APÉNDICE II

Conducting a Standard Setting Study for the Puerto Rico Bar Exam and Notary Exam

F i n a l R e p o r t

January 14, 2021

Submitted By:
Chad W. Buckendahl, Ph.D.
402.770.0085
cbuckendahl@acsventures.com

Contents

ACS Ventures, LLC – *Bridging Theory & Practice*

Executive Summary

The Puerto Rico State Bar conducted a standard setting study October 27-30, 2020 to evaluate the passing scores for the Puerto Rico Bar Exam and Notary Exam. The results from this study serve as an important source of evidence for informing the final policy decision on any potential changes that might be made to the current passing scores for these exams. The study involved gathering judgments from panelists through the application of a standardized methodology, calculating results, and evaluating recommendations for a passing score.

For both examinations, the standard setting study applied two methods: Yes/No modification of the Angoff (1971) method (Impara & Plake, 1997) and a paper selection method (see Plake & Hambleton, 2001). Both methods require panelists to consider the characteristics of a minimally competent candidate (MCC) and then make judgments about expected performance on test questions or written responses to essay questions. Both examinations have both multiple-choice and constructed response items and the selection of these methods reflected consideration of these characteristics as well as the requirements of the standard setting method itself. The paper selection method was designed for examinations that use constructed response questions (i.e., short answers, essay questions) that are designed to measure multiple traits. The responses produced by applicants on the essay questions are examples of constructed response questions for which the paper selection method is applicable.

For the multiple-choice items on the examinations, panelists made independent judgments about whether the MCC would get the question right (Yes) or wrong (No). For the essay questions, panelists were asked to review multiple examples of responses from a range of score points and then identify the paper that best represented the work of the MCC. The process was repeated for each essay question with the results summed across questions to form an individual panelist's recommendation. The results of the two components – multiple-choice and essay – were then combined to form an overall recommendation.

To calculate the recommend scores for the multiple-choice items, the number of Yes responses were counted to determine an expected number correct. To calculate the recommended cut score for the essay questions, the actual scores for the example papers were averaged (i.e., mean, median) across the group. These calculations were summed across the questions. This process occurred for the Bar Examination and the Notary Examination.

The standard setting study results and evaluation feedback supported the validity of the panel's recommended passing scores for use with the Puerto Rico Bar Examination and the Notary Examination. Results from the study were analyzed to create a range of recommended passing scores. Additional factors can be considered by policy makers when determining the final passing score for each examination are discussed in the body of this report.

Introduction and Overview

The purpose of licensure examinations like the Puerto Rico Bar Examination (GBX) and Notary Examination (NPX) is to distinguish competent candidates from those that could do harm to the public. This purpose is distinguished from educational exams in that licensure exams are not designed to evaluate training programs, evaluate mastery of content, predict success in professional practice, or ensure employability. Although other stakeholders may attempt to use scores from the examination for one or more of these purposes, it is important to clearly state what interpretations and uses the test scores are designed to support or not. Therefore, the standard setting process was designed in a way to focus expert judgments about the criterion level of performance that aligns with minimal competence.

Assessment Design

The Puerto Rico Bar Examination and Notary Examination were developed to measure the breadth and depth of content needed by entry level lawyers and notaries. The examinations include both multiple choice and essay questions. Because it represents a typical examination administration, the September 2019 examination questions and data were used to inform the study. A decision about passing or failing is based on the compensatory performance of applicants on the examination and not any single component. This means that a candidate's total score on the examination is evaluated relative to the passing score to determine pass/fail status. The candidate does not need to separately "pass" the multiple-choice questions and the essay questions.

Study Purpose and Validity Framework

The purpose of this study was to recommend passing scores on these examinations that distinguished the performance characteristics of someone who was minimally competent from someone who was not competent. To establish recommended passing scores, Dr. Chad Buckendahl and Dr. Andrew Wiley of ACS Ventures, LLC (ACS) facilitated a virtual standard setting meeting for the Commonwealth of Puerto Rico October 27-30. To complete the study, a committee of lawyers licensed to practice in Puerto Rico were recruited to serve as subject matter experts (SMEs) that would recommend cut scores for these examinations. Dr. Roger Bolus, psychometric consultant to the program, served as an external evaluator to monitor implementation of the process, the discussion, the results, and documentation of the study.

