Since the empirical procedures used in both predictive and concurrent validity are essentially the same (correlation, \( r \)), and since each involve an additional criterion, they are classified together under criterion-related validity.

#### Validity Coefficients

Evaluating the criterion-related validity of an assessment requires examining the magnitude of the correlation coefficient, which is referred to as the validity coefficient. Validity coefficients are merely correlations that are interpreted within the context of validity. Interpretations can be aided further by using the coefficient of determination (\( r^2 \)) and then determining the percentage of variance (a statistical summary of the total differences in

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
<th>Evidence</th>
<th>How to Establish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td>The appearance of the instrument</td>
<td>Nonempirical</td>
<td>Make the instrument appear appropriate</td>
</tr>
<tr>
<td>Content</td>
<td>The adequacy with which the instrument samples the domain of measurement</td>
<td>Nonempirical</td>
<td>Construct a table of specifications</td>
</tr>
<tr>
<td>Criterion</td>
<td>The extent to which the instrument relates to some criterion of importance</td>
<td>Empirical (correlation, regression, exploratory factor analysis, discriminant analysis)</td>
<td>Study the relationship of the scores derived with some criterion of importance</td>
</tr>
<tr>
<td>Construct</td>
<td>The psychological processes that underlie performance on an assessment device</td>
<td>Empirical (correlation, confirmatory factor analysis, structural equation modeling)</td>
<td>Manipulate the test scores through experimental procedures and observe results Study relationships among latent variables (constructs)</td>
</tr>
</tbody>
</table>

### Table 1: Four types of validity

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
<th>Evidence</th>
<th>How to Establish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td>The appearance of the instrument</td>
<td>Nonempirical</td>
<td>Make the instrument appear appropriate</td>
</tr>
<tr>
<td>Content</td>
<td>The adequacy with which the instrument samples the domain of measurement</td>
<td>Nonempirical</td>
<td>Construct a table of specifications</td>
</tr>
<tr>
<td>Criterion</td>
<td>The extent to which the instrument relates to some criterion of importance</td>
<td>Empirical (correlation, regression, exploratory factor analysis, discriminant analysis)</td>
<td>Study the relationship of the scores derived with some criterion of importance</td>
</tr>
<tr>
<td>Construct</td>
<td>The psychological processes that underlie performance on an assessment device</td>
<td>Empirical (correlation, confirmatory factor analysis, structural equation modeling)</td>
<td>Manipulate the test scores through experimental procedures and observe results Study relationships among latent variables (constructs)</td>
</tr>
</tbody>
</table>
the clinical years (multiple analytical scores were the best predictor of performance in performance (multiple scores were found to be predictors of pre-clinical performance. Pre-requisite GPA and GRE analytical sections, overall GPA, and Graduate Records Examination (GRE) scores on the analytical, quantitative, and verbal sections were independent variables. In veterinary school, the dependent variable was grade-point average (GPA) in pre-clinical and clinical years. A recent study examined whether knowledge-based admission requirements predict performance in pre-clinical and clinical years in a veterinary school. The dependent variable was grade-point average (GPA) in veterinary school and the independent variables were Graduate Records Examination (GRE) scores on the analytical, quantitative, and verbal sections, overall GPA, and prerequisite GPA. Pre-requisite GPA and GRE analytical scores were found to be predictors of pre-clinical performance (multiple $R = 0.48, 20.7\%$ of the variance), while GRE analytical score was the best predictor of performance in the clinical years (multiple $R = 0.28; 6.8\%$ of the variance). This finding is consistent with the predictive validity of admissions work in medical schools, where GPA is one of the best predictors for pre-clinical performance, but knowledge-based admissions requirements typically account for a very small amount of variance (weak predictors) in clinical performance.7

