ACS VENTURES
BRIDGING THEORY & PRACTICE

Conducting a Standard Setting Study for the California Bar Exam

Final Report

July 28, 2017

Submitted By:
Chad W. Buckendahl, Ph.D.
Office: 702-586-7386
cbuckendahl@acsventures.com

Contents

Executive Summary

The California State Bar conducted a standard setting workshop[1] May 15-17, 2017 to evaluate the passing score for the California Bar Exam. The results from this workshop serve as an important source of evidence for informing the final policy decision on what, if any, changes to make in the current required passing score. The workshop involved gathering judgments from panelists through the application of a standardized process for recommending passing scores and then calculating a recommendation for a passing score.

The standard setting workshop applied a modification of the Analytic Judgment Method (AJM; Plake & Hambleton, 2001). This method entails asking panelists to classify illustrative responses into defined categories (e.g., not competent, competent, highly competent). The selection of the AJM for the California Bar Examination reflected consideration of the characteristics of the exam as well as requirements of the standard setting method itself. The AJM was designed for examinations that use constructed response questions (i.e. narrative written answers) that are designed to measure multiple traits. The responses produced by applicants on the essay questions and performance task are examples of constructed response questions for which the AJM is applicable.[2]

The methodology involved identifying exemplars of applicant performance that span the observed score scale for the examination. The exemplar performances were good representations of the respective score point such that the underlying score was not in question. The rating task for the panelists was to first broadly classify each exemplar into two or more categories (e.g., not competent, competent, highly competent). Once this broad classification was completed, panelists then refined those judgments by identifying the papers close to the target threshold (i.e., minimally competent). This meant that the panelists identified the best of the not competent exemplars and the worst of the competent exemplars that they had initially classified. The process was repeated for each essay question and performance task with the results summed across questions to form an individual panelist's recommendation.

To calculate the recommended cut score for a given question for a panelist, the underlying scores for the exemplars identified by a respective panelist were averaged (i.e., mean, median) across the group. These calculations were summed across the questions with each essay question being equally weighted and the performance task counting for twice as much as an individual essay question to model the operational scoring that will occur beginning with the July 2017 administration.

Following these judgments, we calculated the recommended score and associated passing rate when considering the written part of the examination. However, we needed to know what score on the total exam corresponded to this same pass rate. To answer this question, another step was needed to transform these

---

[1] Standard setting is the phase of examination development and validation that involves the systematic application of policy to the scores and decisions on an examination. Conducting these studies to establish passing scores is expected by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

[2] Alternative methods that rely on panelists' judgments of candidate work include Paper Selection and Body of Work (see Hambleton & Pitoniak, 2006, for additional details on these and a discussion of the categories of standard setting methods).

judgments to the score scale on the full-length examination. After creating distributions of individual recommendations for the written part of the examination, to estimate the score for the full-length examination we applied an equipercentile linking approach to find the score that yielded the same percent passing as was determined on just the written component of the examination that panelists evaluated. Equipercentile involves finding the equivalent percentile rank within one distribution of scores and transforming to another score distribution to retain the same impact from one examination to another or in this instance, from a part of the examination on which panelists made judgments to the full examination.

The **standard setting meeting results and evaluation feedback generally supported the validity of the panel's recommended passing score for use with the California Bar Examination**. Results from the study were analyzed to create a range of recommended passing scores. However, additional policy factors may be considered when establishing the passing score. One of these factors may include the recommended passing score and impact relative to the historical passing score and impact. The panel's median recommended passing score of 1439 converged with the program's existing passing score while the mean recommended passing score of 1451 was higher.

Additional factors that could be considered in determining the appropriate cut score for California might include the passing rates from other states that have similarly large numbers of bar applicants sitting for the examination. However, the interpretation of these results and the comparability are mitigated by the different eligibility policies among these jurisdictions and **California's more inclusive policies** as to who may sit for the exam [3]along with the downward trend in bar examination performance across the country, particularly over the last few years. In some instances, the gap passing the bar exam between California's applicants and other states has closed and in others, the gap observed in 2007 has remained essentially constant as the trend declined on a similar slope.

An additional factor warrants consideration as part of the policy deliberation. Specifically, the consideration of policy tolerance for different types of classification errors is relevant. 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 cut 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 in nature 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

[3] California has a uniquely inclusive policy as to who may be eligible to take the Bar Exam. Not only those who have graduated from schools nationally accredited by the American Bar Association, but applicants from California accredited and unaccredited law schools are also allowed to take the exam, as well as those who have 'read law.' This sets California apart from virtually all other jurisdictions.

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 I*I 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 (i.e., 144) or adopt a value that is one to two standard errors below the recommendation (i.e., 139 to 141). 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 (i.e., 144) or adopt a value that is one to two standard errors above the recommendation (i.e., 148 to 150). Because standard setting is an integration of policy and psychometrics, the final determination will be policy driven, but supported by the data collected in this workshop and this study more broadly.

Introduction and Overview

The purpose of licensure examinations like the California Bar Exam[4] is to distinguish competent candidates from those that could do harm to the public. This examination purpose is distinguished from other types of 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 inferences 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 level of performance that aligns with minimal competence.

