Standard Setting Study for the ACPV Certification Examination

March 2015

Elizabeth A. Witt, Ph.D.
Witt Measurement Consulting

Prepared for:

The American College of Poultry Veterinarians
12627 San Jose Blvd., Suite 202
Jacksonville, FL 32223-8638
support@acpv.info
www.acpv.info

INTRODUCTION

The American College of Poultry Veterinarians (ACPV) held a standard setting study in Athens, Georgia on March 12-13, 2015. The purpose of this meeting was to establish a standard and passing scores for the new ACPV certification exam, which was recently constructed in alignment with the content outline that resulted from the recent job analysis. This report describes the methodology that was used for the standard setting and the outcome of the study: six passing scores, one for each of three sections on each of two forms of the certification exam.

PARTICIPANTS

Five subject matter experts (SMEs) served as panelists on the standard-setting committee. Ideally, a higher number of SMEs would be preferred. However, it is not always feasible to recruit a large number of qualified volunteers to participate in this type of activity when the number of experts in the field is small. More important than the size of the committee are the qualifications and representativeness of the committee members. In terms of ethnicity, gender, and geographic region, the group was as diverse as can be expected. Poultry medicine in the United States is largely concentrated in the South. Most participants were currently located in the U.S. South, one was from Canada, and at least two others had international experience. All of the SMEs serving on the standard-setting committee were certified poultry veterinarians with at least 10 years of experience, familiarity with the scope of work in the field, and knowledge of the abilities of entry-level certificants. At the beginning of the meeting, the panel discussed areas of expertise and specialty within the field of poultry medicine; they determined that the broad scope of specialties and work was in fact represented among them via their current and prior work experience.

The following individuals were panelists on the standard-setting committee:

Andrés Montoya
Veterinarian
Merck Animal Health

Darko Mitevska
Veterinarian
Poultry Health Services, LTD
Alberta, Canada

Pedro Villegas
Professor Emeritus
University of Georgia (Athens)

Dennis Wages
Professor, Poultry Health Management
College of Veterinary Medicine
North Carolina State University (Raleigh)

Andrea Zedek
Veterinarian
Zedek Poultry Consulting, LLC

The standard setting study was facilitated by Elizabeth Witt, Chief Consultant and Psychometrician, Witt Measurement Consulting. Also present were Janece Bevans-Kerr, Director of Member Services for ACPV and Karen Grogan, Executive Vice President of ACPV and Veterinarian with Chicken Scratch, LLC. These individuals participated in the discussions and provided information as needed but were not part of the decision-making panel.

METHODOLOGY

The five committee members were asked to play the role of candidates and take both forms of the exam prior to the standard-setting meeting. This allowed them to experience the exam from the perspective of a candidate and helped them to develop a realistic understanding of the difficulty of the exam. Three additional certified poultry veterinarians also participated in this pilot administration of the exam. The performance of all eight test takers was used in creating difficulty estimates for test items that had not been previously used.

The two-day standard-setting meeting began with training. The panel of experts was introduced to the concept and purpose of standard-setting and the methodologies that would be employed at this meeting. The difference between norm-referenced testing and criterion-referenced testing was discussed. When a passing score is based on a norm-referenced procedure, the pass/fail decision may depend on what other individuals are taking the exam. Those who score highest would pass, while those who scored lowest would fail. Norm-referenced testing is generally considered inappropriate in licensure and certification. In criterion-referenced testing, the pass/fail decision depends only on whether or not the candidate has met the established standard. Candidates who demonstrate that they have the level of knowledge, skill, and ability needed will pass; those who failed to demonstrate that they have the required level of competence will fail.

A distinction was drawn between the standard and the passing score, or cut score. The standard is the minimum level of knowledge, skill, and ability (KSA) that is needed to practice safely and effectively as a certified poultry veterinarian—that is, the minimum level of competence that is worthy of certification. The passing score is the score on the examination that corresponds to the standard. The passing score takes into account the difficulty of the exam and may differ from one form of the exam to another. The task of the committee was first to define the standard of competence appropriate for certification and then to set a passing score on each section of the new examination that would reflect that standard, given the difficulty of the particular section and form.