This report describes the sources of validity evidence that were collected and reports the study's passing score recommendations. The Puerto Rico Bar is receiving these recommended passing scores within ranges of standard error to contribute to discussions about developing a policy recommendation that will then be provided to the Puerto Rico Supreme Court for final decision-making. These results would serve as a starting point for a final passing score to be established for use with the Puerto Rico Bar Examination and the Notary Examination.

Procedures

For both the GBX and the NPX, the final cut score recommendations were derived using two distinct standard setting methodologies. First, the Yes/No modification of the Angoff (1971) method as described by Impara and Plake (1997) was selected for panelists to make judgments on multiple-choice questions. Second, a paper selection method described by Plake and Hambleton (2001) was used for the essay questions. Both methods

focus on the relationship between empirical difficulty of the question, task, and scoring expectations and examinee performance. It is common for licensure and certification examinations that use multiple question types to also apply multiple methods for collecting standard setting judgments (Buckendahl & Davis-Becker, 2012).

Panelists and Observers

Panelists were recruited by Puerto Rico Bar to represent newly licensed and experienced attorneys and notaries who had knowledge of the content and familiarity with the target population of candidates. The panelists were licensed attorneys and notaries. Panelists were recruited to represent a range of stakeholder groups. These groups were defined as Recently Licensed Professionals (panelists with less than five years of experience) and Experienced Professionals (panelists with ten or more years of experience, but with experience working with newly licensed attorneys). A list of the panelists, practice areas, and years of experience is shown in Table 1 below. All panelists signed nondisclosure agreements that permitted them to discuss the standard setting activities and processes outside the study, but that they would not be able to discuss the specific definition of the minimally competent candidate, the exam content, or any of the preliminary results that they may have heard or observed during the study.

In addition to the panelists, there were also observers who attended the standard setting study. These included an external evaluator with expertise in psychometrics and staff from the Puerto Rico Bar Examination. The external evaluator and observers were included in the process to promote the transparency of the standard setting and to evaluate the implementation of the process by which a passing score would be recommended.

Table 1. List of panelists for the standard setting studies.

Name	Practice Area(s)	Years of Experience
Roberto Abesada Aguet	-Civil and Commercial Litigation -Federal Litigation	20
Albéniz Couret Fuentes	-Brands Law -Product Liability Claims -Contractual Disputes -Laws on Product Distribution	16
María Trelles Hernádez	-Civil and Commercial Litigation	15
Arturo Hernández González	-Commercial Litigation -Appellate Practice	5
Jean René Santiago	-Civil and Commercial Litigation -Corporate Law -Labor Law	3
Manuel Pietrantoni Cabrera	-Commercial Litigation	17
Carla Ferrari Lugo	-Intellectual Property Law -Property Law -Corporate Law -Bankruptcy -Notarial Law	17
José Lamas Rivera	-Civil Law (including torts, insurance, contracts, professional	6

	liability, personal injury, premises liability and commercial litigation)	
Rosamar García Fontán	-Bankruptcy -State Litigation -Federal Litigation	17
Carlos Yamín Rivera	-Public Finance Law -Public Policy Law -Civil Law	5
Dennis Martínez Colón	-Notarial Law -Mortgage Law	43
Leny Cáceres Vazquéz	-Civil and Commercial Litigation -Contractual Disputes -Notarial Law -Administrative Law	4
Francisco Adams Quesada	-Criminal Law in State and Federal Courts	18

Methods

Numerous standard setting methods are used to recommend passing scores on licensure and certification exams (Buckendahl & Davis-Becker, 2012; Hambleton & Pitoniak, 2006). The selection of the Yes/No Angoff and paper selection methods for this study reflected consideration of the characteristics of the exams as well as requirements of the standard setting method itself. The Yes/No Angoff method was designed for multiple-choice items while the paper selection method was designed for essay questions that are designed to measure multiple traits or skills.