Assessment Design

The California Bar Exam is built on multiple components intended to measure the breadth and depth of content needed by entry level attorneys who are minimally competent. These components are the Multistate Bar Exam (MBE), five essay questions, and a performance task[5]. Beginning with the July 2017 examination, the combined score for the examination weights the MBE at 50% and the constructed response components at 50% with the performance task being weighted as twice as much as an essay question.[6] 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 an applicant's total score on the examination is evaluated relative to the passing score to determine pass/fail status. The applicant does not need to separately "pass" the MBE and the constructed response questions.

Study Purpose and Validity Framework

The purpose of this study was to recommend a passing score that distinguished the performance characteristics of someone who was minimally competent from someone who was not competent. To establish a recommended passing score, Dr. Chad Buckendahl of ACS Ventures, LLC (ACS) facilitated a standard setting meeting for The State Bar of California on May 15-17, 2017 in San Francisco, CA. The purpose of the meeting was to enlist subject matter experts (SMEs) to serve as panelists and recommend cut scores that designate the targeted level of minimally competent performance.

---

[4] Note that the California Department of Consumer Affairs is responsible for managing the licensure process for many professions and consults with many others. As such, a representative from the Department was asked to serve as an external reviewer for this study.

[5] The performance task is designed to measure skills associated with the entry level practice of law (e.g., legal analysis, reasoning, written communication) separate from the domain specific application of these skills to specific subject areas as are measured in the essay questions.

[6] Prior to the July 2017 exam, MBE accounted for 35% of the exam, with the constructed response components weighted 65% of the total. Previously, constructed responses consisted of six essay and two performance task questions. While the papers used in the workshop were originally administered according to the old format, in anticipation of the new cut score potentially applied to exams from July 2017 based on the new format, the five essay and one performance test questions were used in the workshop to conform with the new exam structure.

To evaluate the cut score recommendations that were generated from this study, Kane's (2001) framework for evaluating standard setting activities was used. Within this framework, Kane suggests three sources of evidence 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 cut scores.

This report describes the sources of validity evidence that were collected and reports the study's passing score recommendations. The California Bar is receiving these recommended passing score within ranges of standard error to contribute to discussions about developing a policy recommendation that will then be provided to the California 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 California Bar Exam.

Procedures

The standard setting study used a modified version of the Analytic Judgment Method (AJM; Plake & Hambleton, 2001). The AJM approach is characterized as a test based method (Hambleton & Pitoniak, 2006) that focuses on the relationship between item difficulty and examinee performance on the test. It is appropriate for tests that use constructed response items like the essay questions and performance task that are part of the written part of the California Bar Exam (see Buckendahl & Davis-Becker, 2012). The primary modification for the study was to reduce the number of applicants' performances that panelists reviewed from 50 to 30 given the score scale for each essay question and the performance task.

Panelists and Observers

A total of 20 panelists participated in the workshop[7]. The panelists were licensed attorneys with an average of 14 years of experience in the field. 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), Experienced Professionals (panelists with ten or more years of experience), and Faculty/Educator (panelists who are employed at a college or university). Note that some panelists were associated with multiple roles. Some of the experienced attorneys also served as adjunct faculty members at law schools. In listing their employment type in the table below, we have documented the primary role indicated by panelists. A summary of the panelists' qualifications is shown in Table 1.

In addition to the panelists, there were also observers who attended the in-person standard setting workshop. These included an external evaluator with expertise in standard setting, a representative from the California Department of Consumer Affairs, representatives from California Law Schools, a representative from the Committee on Bar Examinations, and staff from the California Bar Examination. Observers were instructed

[7] Nominations to participate on the standard setting panel were submitted to the Supreme Court who selected participants to represent diverse backgrounds with respect to experience, practice areas, size of firms, geographic location, gender, and race/ethnicity.

during the orientation of the meeting that they were not to intervene or discuss the standard setting activities with the panelists. All panelists and observers signed confidentiality and nondisclosure agreements that permitted them to discuss the standard setting activities and processes outside the workshop, but that they would not be able to discuss the specific definition of the minimally competent candidate or any of the preliminary results that they may have heard or observed during the study. External evaluators and observers were included in the process to promote the transparency of the standard setting and to critically evaluate the fidelity of the process by which a passing score would be recommended.

Table 1. Summary of panelist demographic characteristics.

Race/Ethnicity	Freq.	Percent
Asian	3	15.0
Asian/White	1	5.0
Black	4	20.0
Hispanic	2	10.0
White	10	50.0
Total	20	100.0

Gender	Freq.	Percent
Female	9	45.0
Male	11	55.0
Total	20	100.0

Years of Practice	Freq.	Percent
5 Years or Less	10	50.0
>=10	10	50.0
Total	20	100.0

Nominating Entity	Freq.	Percent
ABA Law Schools	3	15.0
Assembly Judiciary Comm.	1	5.0
Board of Trustees	2	10.0
BOT - CBE*	1	5.0
BOT - COAF*	8	40.0
BOT - CYLA*	2	10.0
CALS Law Schools	1	5.0
Governor	1	5.0
Senior Grader	1	5.0
Total	20	100.0

* Committee of Bar Examiners; Council on Access and Fairness; California Young Lawyers Association.