Participants also spent some time discussing why it is problematic to arbitrarily set a cut score, for example, at 65% correct, without taking into account the difficulty of the questions on that particular test form. Although such a cut could be described as “criterion-referenced,” the “criterion” is not a carefully defined standard of competence but simply a number of questions answered correctly. Indeed, the level of competence required to pass could vary from one form to another, while the cut remained the same. In that case, a candidate’s pass/fail outcome may depend on the specific form of the exam taken. Even if there is only one form of the exam, this kind of standard is not directly defined in terms of the KSAs needed for competent performance as a certified professional, but is dependent on many factors that affect the difficulty of the specific questions on the exam.

In order to define the standard, committee members first reviewed the purpose of the examination as described in the job analysis report, and discussed the purpose of certification. This was done to ensure

that participants shared a common understanding of why we certify poultry veterinarians and what that certification means.

Next, a significant amount of time was spent discussing the meaning of borderline performance. A "borderline" candidate is one whose performance is just good enough to be worthy of certification. This is generally not the average candidate, but may well represent a level of competence that is lower than that of the average candidate. At the same time, the borderline candidate is not an individual with a low level of knowledge and skill; indeed, a fairly high level of competence is required for certification, and the borderline candidate is one who demonstrates that level of competence – but just. We might say that the borderline candidate is one who is "just sufficiently qualified." Panelists discussed the qualities of the borderline candidate, not only in general, but in connection with each KSA listed in the examination content outline. The group walked through the content outline and, for each KSA listed, they described the level of competence the borderline candidate would exhibit, as compared with the performance of either an unqualified candidate or a highly qualified candidate. Through this process, they defined the standard of competence appropriate for certification in terms of the qualities of the borderline certified professional. When the group was satisfied that they clearly understood the level of competence of the borderline candidate, they moved on to evaluate the difficulty of each section of the exam in terms of the performance of the borderline certified poultry veterinarian.

Passing scores were set separately for each section and form of the new exam, using a combination of a modified bookmark and a modified Angoff method. The Angoff and bookmark methods are the most widely used standard-setting methods in educational and psychological testing, including licensure and certification. Ideally, the Angoff method might have been used for all sections of the examination. However, because of the sheer length of the exam and the limited time available for the meeting, it was not feasible to use a method that would require the evaluation of every individual item, as the Angoff does. Each form of the examination contains 200 multiple-choice items, 100 image items, and a 100-point practical section. In order to set a cut on two forms, 600 items would have to be evaluated individually – not counting the practical sections. Therefore, the bookmark method, which requires evaluating item difficulty only near the cut, served as the basis of the procedure used to determine the cut scores for the multiple-choice and slide/image sections. A polytomous version of the Angoff method was applied to determine the passing scores on the practical sections.

Throughout the standard-setting process, committee members were reminded to consider how the borderline candidate, as they had defined borderline performance, *would* respond to the test questions, not how their expectations might leave them to believe the borderline candidate *should* respond. As a reality check, the performance of the eight certified poultry veterinarians who participated in the pilot exam was presented to the committee.

In the traditional bookmark standard-setting procedure, empirical estimates of item difficulty are used to rank order all items from easiest to hardest. Subject matter experts (SMEs) draw a line or insert a bookmark between the questions that a borderline candidate would answer correctly and those that he or she would answer incorrectly. For this certification exam, empirical item difficulties were not available for all items, as approximately half of the questions on the multiple-choice and slide/image sections were new or had changed substantially. However, reasonable estimates of the relative difficulty of the new items were obtained by having a group of certified poultry veterinarians take the exam, then using the relationship between this group's performance and the performance of candidates to obtain statistical estimates of candidate item difficulty. Eight certified veterinarians took both forms of the exam. A simple linear regression was performed to obtain an equation for predicting candidate

difficulty from the performance of the new group. For the final difficulty estimates, empirical candidate difficulties were used wherever available, and the predicted values obtained from the regression were used for the new items. In addition, for a small number of items on each form of the multiple-choice and slide/image sections, Angoff estimates were obtained. These were previously used items that had changed substantially so that the empirical difficulty estimates were no longer valid. Committee members were asked to estimate the percentage of borderline candidates who would answer each of these items correctly; the consensus of the committee determined the value of the final difficulty estimate for that item. Final difficulties were then used to rank order the items.