For the multiple-choice questions, the rating task is to make a judgment about whether the minimally competent candidate (MCC) will get the question right (Yes) or wrong (No). For the essay questions, the rating tasks for the panelists was to evaluate candidate responses that represented a range of performance, but without knowing the score and to select the paper that best represented the work of the MCC. To calculate the recommended passing score, the sum of the number of Yes judgments combined with the actual scores for the essay papers that were selected by panelists to represented MCC performance was determined.

Study Activities

The Puerto Rico Bar Examination and Notary Examination standard setting studies were conducted October 27-30, 2020 using the Zoom online meeting platform. Prior to the meeting, participants were informed that they would be engaging in tasks that would result in a recommendation for a passing score for each examination. The standard setting procedures consisted of orientation and training, operational standard setting activities for each multiple-choice question and essay question, and a written evaluation to gather panelists' opinions of the process. Study orientation materials are provided in Appendix B.

Orientation

The meeting commenced on October 27th with Dr. Buckendahl providing a general orientation for all panelists that included the goals of the meeting, an overview of the paper selection method and its application. The opening orientation also described how passing scores would ultimately be determined through

recommendations to the Puerto Rico Bar. In addition, a generic scoring guide/rubric was shared with the panelists to provide a framework for how essay questions for the Bar Examination would be scored. The different areas of the scoring criteria were a) Issue spotting, b) Discussion of applicable law(s), c) Analysis and application of law(s) to fact pattern, and d) Drawing conclusions with supporting reasoning.

Part of the orientation was a discussion around the expectations for an entry-level lawyer who is a minimally competent and capable of passing the examinations. The process for defining minimum competency is policy driven and started with a draft definition produced by the Puerto Rico Bar.

Based on the input from multiple stakeholder groups and relying on best practice as suggested by Egan et al. (2012), the Puerto Rico Bar provided the following description of minimally competent candidate (MCC).

> A minimally qualified candidate will be able to demonstrate the following knowledge, skill, and legal reasoning ability with responses that contain some errors of fact and judgment:
>
> (1) Foundational knowledge of legal rules and principles in common practice areas. May need assistance to identify all elements or dimensions of rules.
>
> (2) Ability to distinguish relevant from irrelevant information when assessing a legal rule and identify information that would be helpful in making the assessment.
>
> (3) Ability to apply a legal rule or rules to a given set of facts but may miss some dimensions of the relationship between fact and law.
>
> (4) Construct and communicate legal conclusions and reasoning given the context.

Additionally, the facilitator guided the panel through a process where panelists further discussed the MCC by answering the following questions:

- What knowledge, skills, and abilities are representative of the work of the MCC?

- What knowledge, skills, and abilities would be easier for the MCC?

- What knowledge, skills, and abilities would be more difficult for the MCC?

Operational Standard Setting Judgments
After completing the training activities panelists began their ratings with the essay questions for the Bar Examination. The questions and candidate responses from the September 2019 test examinations were used throughout the workshop. For each question, panelists reviewed exemplar papers (i.e., candidate responses) that were distributed across score points. Their task was to select the paper that was most representative of the work of the MCC. For the study, these exemplars were randomly ordered and only identified with a code that represented the score that the exemplar received during the grading process in 2019. Panelists were not told the scores on the exemplars to maintain their focus on the criterion-referenced definition of performance and the content rather than an intuitive perception of a given score. To facilitate panelists' consistency, panelists were told which essay questions were similar in terms of empirical difficulty without revealing the raw scores.

To convert the panelists' ratings into scores, the first step was to use a look up table to determine the underlying raw score associated with an exemplar code. This was done for each essay question. The conversion of the exemplar codes into the scores then allowed for summation across panelists to calculate average (i.e., mean, median) values along with estimates of variability (e.g., standard deviation, frequency distribution).