Primary Employment Type	Freq.	Percent
Academic	2	10.0
Court	1	5.0
District Attorney	1	5.0
Large Firm	4	20.0
Non Profit	3	15.0
Other Govt.	3	15.0
Public Defender	1	5.0
Small Firm	3	15.0
Solo Practice	2	10.0
Total	20	100.0

Practice Areas	Freq.	%
Business	12	17%
Personal Injury	6	9%
Appellate	5	7%
Criminal	5	7%
Labor Relations	4	6%

Juvenile Delinquency	3	4%
Probate	3	4%
Real Estate	3	4%
Antitrust	2	3%
Disability Rights	2	3%
Employment	2	3%
Environmental Law	2	3%
Family	2	3%
Insurance Coverage	2	3%
Intellectual Property	2	3%
Administrative Law	1	1%
Civil Rights	1	1%
Contract Indemnity Litigation	1	1%
Education	1	1%
Elder Abuse	1	1%
General Commercial Litigation	1	1%
Government Transparency	1	1%
Immigration	1	1%
Legal Malpractice	1	1%
Mass Tort	1	1%
Nonprofit Law	1	1%
Policy Advocacy	1	1%
Product Liability	1	1%
Public Interest	1	1%
Total	69	100%

Method

Numerous standard setting methods are used to recommend passing scores on credentialing[8] exams (Hambleton & Pitoniak, 2006). The selection of the Analytical Judgment Method (AJM; Plake & Hambleton, 2001) for the California Bar Exam reflected consideration of the characteristics of the exam as well as requirements of the standard setting method itself. The AJM was designed for examinations that use constructed response questions that are designed to measure multiple traits. The responses produced by the applicants on the essay questions and performance task of the California Bar Exam are examples of constructed response questions where the AJM is applicable.

The methodology first involves identifying exemplars of applicant performance that span the observed score scale for the examination. The exemplar performances should be good representations of the respective score point such that the underlying score should not be in question. Plake and Hambleton (2001) suggested using

[8] Credentialing is an inclusive term that is used to refer to licensure, certification, registration, and certificate programs.

50 exemplars to ensure that there was sufficient representation of the score scale. Once these exemplars have been identified, they should be randomly ordered and coded to de-identify the score for the standard setting panelists. The goal is to have the panelists focus on the interpretation of the performance level descriptor of minimum competency and not the score of the paper.

The rating task for the panelists is to then broadly classify each exemplar into two or more categories (e.g., not competent, competent, highly competent). Once this broad classification is completed, panelists are asked to then refine those judgments by identifying the papers close to one or more thresholds. For example, if the target threshold is minimum competency, then panelists would identify the best of the not competent exemplars and the worst of the competent exemplars. To calculate the recommended cut score for a given question, the underlying scores for these exemplars are averaged (i.e., mean, median) to determine a value for this question. The process is then repeated for each essay question and performance task with the results summed across questions to form an individual panelist's recommendation.

In the operationalization of this method for this study, two modifications of the methodology were used. First, rather than having 50 exemplars for each question, panelists evaluated 30 exemplars for each question. This modification was applied primarily due to the width of the effective scale. Meaning, although the theoretical score scale for each essay question spans from 0-100, the effective score scale only ranges from approximately 45-90 and is limited to increments of 5 points. This reduces the number of potential scale score points and thereby reduces the number exemplars necessary for each score point to illustrate the range. The second modification of the process involved sharing with the panelists a generic scoring guide/rubric as opposed to specific ones for each question. This was done to avoid potentially biasing the panelists in their judgments and to focus on the common structure of how the constructed response questions were scored.

In the rating task, panelists were asked to review examples of performance and categorize each example as either characteristic of *not competent*, *competent*, or *highly competent* performance. Even though the only target threshold level was *minimally competent*, the use of *highly competent* as a loosely defined category was meant to filter out exemplars that would not be considered in the refined judgments. Following the broad classification, these initial classifications were then refined to identify the papers that best represented the transition point from not competent to competent (i.e., minimally competent). Once these papers were identified by the panelists (i.e., the two best not competent exemplars and the two worst competent exemplars), the actual scores that these exemplars received during the actual, original grading process were used to calculate the average values of the panelists' recommendations for each question and then summed across questions.

Workshop Activities

The California Bar Exam standard setting meeting was conducted May 15-17, 2017 in San Francisco, CA. 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 the examination. The standard setting procedures consisted of orientation and training, operational standard setting activities for each essay/performance task, and successive evaluations to gather panelists' opinions of the process. Chad Buckendahl, Ph.D., served as the facilitator for the meeting. Workshop orientation materials are provided in Appendix B.

Orientation

The meeting commenced on May 15th with Dr. Buckendahl providing a general orientation for all panelists that included the goals of the meeting, an overview of the Analytical Judgment Method and its application, and specific instructions for panel activities. Additionally, the opening orientation described how cut scores would ultimately be determined through recommendations to the California State Bar. In addition, a generic scoring guide/rubric was shared with the panelists to provide a framework for how essay questions and the performance task would be scored. The different areas of the scoring criteria were a) Issue spotting, b) Identifying elements of applicable law, c) Analysis and application of law to fact pattern, d) Formulating conclusions based on analysis, and e) Justification for conclusions. Each essay question and performance task had a unique scoring guide/rubric for the respective question, but followed this generic structure.