### Table 2: Methods of estimating reliability

<table>
<thead>
<tr>
<th>Method</th>
<th>Reliability Measure</th>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest</td>
<td>Stability over time</td>
<td>Give the exact same test to the same group of students at different times (e.g., hours, days, months, years)</td>
</tr>
<tr>
<td>Parallel forms (same time)</td>
<td>Form equivalence</td>
<td>Give two forms of the same test to the same group at the same time</td>
</tr>
<tr>
<td>Parallel forms (different times)</td>
<td>Form equivalence and stability over time</td>
<td>Give two forms of the same test with a time interval between the two tests</td>
</tr>
<tr>
<td>Split-half</td>
<td>Internal consistency</td>
<td>Split the test into odd and even halves and correlate the resulting scores. Adjust with the Spearman–Brown formula</td>
</tr>
<tr>
<td>KR20, Cronbach’s $\alpha$</td>
<td>Internal consistency</td>
<td>Give the test once and apply KR20 or the Cronbach’s $\alpha$ coefficient formula.</td>
</tr>
<tr>
<td>Generalizability theory</td>
<td>–</td>
<td>Apply analysis of variance methods to get variance components</td>
</tr>
</tbody>
</table>


test scores) that is accounted for in the criterion by the assessment. When the correlation is used within the context of predictive validity it is called the predictive validity coefficient. In the context of concurrent validity, it is called the concurrent validity coefficient.

Predictive validity studies can also be performed using regression analysis, where multiple regression is used to determine which variables predict performance in clinical rotations. A recent study examined whether knowledge-based admission requirements predict performance in pre-clinical and clinical years in a veterinary school.7 The dependent variable was grade-point average (GPA) in veterinary school and the independent variables were Graduate Records Examination (GRE) scores on the analytical, quantitative, and verbal sections, overall GPA, and prerequisite GPA. Pre-requisite GPA and GRE analytical scores were found to be predictors of pre-clinical performance (multiple $R = 0.48, 20.7\%$ of the variance), while GRE analytical score was the best predictor of performance in the clinical years (multiple $R = 0.28; 6.8\%$ of the variance). This finding is consistent with the predictive validity of admissions work in medical schools, where GPA is one of the best predictors for pre-clinical performance, but knowledge-based admissions requirements typically account for a very small amount of variance (weak predictors) in clinical performance.7

### Construct Validity

Construct validity focuses on the truth or correctness of a construct and the measurements that test it. What is a construct? A construct is an “entity, process, or event which is itself not observed,”8 but which is proposed to summarize and explain facts, empirical laws, and other data.

In educational and other health sciences measurement, examples of constructs include professionalism, communication skills, empathy, collegiality, and team participation. Construct validity requires the accumulation of information from a variety of sources. In effect, it is a special instance of the general procedure of validating a theory in any scientific endeavor. The ultimate purpose of validation is explanation, understanding, and prediction. For instance, Fuentealba and Hecker assessed the reliability and validity of a standardized clinical rotation evaluation form.9 To determine construct validity, exploratory factor analysis was used to determine if specific questions regarding “clinical skills,” “knowledge,” “professionalism,” and “interpersonal skills” clustered together. Of the 21 items asked in the four categories, it was determined that they clustered in three factors—“professionalism,” “knowledge,” and “clinical skills”—leading the authors to conclude that certain questions in “clinical skills” and “interpersonal skills” were assessing either “professionalism” or “knowledge.” Overall, the analysis provided preliminary evidence of construct validity of the clinical rotations’ assessment form.

In recent years, there have been important measurement and statistical innovations for the further exploration of construct validity. These include the development of structural equation modeling approaches, particularly confirmatory factor analyses.10 The main advantages of these new approaches over more conventional ones such as exploratory factor analysis and multiple regression, for example, is that they can be used to test complex theoretical models specified a priori. Accordingly, they provide much stronger theoretical underpinnings to construct validity than the more traditional exploratory approaches.