After completing their ratings on the first question, the facilitator led a discussion of the rationale for why panelists selected the paper that they did using the performance level descriptor as their reference point. This process of discussion occurred as a full group and was intended to reinforce the methodology and the need to use the definition of minimum competency to inform the judgments about exemplar classification. Following this discussion, the judgment process was independently replicated for each of the subsequent essay questions. For the Bar Examination, panelists reviewed each of the eight essay questions before transitioning to the Notary Examination where they reviewed two essay questions. Similar discussions about the MCC occurred for the Notary Examination.

Following ratings and discussions over the first two days of the process, panelists then convened at the Puerto Rico Bar offices to rate the multiple-choice questions over the final two days of the study. Because the multiple-choice questions are maintained in a secure environment, staff from Puerto Rico Bar managed logistics and materials onsite while facilitators led the group using the online meeting software. At the beginning on the third day, Dr. Buckendahl provided additional training to the panelists that described how panelists would complete the Yes/No process. Once all panelists completed their ratings, all ratings were sent to ACS to evaluate the first round of ratings. On the last day, Dr. Buckendahl reviewed the initial recommendations as well as the ratings for some specific items. After this discussion, panelists were permitted time to complete a 2nd round of ratings for all items. Following completion of the rating activities, panelists completed a written evaluation form of the process.

Analysis and Results

Panelists' Recommendations

To calculate the recommended passing score on the examination from the panelists' judgments, the individual recommendations for each panelist were summed across the essay and multiple-choice questions for each examination. Analyses were conducted to convert the recommendations that panelists made on a raw score scale to the scale score. This was accomplished by applying the conversion formula provided by Puerto Rico Bar. The goal in creating a scale score is to convert the raw score to an interpretive scale that allows the program to maintain the meaning of the scale over time. We calculated the average scale score for the group and calculated the standard error of the mean to account for measurement error that may be part of the judgment process.

Because mean and median recommendations for the Bar Examination and the Notary Examination generally converged, we are reporting the mean values as the more common representation of an average score. After calculating the mean and the standard error of the mean, we evaluated the impact of the recommendation. Specifically, we calculated the estimated pass rate for the examination based on the group's recommendation using a cumulative percent distribution. Using the panelists' recommended scale score for the Bar Examination (575) as an example, we looked at the location within the cumulative percent distribution to determine what percent of candidates scored below this value. To illustrate the range of recommendations,

we then looked for the impact at locations that were +/- 1 and 2 standard errors of the mean above and below this recommended value.

The summary results for each examination and each Round of judgments are presented in Table 2. The panel's recommended mean and median with the associated standard errors are included along with the impact along with a +/- 2 standard error of mean. Individual ratings for each essay question, the multiple-choice questions, and the summary calculations are included in Appendix C. Data were de-identified to preserve anonymity of individual panelists. The summary results of these analyses are shown here in Table 2.

Table 2. Summary scale score results with range of recommendations for the Bar Examination and Notary Examination.

	Bar Exam – Mean Scale Score	Bar Exam – Pass Rate	Notary Exam – Mean Scale Score	Notary Exam – Pass Rate
Round 1				
-2 SE$_{Mean}$	574	42.8%	588	47.9%
-1 SE$_{Mean}$	579	39.4%	592	47.6%
Recommended scale score (SE$_{Mean}$)	584 (5.08)	35.8%	596 (4.30)	47.0%
+1 SE$_{Mean}$	589	35.5%	600	43.0%
+2 SE$_{Mean}$	594	34.7%	604	41.5%
Round 2				
-2 SE$_{Mean}$	569	46.4%	591	47.6%
-1 SE$_{Mean}$	572	43.6%	595	47.0%
Recommended scale score (SE$_{Mean}$)	575 (3.11)	41.5%	599 (4.08)	43.0%
+1 SE$_{Mean}$	578	39.4%	603	41.5%
+2 SE$_{Mean}$	581	37.3%	607	38.1%

Process Evaluation Results

Panelists completed an evaluation that included rating scale and open-ended questions. The responses to the questions are included in Table 3 and the comments provided are included in Appendix C. With the exception of questions related to time allocation that was rated on a 3-point scale (1 = not enough, 2 = about right, 3 = too much), ratings closer to 4.0 can be interpreted as more positive perceptions of the question (e.g., success of training, confidence in ratings, appropriate time) versus values closer to 1.0 which suggest perceptions that are more negative with respect to these questions.