Part of the orientation was a discussion around the expectations for someone who is a minimally competent lawyer and therefore should be capable of passing the exam. The process for defining minimum competency is policy driven and started with a draft definition produced by the California Bar. Feedback was solicited from law school deans, the Supreme Court of California, and the workshop facilitator for substance and style.

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

> A minimally competent applicant will be able to demonstrate the following at a level that shows meaningful knowledge, skill and legal reasoning ability, but will likely provide incomplete responses that contain some errors of both fact and judgment:
>
> (1) Rudimentary knowledge of a range of legal rules and principles in a number of fields in which many practitioners come into contact. May need assistance to identify all elements or dimensions of these rules.
>
> (2) Ability to distinguish relevant from irrelevant information when assessing a particular situation in light of a given legal rule, and identify what additional information would be helpful in making the assessment.
>
> (3) Ability to explain the application of a legal rule or rules to a particular set of facts. An applicant may be minimally competent even if s/he may over or under-explain these applications, or miss some dimensions of the relationship between fact and law.
>
> (4) Formulate and communicate basic legal conclusions and recommendations in light of the law and available facts.

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?

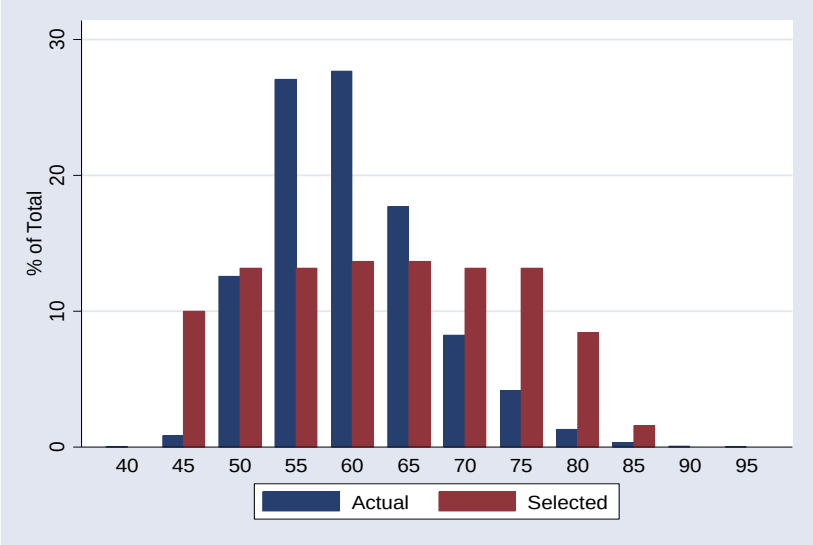
The results of this discussion and the illustrative characteristics of MCC performance for each of the subject areas that were included in this study are included as an embedded document in Appendix C.

Training/Practice with the Method

Panelists also engaged in specific training regarding the AJM. This involved a discussion about the initial task of broadly classifying exemplars into one of three categories – not competent, competent, or highly competent – and using the performance level descriptor (PLD) of the MCC to guide those judgments. In addition, prior to the operational ratings, panelists were given an opportunity to practice with the methodology. The practice activity replicated the operational judgments with two exceptions: a) panelists were only given 10 exemplars distributed across the score scale to review and b) panelists only identified one exemplar that represented the best not competent and the worst competent. Panelists then discussed their selections and the reasoning for why their judgments reflected the upper and lower bound of the expected performance of the MCC.

Written Exam Score Distributions - Actual and Sample Selected for Workshop

	Actual		Selected	
Score	Freq.	*%*	Freq.	*%*
40	29	*0.1*	0	*0.0*
45	436	*0.8*	19	*10.0*
50	6,669	*12.6*	25	*13.2*
55	14,354	*27.1*	25	*13.2*
60	14,678	*27.7*	26	*13.7*
65	9,383	*17.7*	26	*13.7*
70	4,365	*8.2*	25	*13.2*
75	2,206	*4.2*	25	*13.2*
80	689	*1.3*	16	*8.4*
85	178	*0.3*	3	*1.6*
90	33	*0.1*	0	*0.0*
95	3	*0.0*	0	*0.0*
Total	53,023	*100.0*	190	*100.0*



Operational Standard Setting Judgments

After completing the training activities panelists began their ratings by independently classifying the 30 exemplars that were selected for the first question. The 30 exemplars for each question were selected to approximate a uniform distribution (i.e., about the same number of exemplars across the range of observed scores). Figure 1 below shows the distribution of scores for the written section of the examination along with the distribution of exemplars that were selected for this study.

Figure 1. Distribution of observed scores and selected exemplars for the written section of the California Bar Examination from July 2016.

For the study, these exemplars were then randomly ordered and only identified with a code that represented the score that the exemplar received during the grading process in 2016. Panelists were not told the scores on the exemplars to maintain their focus on the content rather than an intuitive perception of a given score. After panelists made their initial, broad classification, they identified the ***two best not competent exemplars*** and the ***two worst competent exemplars*** from their initial classifications. The selection of these specific exemplars is used to estimate the types of performance that would be demonstrated by a MCC. Panelists used a predeveloped rating form to indicate the codes on the exemplars that aligned with these instructions.