Each form and section of the exam was evaluated by the committee separately. For the multiple-choice and slide/image sections, the SMEs were asked to review a section, ordered by question difficulty, and draw a line that would divide the section into two parts: questions that the borderline candidate would answer correctly, and questions that the borderline candidate would get wrong. The methodology differed slightly from the traditional bookmark procedure in that SMEs were fully informed regarding the imperfect nature of the difficulty estimates and were allowed to move some items above or below their line if the difficulty estimate appeared to be inaccurate. SMEs were cautioned that this should not be done lightly. Items should be moved above or below the line only with good reason and following discussion by the committee. Differences in bookmark settings were discussed, and all participants were given the opportunity to change their mind. The final bookmark for each section was determined by taking the mean position of the marks set by all committee members.

This modified version of the bookmark method is similar to the Ordered Item Book method described by Smith and Becker (2014) in the *Journal of Applied Testing Technology*.

For the practical section on both forms, all items were new or substantially changed since their last use. Although each practical section contributed 100 points, there were less than 50 subscores on each form. Following the Angoff methodology, SMEs were asked to evaluate the difficulty of each scored piece. Specifically, committee members were asked, “How many points would the borderline candidate get on this part of this question?” Data were available to show how the committee and the other pilot exam participants performed on each score point, as well as the overall performance on the pilot exam. The final passing score for the practical section of each form was determined by summing the borderline candidate scores for each SME and taking the mean total score.

As a final step in evaluating the standard and passing scores, the group discussed the range of pass rates that might be considered reasonable, given their understanding of the candidate population as well as the difficulty of the examination. Committee members felt strongly that they could not set a reasonable range because of the small number of people testing each year. They reviewed the six separate passing scores they had set, comparing across forms and sections, and felt that the final cuts were very reasonable, given the content and difficulty of the specific questions on the exam.

RESULTS

The raw score means established by the committee using the procedures described were as follows: Multiple-Choice Form A, 50.6; Multiple-Choice Form B, 50.1; Slide/Image Form A, 62.0; Slide/Image Form B, 61.4; Practical Form A, 57.2; Practical Form B, 54.9. (Note: the multiple-choice sections each had 200 items, but each item was worth only one-half point.) Potential rounding rules were discussed, and the committee unanimously agreed to use conventional rounding. Thus the final passing scores

were set as shown in the table below. Committee members perceived this exam as difficult in comparison with the previous test.

Section and Form	Passing Score		Consistency		
	Raw Score	**As Percent**	SE	Cut-1SE	Cut+1SE
Multiple-Choice Form A	**51**	**51%**	2.3	48	53
Multiple-Choice Form A	**50**	**50%**	2.8	47	53
Slide/Image Form A	**62**	**62%**	3.3	59	65
Slide/Image Form B	**61**	**61%**	2.6	59	64
Practical Form A	**57**	**57%**	2.3	55	60
Practical Form B	**55**	**55%**	1.6	53	57

RECOMMENDATION

The psychometric consultant recommends that the ACPV Board accept all six passing scores determined by the committee, without adjustment. Standard errors (SEs) are provided for each section as measures of consistency of the SME ratings. (The lower the SE, the greater the consistency.) The SE is sometimes used to adjust a passing score up or down if there is strong reason to believe that the cut determined by the committee is inappropriate. (For example, if all candidates were to fail and there is nothing about the group of candidates tested that would suggest they are less capable than usual, it could be that the cut was set too high.) In such a case, the cut is typically adjusted up or down by one SE. The basic rationale is that, if different individual SMEs were to serve on the committee, the final passing score may differ, but would be likely to fall within ± 1 SE of the cut set by the existing committee. However, such an adjustment should be made only if there is compelling evidence that the committee's cut is too high or too low. There was nothing in the group dynamic of the meeting or in the ratings of any SME on the committee to suggest that the passing scores should not be accepted as they are. Furthermore, the number of candidates testing each year tends to be small enough that pass rates are not a reliable indicator of the reasonableness of the passing score. The cuts established by the committee appear to be realistic for this rigorous certification exam.

REFERENCE

Smith, R.W., Davis-Becker, S.L., and O'Leary, L.S. (2014). Combining the best of two standard setting methods: the Ordered Item Booklet Angoff. *Journal of Applied Testing Technology,* 15(1), 18-26.