### RELIABILITY

Reliability has to do with the consistency of measurement. Specifically, we would like to know that if we were to repeat an examination with the same students under the same conditions, would we get the same results? Reliability is a necessary condition for validity. While reliability is a precondition for validity, it does not guarantee it. That is, an assessment device that is reliable is not necessarily valid. What is meant by the consistency of measurement? What factors can lead to the inconsistency of measurement? How can you tell if a measurement is consistent or not? Reliability is a central problem in assessing the knowledge, attitudes, and skills or clinical competencies of veterinary students. Reliability is a multifaceted concept rather than a singular idea. Several factors can influence reliability, such as examination questions that are unclear, unstandardized test conditions, nervous examinees who misinterpret
questions, length of tests, examination format, and difficulty of the examination. There are several ways of thinking about and discussing reliability.\textsuperscript{4, 11} Four methods of establishing reliability are usually recognized.

**Four Methods for Determining Reliability**

The six methods or techniques for determining the reliability of a measurement instrument are (1) test–retest; (2) parallel forms—given at the same time; (3) parallel forms—given at different times; (4) split-half; (5) Kuder–Richardson formula 20 (KR20), Cronbach’s alpha or \( \alpha \) and (6) generalizability theory (see Table 2).

These techniques are not only different methods of establishing reliability, but each produces a somewhat different type of reliability as well. While all forms of reliability focus on consistency, there are different aspects of the testing to which the consistency is relevant. The consistency on the test–retest method, for example, focuses on time. That is, does the assessment produce consistent results from one measurement to another at a different time?

A second type of reliability is estimated by the parallel forms (also called equivalent forms) method. Here the focus is on the consistency of measurement across different forms (parallel or equivalent) of the same test. Because the different forms can be given either simultaneously or at different times, stability of the assessment over time can also play a role in this method.

The third type of reliability that is estimated both by split-half method and the internal consistency method (e.g., Cronbach’s \( \alpha \)), deals with the internal consistency of the assessment or the consistency of the measurement across different items within the instrument. Thus, neither time nor form equivalence is relevant here. Split-half is a special version of internal consistency. Here we are concerned about the extent to which each item (or group of items) consistently measures in the direction of the other items on the test. Cronbach’s (or coefficient) \( \alpha \) applies to the reliability of how consistently students perform across items within a single examination.\textsuperscript{12}

**Interpreting the Reliability Coefficient**

There are several useful rules of thumb that can help to interpret the reliability coefficient (\( r_{xx} \)). Different standards apply to different assessment situations. For high-stakes tests (for screening, diagnosis, admissions, or licensing), reliability coefficients are generally expected to exceed \( r_{xx} = 0.80 \). This means that 80% of the variation in repeated measurements is due to consistent and stable measurements. The remaining 20% \((1−r_{xx})×100 = 20\%\) is due to errors of measurement. An assessment with a reliability of less than \( r_{xx} = 0.70 \) has 30% of the variance due to errors of measurement.

Different standards, however, must guide the adjudication of assessment instruments, especially when in development. Generally, any assessment—an instrument in development or otherwise—is considered poor if \( r_{xx} \) is less than 0.50. This means that only 50% of the variance in the measurement is consistent and a true measure, while 50% is error. An instrument with \( r_{xx} = 0.40 \) has 60% error of measurement and only 40% consistent measurement. Rating scales, self-report instruments, and multiple-choice tests in development with reliabilities of 0.50–0.60 are adequate, those with reliabilities of 0.61–0.70 are good, and those with reliabilities of 0.71 or more are very good. Excellent is above 0.80.\textsuperscript{4, 11}

**Generalizability Theory**

A comprehensive and fruitful approach to estimating reliability and error variance in situations in which several sources of measurement error are introduced is generalizability theory.\textsuperscript{13} In a situation where several assessors rate candidates on parallel forms of assessments (e.g., an OSCE), generalizability approaches permit the estimation of several sources of error of measurement (i.e., assessors and different forms of the instrument). When veterinary students are assessed by two or more examiners on a communications examination test that utilizes different standardized clients, generalizability approaches allows the simultaneous estimation or variability of measurement due to assessors; estimation of variability due to client; and estimation of individual student variability due to differences in performance. Recently, generalizability analysis has been employed to assess the reliability of an interview technique for human and veterinary medical schools, where applicants were assessed over several interview stations, with either one or two interviewers per station.\textsuperscript{14, 15} The authors were interested in variability due to applicants, stations, and interviewers, and the results suggested that interviewers were able to distinguish between applicants within stations, stations differentiated between applicants, and interviewers were consistent in their ratings of applicants.