To convert the panelists' ratings into numerical values to then calculate the recommendations, the first step was to use a look up table to determine the underlying score associated with a given exemplar code. This was done for each question and each panelist. The conversion of the exemplar codes into the scores that each exemplar received permitted the summation of the values, calculation of averages (i.e., mean, median) across panelists.

After completing their ratings on the first question, the facilitator led a discussion of the rationale for why they selected the exemplars that they did. 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 replicated for each of the subsequent essay questions and the performance task with an exception that a group discussion did not occur after each question. For logistics purposes, the remaining four essay questions were evaluated by half the group as a split panel. Following their ratings on the essay questions, the full panel then replicated the judgment process for the performance task. After completing key phases in the process (e.g., orientation/training, operational rating) panelists completed a written evaluation form of the process.

Analysis and Results

Following the design of the process, each panelist reviewed 3 essay questions (1 as a full group and then 2 as part of their subgroup) and the performance task. For each, panelists were asked to select four borderline papers that represented the best non-competent responses (2) and the best competent responses (2). After the study, the scores for each of the selected borderline papers were identified and used to determine the level of performance expected for candidates at this level.

To calculate the recommended passing score on the examination from the panelists' judgments, the individual recommendations for each panelist were summed across the questions with each essay question being equally weighted and the performance task counting for twice as much as an individual essay question to model the operational scoring that will occur beginning with the July 2017 administration. Because some essay questions were evaluated by half the group per the design, mean and median replacement were used to estimate the individual recommendations. Mean and median replacement are missing data techniques that are used to approximate the missing values when panelists do not make direct judgments.

The strategy first calculates the mean or median for the available data and then replaces the missing values with the calculated values. This approach retained the recommended values across questions for the panelists while permitting calculations of the standard error of the mean and standard error of the median. The standard error is an estimate of the variability of the panelists' recommendations adjusted for the sample size

of the group. These values provide additional information for interpreting the results of the panelists' recommendations.

Following these judgments, we calculated the recommended score and associated passing rate when considering the written part of the examination. However, we needed to know what score on the total exam corresponded to this same pass rate. To answer this question, another step was needed to transform these judgments to the score scale on the full-length examination. After creating distributions of individual recommendations for the written part of the examination, to estimate the score for the full-length examination we applied an equipercentile linking approach to find the score that yielded the same percent passing as was determined on just the written component of the examination that panelists evaluated.

This methodology is characterized as equipercentile because the goal is to find the equivalent percentile rank within one distribution of scores and transform it over to another score distribution to retain the same impact from one examination to another or in this instance, from a part of the examination on which panelists made judgments to the full examination. This linking occurred applying the weight that 50% of the total score would be contributed by each component – written and MBE.

There are two important assumptions when applying equipercentile linking. First, we assume that the same or a randomly equivalent group of candidates are used to create the two score distributions. Second, we assume that the examinations are sufficiently correlated to support the interpretation. In this application, the same candidate scores were used from the written part to the full-length examination. In addition, the correlation between the written scores and the total score (of which the written scores are a part) was 0.97 suggesting a strong relationship between the distributions to support applying an equipercentile linking approach.

The summary results are presented in Table 2. The panel's recommended mean and median with the associated standard errors are included along with the impact and combined score associated with the recommendation, along with a +/- 2 standard error of mean or median. Individual ratings for each essay question, the performance task, and the summary calculations are included in Appendix C and have been de-identified to preserve anonymity of individual panelists. The summary results of these analyses are shown here in Table 2.

Table 2. Summary results with range of recommendations on written and combined score scales with impact (i.e., pass rate).

	Written Score - Mean	Combined Score – Mean (pass rate)	Written Score – Median	Combined Score – Median (pass rate)
-2 $SE_{Mean/Median}$	419	1414 (53%)	414	1388 (60%)
-1 $SE_{Mean/Median}$	424	1436 (47%)	419	1414 (53%)
Recommended score ($SE_{Mean/Median}$)	428 (4.47)	1451 (43%)	425 (5.60)	1439 (45%)
+1 $SE_{Mean/Median}$	432	1480 (36%)	431	1477 (37%)
+2 $SE_{Mean/Median}$	437	1504 (31%)	436	1504 (31%)

Panelists' Recommendations

Interpreting the results of the panelists' recommendations involves a combination of sources of evidence and related factors. The results shown in this section represent one of those sources, specifically, the ratings provided by subject matter experts on exemplars of performance from the California Bar Examination. Additional discussion of empirical and related policy considerations is provided in the *Evaluating the Cut Score Recommendations* section below.

The goal in analyzing the results of the panelists' judgments was to best represent the recommendation from the group. There are different ways this could have been done, each involving a measure of central tendency (e.g., mean, median). The mean calculation is the arithmetic average that most people are familiar with, however, it may not be the best representation of the group's recommendation when the distribution is skewed. For smaller samples or when extreme scores are observed in a distribution, the mean may be higher or lower than the group would have otherwise intended. In these instances, the median is calculated at the point where half the recommendations are above the value and half the recommendations are below the value to balance the effects of an extreme or outlier recommendation. When the mean and median do not converge, it is generally recommended that the median be used as the better representation of the central tendency of the observed score distribution. This approach is analogous to the data that are often shared with respect to housing prices in cities where a median is used to offset the effects of outliers on upper and lower end of the distribution.