Table 3. Written Process Evaluation Summary Results

	Median	1 - Lower	2	3	4 - Higher
1. Success of Training					
Training provided	4	0	0	1	11
Purpose of the study	4	0	0	3	9
Information about the exams	4	0	0	1	11
Discussion of the Minimally Qualified Candidate	4	0	0	4	8
Methodology for essay questions	4	0	0	2	10
Methodology for multiple-choice questions	4	0	0	3	9
2. Time allocation to Training	2	1	9	2	N/A
3. Confidence moving from Practice to Operational	4	0	0	5	7
4. Time allocated to Discuss Initial Essay and Multiple-Choice Questions	2	4	8	0	N/A
6. Time to make Round 1 judgments	2	1	11	0	N/A
7. Helpfulness of feedback data provided after first Round judgments	4	0	0	2	10
9. Primary focus in modifying judgments	3	0	0	11	1
1 = colleagues' explanation; 2 = colleagues' ratings; 3 = more appropriate for the exam; 4 = maintaining my recommendation					
10. Confidence in Bar Examination recommendation	4	0	0	5	7
12. Confidence in Notary Examination recommendation	3.5	1	1	4	6
14. Overall success of the study	4	0	0	0	12

Collectively, the results of the panelists' evaluation suggested generally positive perception of the activities for the study, their ratings, and the outcomes. The ratings regarding the time allocation were generally lower which can be attributed to the intensity of the task and the amount of work. Future studies may benefit from an additional day or two to permit more reasonable workload for the panelists.

Evaluating the Passing Score Recommendations

To evaluate the passing score recommendations that were generated from this study, we applied Kane's (1994; 2001) framework for validating standard setting activities. Within this framework, Kane suggested three sources of evidence that should be considered in the validation process: procedural, internal, and external. When evaluating procedural evidence, practitioners generally look to panelist selection and qualification, the choice of methodology, the application of the methodology, and the panelists' perspectives about the implementation of the methodology as some of the primary sources. The internal evidence for standard setting is often evaluated by examining the consistency of panelists' ratings and the convergence of the recommendations. Sources of external evidence of validity for similar studies include impact data to inform the reasonableness of the recommended passing scores. Evidence within each of these areas that was observed in this study is discussed here.

Procedural

When evaluating procedural evidence, practitioners consider panelist selection and qualifications, the choice of methodology, the application of the methodology, and the panelists' perspectives about the implementation of the methodology as some of the primary sources. For this study, the panel that was recruited and selected by Puerto Rico Bar to represent a range of stakeholders: newer and more experienced attorneys having familiarity with the content and target population of candidates. The choice of methodology aligned with the different measurement strategies to use multiple-choice and essay questions. Panelists' perspectives on the process were collected with evaluation responses showing positive reactions to the process and confidence in the ratings.

Internal

The internal evidence for standard setting is often evaluated by examining the consistency of panelists' ratings and the convergence of the recommendations. The standard error of the mean on which the recommendation was based (approximately 3.0 for the Bar Examination and 4.0 for the Notary Examination) suggests a reasonable range. These observations suggest that panelists were generally in agreement regarding the expectations of which applicant responses were characteristic of the Minimally Competent Candidate.

External

External evidence for licensure programs is particularly difficult to collect because comparable, psychometrically sound alternatives that measure the same construct are rarely available. The interpretation of scores from these examinations is criterion-referenced which means that anyone can meet the performance threshold; there is not a predefined distribution of pass/fail decisions that is expected. The use of historical data can inform the reasonableness of the recommendations. Puerto Rico may also want to consider looking at performance from examinations given at different times of year if there is an expectation of sample differences between exam administrations.