Generalizability approaches that result in generalizability coefficients (\( Ep^2 \)) therefore allow greater analysis of complex and multifaceted measurement problems than the more conventional internal consistency and retest approaches. This is particularly useful for determining the reliability of admissions procedures, specifically interviews, OSCEs, and clinical assessments such as clinical evaluation exercises or in-training evaluation ratings.

**DEFENSIBILITY: PROFESSIONAL STANDARDS AND THE LAW**

Probably no activity that educators and measurement experts perform has come under as much legal scrutiny as testing and assessment. Since the 1950s a number of important court decisions affecting the use of tests have been handed down. Some of the most important themes that have resulted from decisions include: 1) test bias resulting in discrimination against some groups, 2) group quotas and the “affirmative action” principle, 3) consumerism and test result disclosure, and 4) evidence of validity and reliability. This last theme is particularly important for the purpose of assessment in undergraduate veterinary medical education.

A joint committee of the American Psychological Association (including the Canadian Psychological Association), the National Council on Measurement in Education, and the American Educational Research Association has prepared a document, *Standards for Educational and Psychological Testing*,\textsuperscript{16} that describes validity and sets standards for test use and assessments. The *Standards* address professional and technical issues of test development and use in education, health professions, employment, and psychology and are widely used worldwide.
All assessments, particularly high-stakes assessments, are subject to challenges of reliability and validity. If challenged in a court of law, the developers and users of assessments need to demonstrate that the assessments and their use conform to the basic standards of reliability and validity as set out in the Standards for Educational and Psychological Testing and described briefly in the present report.

SUMMARY AND CONCLUSIONS

The critical characteristics of instruments in educational measurement are validity and reliability. Both of these are multifaceted, complex aspects of assessment devices. The focus of validity is on the extent to which an instrument measures what it is intended to measure, while reliability deals with the consistency of measurement.

Several levels of analysis of evidence for validity have been identified: face, content, criterion-related, and construct. The first two levels of analysis do not require the accumulation of empirical evidence. Conversely, empirical evidence is necessary for establishing criterion-related and construct validity. Multivariate analyses and procedures for investigating these levels of analysis of validity include multiple regression, exploratory factor analysis, confirmatory factor analysis, and, more recently, structural equation modeling.

There are several methods for estimating reliability: test–retest, parallel forms, split-half, and internal consistency. While each of these focuses on somewhat different aspects of consistency of measurement, each deals with some important element of it. Approaches that involve retesting after some lapsed time (i.e., test–retest, parallel forms at different times) estimate the stability of the measurement. Generalizability theory provides a new approach to reliability problems that allows the simultaneous estimation of variability contributed by various facets in the assessment situation.

Veterinary educators should have an understanding of each of these concepts because evaluation of veterinary students and performance measures provides evidence of the utility of assessment tools, informs teaching practices, and can guide curriculum development and revision.

REFERENCES


AUTHOR INFORMATION

Kent Hecker, PhD, is Assistant Professor, Veterinary Medical Education Research, G380 Health Sciences Centre, Department of Veterinary Clinical and Diagnostic Science, Faculty of Veterinary Medicine, Department of Community Health Sciences, Faculty of Medicine, University of Calgary, 3330 Hospital Drive, NW, Calgary, AB Canada, T2N 4N1. E-mail: kghecker@ucalgary.ca.
Claudio Violato, PhD, is Professor, HMG14 Health Sciences Centre, Medical Education and Research Unit, Department of Community Health Sciences, Faculty of Medicine, University of Calgary, 3330 Hospital Drive, NW, Calgary, AB Canada, T2N 4N1. E-mail: violato@ucalgary.ca.