Although the values calculated for the panelists were close, the mean and median recommendations did not converge. Therefore, the median likely serves as a better indicator of central tendency of the recommendation of the panelists. The median recommended cut score for the written portion of the exam based on all panelists' judgments was 423.75 and was rounded to the nearest observable score of 425 on a theoretical scale that ranges from 0 to 700 (i.e., 100 points for each essay question, 200 points for the performance task). To then determine how this recommendation would be interpreted with respect to a pass/fail decision, we evaluated the impact on a cumulative percent distribution using only the written component performance by applicants who took the July 2016 California Bar Examination.

To evaluate the impact of this recommendation, we found the location in the cumulative percent distribution of the written scores that corresponded with this value (i.e., 425). This value resulted in an overall impact of 46% pass and 54% fail based on the applicants who took the July 2016 California Bar Examination. To then determine the score on the full examination that corresponded to this impact, we then used an equipercentile linking approach to find the value on the combined score that corresponded to the same impact (i.e., 46% pass and 54% fail), and the corresponding value in the distribution yielded a score of 1439. The same process was followed in evaluating the mean score that was calculated for the group.

When collecting data from a sample, it is important to acknowledge that the results are an estimate. For example, when public opinion polls are conducted to gather perceptions about a given topic (e.g., upcoming elections, customer satisfaction), the results are reported in conjunction with methodology, sample size, and margin of error to illustrate that there is a level of uncertainty in the estimate. In selecting a representative sample of panelists for this study, we similarly collected data that resulted in a distribution of judgments from which we could calculate an estimate of the recommendation of the group.

Because the mean and median were calculated from a distribution of scores, it is also appropriate to estimate the variability in those recommendations to produce a range within which policymakers may consider the panel's recommendation. This range was calculated using the standard error of the mean and median. The standard error is an estimate of the standard deviation (i.e., variability) of the sampling distribution. To calculate the standard error of the median (SE_{median}), the standard error of the mean is first calculated and can then be approximated by multiplying that value by the square root of pi (i.e., 3.14159. . .) divided by two which produces a slightly wider range than the standard error of the mean. Though technical in nature, the Standard Error of the Median can also be interpreted conceptually as the margin of error in the judgments provided by the panel.

Given a median recommendation of 425 on the written section with a SE_{median} of 5.60, the range of recommended passing scores on the written score scale would be 414 to 436 which translates to a range of 1388 to 1504 on the combined score scale. This range would correspond to the interpretative scale of 139 to 150. If the mean recommendation range was used, it would correspond to a 1414 to 1504 which on the interpretive scale would be 141 to 150.

Process Evaluation Results

Panelists completed a series of evaluations during the study that included both multiple-choice questions and open-ended prompts. The responses to the questions are included in Table 3 and the comments provided are included in Appendix D. With the exception of Question 2 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					
Orientation to the workshop	4	0	0	9	11
Overview of the exam	3	0	0	12	8
Discussion of the PLD	4	0	1	5	14
Training on the methodology	3.5	0	2	8	10
2. Time allocation to Training	2	4	16	0	N/A
3. Confidence moving from Practice to Operational	3	1	1	15	3
4. Time allocated to Practice	3	1	6	10	3
6. Confidence in Day 1 recommendations	3	1	2	11	6
7. Time allocated to Day 1 recommendations	2	5	6	9	0
9. Confidence in Day 2 recommendations	3	0	1	11	6
10. Time allocated to Day 2 recommendations	3	1	3	8	6
12. Confidence in Day 3 recommendations	4	0	0	5	15
13. Time allocated to Day 3 recommendations	3	2	1	8	9
14. Overall success of the workshop	3	0	1	12	7
15. Overall organization of the workshop	4	0	0	7	13

Collectively, the results of the panelists' evaluation suggested generally positive perception of the activities for the workshop, 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 Cut 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. Threats to validity that were observed in these areas should inform policymakers' judgments regarding the usefulness of the panelists' recommendations and the validity of the interpretation. Evidence within each of these areas that was observed in this study is discussed here.

Procedural

When evaluating procedural evidence, practitioners generally look to 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 the Supreme Court represented a wide range of stakeholders: newer and more experienced attorneys and representatives from legal education who collectively included diverse professional experiences and backgrounds. The choice of methodology was appropriate given the constructed response aspects of the essay questions and performance task. Panelists' perspectives on the process were collected and the evaluation responses were very positive.

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 median on which the recommendation was based (5.60) was reasonable given the theoretical range of the scale (0-700) for the written component of the examination. This means that most panelists' individual recommendations were within about six raw score points of the median recommended value. Even considering the effective range of the scale (approximately 280-630), the deviation of scores across panelists did not vary widely. Similar variation was also observed for the mean recommendation. These observations suggest that panelists were generally in agreement regarding the expectations of which applicant responses were characteristic of the Minimally Competent Candidate.

External

Although external evidence is difficult to collect, some sources were available for this study that will be useful for policy makers in their consideration of the recommendations of the group. The use of impact data from applicants in California from the July 2016 examination can be used as one source of evidence to inform the reasonableness of the recommended passing score. In addition, the application of the recommendation to scores from other exams (e.g., February 2016, February 2017, July 2017) would also be useful to evaluate the potential range of impact. **This would be particularly valuable given the different ability distributions of applicants who take the examination in February versus July**. In addition, consideration of first time test takers versus repeat test takers is another potential factor because applicants who are repeating the exam do not represent the full range of abilities.