Because Puerto Rico's Bar Examination and Notary Examination are unique to the Commonwealth, relying on performance from states may not be as informative. However, it may provide some value to collect external data candidates from Puerto Rico who take Bar Examinations in states to evaluate pass/fail performance as a check on the reasonableness. In doing so, though, caution is urged because it is unknown whether states used for comparison have conducted formal standard setting studies or whether the performance level description is comparable to the one adopted by Puerto Rico.

Determining Final Passing Score Recommendations

The next step in the process is for Puerto Rico Bar to consider the results from this study in combination with related policy factors to make a final determination of the passing score for the Bar Examination and the Notary Examination. One of those factors is the policy tolerance for different types of classification errors. Because we know that there is measurement error with any test score, when applying a passing score to make an important decision about an individual, it is important to consider the risk of each type of error. A *Type I* error represents an individual who passes an examination, but whose true abilities are below the passing score. These types of classification errors are considered false positives. Conversely, a *Type II* error represents an individual who does not pass an examination, but whose true abilities are above the passing score. These types of classification errors are known as false negatives. Both types of errors are theoretical because we cannot know which test takers in the distribution around the passing score may be false positives or false negatives.

A policy body can articulate its rationale for supporting adoption of the group's recommendation or adjusting the recommendation in such a way that minimizes one type of misclassification. The policy rationale for licensure examination programs is based primarily on deliberation of the risk of each type of error. For example, many licensure and certification examinations in healthcare fields have a greater policy tolerance for *Type II* errors than *Type I* errors with the rationale that the public is at greater risk for adverse consequences from an unqualified candidate who passes (i.e., *Type I* error) than a qualified one who fails (i.e., *Type II* error).

In applying the rationale, if the policy decision is that there is a greater tolerance for *Type I* errors, then the decision would be to accept the recommendation of the panel or adopt a value that is a standard error below the recommendation. Conversely, if the policy decision is that there is a greater tolerance for *Type II* errors, then the decision would be to accept the recommendation of the panel or adopt a value that is a standard error above the recommendation. Because standard setting studies represent an integration of policy and psychometrics, the final determination will be policy driven, but supported by the data collected in this study and related factors.

We recommend that the program consider a final passing score for the Bar Examination somewhere in the range of the recommended passing score (575) and a score that is two standard errors of the mean below this score (569). The rationale for this recommendation is that the reference point for the panelists during the study was the Minimally Competent Candidate and panelists made judgments to predict how these candidates would perform on the multiple-choice questions and essay questions for the examination. This means that the distribution of reference candidates was all intended to be minimally competent. In creating that distribution, the lower bound would likely best represent the threshold of minimum competency suggested by the panelists. Setting the passing score at 569 would mean that approximately 46.4% of candidates would pass the examination while setting the passing score at 575 would mean that approximately 41.5% of candidates would pass. This range is consistent with the recommendations of the panelists as characterizing the performance of the minimally competent candidate.

Similarly, for the Notary Examination, we recommend applying the same decision rule where the final passing score would be set somewhere between the group's recommended value (599) and a value that is two standard errors of the mean below this value (591). This would produce estimated passing rates that range from 43.0% to 47.6%. The rationale for considering this range mirrors the reasoning described above for the Bar Examination.

References

Buckendahl, C. W. & Davis-Becker, S. (2012). Setting passing standards for credentialing programs. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 485-502). New York, NY: Routledge.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. J. Cizek (Ed.), Setting performance standards: Foundations, method, and innovations (2nd ed., pp. 79-106). New York, NY: Routledge.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education and Praeger.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64* (3), 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Plake, B. S. & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

Appendix A – Standard Setting Materials

Study Agenda

Study Orientation
Slides

Study Rating Forms

Appendix B – Standard Setting Data

GBX Data -
Technical Report

Notary Data -
Technical Report

Appendix C – Evaluation Comments

- Outstanding workshop
- I want to congratulate the BAR and ACS Ventures for this excellent and professional work. Very proud to work with these amazing colleagues in this effort for the Bar.
- It would have been helpful to discuss in more details the topic of each question.