A limitation of the study was the inability to include items from the MBE as part of the judgmental process. Although it would have been a desired part of the standard setting design, the MBE was not made available to California for inclusion in the study. In using half of the examination for the study, we can make a reasonable approximation of a recommendation for the full examination (see, for example, Buckendahl, Ferdous, &

Gerrow, 2010). The correlation between the written and MBE scores is approximately 0.72 suggesting moderate to strong correlation, but with some unique variance contributed by each component of the examination.

In addition, passing scores on bar examinations from other states can also be used to inform the final policy. However, the use of data from other states should be done with caution for multiple factors. First, it is unclear whether other states have conducted formal standard setting study activities, so to evaluate comparability based solely on the passing standard may not support California's definition of minimum competency. Second, California has different eligibility criteria than other states that will have an impact on the ability distribution of the population of applicants. Specifically, California has a more inclusive eligibility policy than most jurisdictions with respect to the legal education requirements. Third, each jurisdiction may have a different definition of minimum competency as to how it is applied to their examination. These can contribute to different policy decisions.

To illustrate how California passing score compares with other, larger population jurisdictions, Table 4 is shown here for comparison purposes. The overall test taker passing rates are shown from 2007 to 2016 to illustrate the current rate, but also the trend in performance over time.

Table 4. Overall passing rates in selected states and nationally from 2007-2016.[9]

Jurisdiction	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
California	49%	54%	49%	49%	51%	51%	51%	47%	44%	40%
Florida	66%	71%	68%	69%	72%	71%	70%	65%	59%	54%
Illinois	82%	85%	84%	84%	83%	81%	82%	79%	74%	69%
New York	64%	69%	65%	65%	64%	61%	64%	60%	56%	57%
Texas	76%	78%	78%	76%	80%	75%	80%	70%	65%	66%
National Average	67%	71%	68%	68%	69%	67%	68%	64%	59%	58%

Note that across jurisdictions and for the nation, there has been a consistent, downward trend in overall passing rates beginning in 2014. Similar trends were observed for first-time test takers.[6] With passing scores for jurisdictions being held constant through policy and statistical equating, the changing variables of ability within the candidate population in terms of law school admissions, matriculation, as well as any influence on curriculum and instruction have likely contributed to this observed pattern. These data reinforce the caution of not simply relying on current passing scores used in other jurisdictions.

[9] Data for Table 4 were obtained NCBE 2016 Statistics document (pp. 17-20) and represent the combined pass rate for a given year across the February and July administrations. This report can be accessed: http://www.ncbex.org/pdfviewer/?file=%2Fdmsdocument%2F205.

Determining a Final Passing Score

The **standard setting meeting results and evaluation feedback generally support the validity of the panel's recommended passing score for use with the California Bar Examination**. Results from the study were analyzed to create a range of recommended passing scores. However, additional policy factors may be considered when establishing the passing score. One of these factors may include the recommended passing score and impact relative to the historical passing score and impact. The panel's median recommended passing score of 1439 (effectively 144 on the interpretative scale) converged with the program's existing passing score with the mean recommended passing score being slightly higher.

Factors that could be considered include the passing rates from other states that have similarly large numbers of bar applicants sitting for the examination. However, the interpretation of these results and the comparability are mitigated by the different eligibility policies among these jurisdictions and **California's more inclusive policies** along with the downward trend in bar examination performance across the country, particularly over the last few years. In some instances, the gap between California's applicants and other states has closed and in others, the gap observed in 2007 has remained essentially constant as the trend declined on a similar slope.

An additional factor warrants consideration as part of the policy deliberation. Specifically, the consideration of 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 cut 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 in nature 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 (i.e., 144) or adopt a value that is one to two standard errors below the recommendation (i.e., 139 to 141). 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 (i.e., 144) or adopt a value that is one to two standard errors above the recommendation (i.e., 148 to 150). Because standard setting is an integration of policy and psychometrics, the final determination will be policy driven, but supported by the data collected within this workshop and for this study more broadly.

References

Buckendahl, C., Ferdous, A., & Gerrow, J. (2010). Recommending cut scores with a subset of items: An empirical illustration. *Practical Assessment, Research & Evaluation, 15*(6). Available online: http://pareonline.net/getvn.asp?v=15&n=6.

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 – Panelist Information

Standard setting panelists.xlsx

Appendix B – Standard Setting Materials

The nomination form for panelists and documentation used in the standard setting are included below.

State Bar Standard Setting Study Nomin

Agenda

Training

Evaluations

Appendix C – Standard Setting Data

PLD Discussion for Minimally Competen

California Bar Standard Setting Da

Appendix D – Evaluation Comments

Each panelist completed an evaluation of the standard setting process that included several open-ended response questions. The responses provided to each are included below.

Day 1 – Training

- Lots of reading
- More time could easily be spent on the practice rating, but I doubt that it would make a difference in the outcome.
- Dr. Buckendahl trained us very effectively. He is engaging, clear, and attentive. I have confidence in him and the process. Good work!
- Perhaps it was the result of the lively discussions we were having, but a little more time for practice would have been ideal as I felt I was a bit rushed.
- More background information before initiating the process would be helpful
- Perhaps additional time spent as a group discussing not the themes/genres of knowledge for each subject, but on what it means to read an essay and decide whether a discussion of the theme is sufficient to communicate minimal competency.
- Not convinced this methodology is valid. Many of us clearly do not know some applicable law and these conclusions may therefore determine that incompetent answers amounting to malpractice are nevertheless passing/competent.
- Great and important discussion about minimal competencies on each exam answer discussed.
- It would have been helpful at the top to have a broader discussion about why the study is being done, what the Bar is hoping to learn, and how the individuals (participants) were selected.
- Would be helpful if watchers could be talking outside [the] room instead of in during review of essays.
- [Related to confidence rating] - only because some of my ratings were different from the majority. Otherwise, very confident.
- [Related to time rating] - Had to rush in order to have time for lunch.
- I think a broader discussion at the outset before the practice/identification of key issues would have been helpful. We all seemed to struggle with our own lack of knowledge and addressing that more up front may have helped us move along more efficiently.

Day 1 – Standard Setting

- I would have liked to know ahead of time that I would be "grading" 40 essays when I came in.
- I did not finish and felt rushed. More time for first question.
- Snacks for end of day grading would help :) I feel like I'm in a groove now and understand the concept of what I'm doing, but 30 tests to read is a lot at the end of a long day. Grateful we can finish in the a.m.!
- More time please
- I'm still not completely certain that I understand how we are qualified to do this without answers. It seems like this could have the overall effect of making it easier to pass?

- Although a lot of folks complained that we didn't "know enough" of subject matter, after reading 30 tests, yes we are - it became easier to spot the competent from the not competent. Perhaps this could be talked about at the outset to avoid this needless discussion altogether.
- I am concerned that an unprepared attorney, without the benefit of experience, studying, or a rubric, is not a good indicator of a minimally competent attorney. We all have an ethical duty to become competent. New lawyers/3 Ls do that by preparing for the exam. A more seasoned lawyer does that by refreshing recall of old material or by resort[ing] to practice guides. Having neither the benefit of studying nor outside sources, at least some of us may be grading with lack of minimum adequate knowledge. By studying for the exam, test-takers are becoming competent and gaining that minimal competency. Practicing professionals who become specialized may lose/atrophy that competence in certain field, which needs to be refreshed by CLG and other sources. So these scores may be of limited utility.
- It's too much. Too many questions to review.
- No changes
- Got 24/30 done [on the first day]

Day 2 – Standard Setting

- It was very difficult to read 60 essays in one day
- The discussion about where certain papers fall on the spectrum is helpful to let us know we are on the right track.
- We need breaks to stretch our bodies and we need to go outside, so our brains can get fresh air.
- It might be helpful to have some kind of "correct" sample answer to avoid having to go back and re-score or re-read for lack of knowing "the correct answer."
- I do NOT like being tricked into grading/reading 130 frigging essays! We should have been told that this is what the project was.
- Snacks were a great addition to the day.
- Thanks for the afternoon snacks!
- We did not follow the agenda which indicated we should build an "outline" for the "question." Instead, on Day 1, we outlined subject areas. There will not be consistency among the group. This was clear this AM when there was no agreement regarding Question 1. Each of the 30 essays was marked as the best no-pass or worst pass by at least one person. We should have outlined as a group.
- After initial "calibration" session on Day 1; and with more time, I feel confident about my ability to apply the PLDs to these essays.
- No changes

Day 3 – Standard Setting and Overall Evaluation

- This no doubt took a lot of work, so thank you to all staff and State Bar folks!
- The early activities and group discussion were helpful in allowing me to orient and direct what I ought to be doing for my recommendations. Perhaps a few more panelists to ease the burden would be helpful for the future!
- No changes

- I really found the time available to review the subject-matter answers to be very challenging. Trying to discriminate among those last four papers and a few on either side of them was difficult. An idea: have readers make their 3 initial stacks and identify not more than x (10?) papers that fall closer to the borderline. Do that for all answers. Then have readers spend last session choosing the "two and two" all at once.
- I'm not entirely sure I understand how what feels like an arbitrary process by 20 graders/panelists results in a less arbitrary cut score. Perhaps some additional information or process would be helpful.
- Although providing a scoring rubric would make categorization more consistent, it would do so in view of the thoughts of the author and not of the 20 panelists. Having no rubric was tough, but appropriate.
- Breaks between assignments
- Work with Dr. Buckendahl again. He was very careful, clear, and engaging. Well done!
- The performance test, unlike subject matter knowledge tests (essays) is much more amenable to this sort of standard setting. While, as with essays, we did not outline/rubric/calibrate, that is less necessary because of closed universe and the skills being tested.
- Overall, I think this process made sense. I was troubled that at least one of the panelists had clear familiarity with the existing exam and process and a clear knowledge of "right" answers as currently graded. I'm not sure everyone had a clear understanding of "minimally competent attorney" so we may have had different standards in mind.
- I'd like to be included in next steps or discussions. Other than just more grading/reading essays.
- I had a hard time with the time limit to review each answer. I am not clear if I was being too thorough, or I missed the lesson on how to move through answers at a quicker pace.