OREGON STATE BOARD OF BAR EXAMINERS

16037 SW Upper Boones Ferry Road, PO Box 231935, Tigard, OR 97281-1935
(503) 620-0222 or (800) 452-8260 • www.osbar.org

June 23, 2021

Chair, Joanna Perini-Abbott
Oregon State Board of Bar Examiners
16037 SW Upper Boones Ferry Road
Tigard, OR 97224

Re: Standard Setting Task Force Pass Score Recommendation for the Oregon State Bar Examination

Dear Board of Bar Examiners:

On September 14, 2020, Chief Justice Walters asked the Board of Bar Examiners (BBX) for its views as to the appropriate passing score on the UBE for Bar admission in Oregon. To provide depth to its analysis, the BBX established a 2021 Standard Setting Task Force (SSTF or Task Force), which was comprised of members and representatives of the court, BBX, Oregon law school deans, and Oregon State Bar staff. This SSTF met four times via remote meetings and reviewed several articles, reports, documents and data worksheets to encourage robust policy discussions around standard setting in the legal field.

2017 Task Force

Oregon became a Uniform Bar Exam (UBE) state in 2017, the same year an initial "Cut Score Task Force" was established to make a recommendation on Oregon's previous, non-UBE, pass score. That task force reviewed recent trends in Oregon's pass rates and the pass scores of other jurisdictions to recommend lowering the pass score from 284 to 274, which the BBX in turn recommended to the Supreme Court. The BBX also noted that, since the establishment of the New Lawyer Mentoring Program (NLMP), passing the bar examination was not the only rite of passage for new members. The NLMP provides hands-on, one-on-one learning opportunities for new members to acquire necessary skills and become

familiar with procedures and practices not addressed by the bar examination, whatever the pass score. The court approved that pass score change, resulting in a current cut score of 274 (with an exception of a lower, court-approved score (266) applied to the July and fall 2020 exams conducted during the pandemic).

2021 Task Force

The BBX determined that it would like a more scientific approach to be taken in this year's recommendation to the Oregon Supreme Court. The BBX, therefore, requested that the services of a psychometrician be engaged; one with expertise in conducting standard setting exercises for the bar exam in other jurisdictions.

The Oregon State Bar hired ACS Ventures to conduct the standard setting exercises that provided insight to the Task Force on pass scores. Many policy considerations go into the final pass score level, but it is generally recommended that it begin with a study by practitioners in the state bar who have familiarity with the work product in the legal marketplace and can spot the minimum level of competence that is acceptable, using the Rules for Admission's "essential eligibility requirements" as a framework for assessing minimum competence (see generally RFA 1.25).

The study was conducted on May 17 & 18, 2021, and consisted of panels of practitioners reviewing answers from a recent bar exam and determining if the answers meet the minimum level of competence. The panel was made up of lawyers from three categories: 1) lawyers who work in mid-to-large firms that oversee the work of associates who are new to the practice of law; 2) young lawyers who are new to the practice of law; and 3) lawyers who have a solo practice or work in small-firms with less than five lawyers (firms with no associates).

The completed psychometrician's report is attached to this letter for reference as Exhibit 1.

Recommendation

The 2021 Task Force was asked to review the current pass score, provide feedback, and make any appropriate recommendations. We relied heavily on the

psychometrician's report in our conclusion that a score between 268 and 273 falls within an acceptable passing score range. The Task Force found support for this range because at least two-thirds of the panelists within the psychometrician's study found that these scores represented the minimum competence standard. However, there was consensus to recommend a pass score of 270 based on a variety of factors, including consumer protection, the UBE pass scores of other Western States, the need for more Oregon lawyers, and issues surrounding equity and access to justice. Lowering the score from 274 to 270 would have resulted in a 3.8% increase in bar passage for those who took the bar exam in July 2019 (73.3% passage rate increased to a 77.1% passage rate).

The following documents were reviewed and relied upon by the SSTF in reviewing the previously mentioned policy considerations and making its recommendation and are provided as attachments to this letter:

- Psychometrician's report (Exhibit 1)
- 2017 California Standard Setting Exercise (Exhibit 2)
- Article—Need for Standards in Profession (Exhibit 3)
- Article—Standard Setting for High Stakes Professional Exams (Exhibit 4)
- California Supreme Court Lowering Score Based on 2017 Study (Exhibit 5)
- California Cut Score Simulation Analysis (Exhibit 6)
- Article—New York Study of the UBE (Exhibit 7)
- Oregon Membership versus Population spreadsheet (Exhibit 8)
- Legal Needs Executive Summary (Exhibit 9)
- Western UBE Bar Passage Scores (Exhibit 10)

Supplemental Information

A bar exam pass score is intended to represent the numerical definition of the minimum level of competence. However, given the number of complex factors that affect assessing competence through the bar exam, it is impossible to establish a pass score that will provide a fool-proof separation between those who rise to the level of the minimally competent and those who do not. The nature of the exam and the assessment of its answers lead to occasional false positives and false negatives.

A false positive occurs when an examinee passes the bar exam, but their true knowledge and skill do not meet the minimum competence standard. A false negative occurs when an examinee fails the bar exam, but they have requisite knowledge and skill to meet the minimum competence standard. In standard setting, false positives and negatives are assumed to occur on every exam; generally are represented by a bar exam score near the pass score level; and represent statistical anomalies that make up a small percentage of all examinees.

While false positives and negatives might be statistical anomalies, the life-altering impact they have on applicants who fail the bar exam or on legal consumers harmed by a lawyer who is not minimally competent should be addressed in the setting of the standard. There are many reasons and theories about the existence of false positives and negatives, but they are generally believed to result from at least one of the following factors:

1. Issues related to the questions asked on a particular exam;
2. Issues with the testing environment for a particular exam;
3. Events that occurred in an examinee's life temporally close to a bar exam;
4. Access to adequate exam preparation materials and study environments;
5. Sufficient resources and time to adequately prepare for an exam;
6. The questions presented on a particular exam perfectly match an unqualified examinees knowledge; and
7. Cheating on the bar exam.

While the National Conference of Bar Examiners and the BBX regularly look at ways to remediate the effects of the above referenced factors, no institution has found a solution that completely eliminates the possibility of false positives or negatives. However, changes in the pass score are generally believed to have a direct impact on the chances of false positives or negatives occurring on a bar exam. The lowering of a pass score is generally believed to increase the possibility of a false positive, while also lowering the possibility of a false negative. The inverse is true if a pass score is raised.

While the pass score ultimately represents a standard of competence, that score also represents decisions made on certain policy issues or the weight given to

those policy issues, and the standard setting body's tolerance for false positives or false negatives. While the Task Force reviewed many policy issues and factors in reaching its decision, the following list represents the issues and factors deemed most influential for this recommendation (materials found to be persuasive on these issues and factors are referenced after each issue and/or factor):

1. Protection of the general public from the actions of lawyers who are not minimally competent and the role standards play in this protection (See Exhibits 2, 3, 4);
2. Disparities in bar exam outcomes between examines from the dominant culture population and various non-dominant culture populations (See Exhibits 5, 6, 7);
3. The need to increase the total number of active bar members to address access to justice issues created by Oregon's growing general population (See Exhibits 8 and 9);
4. A lack of available legal services in certain non-dominant cultures and populations, and the lack of bar members from these same cultures and populations (See Exhibit 9); and
5. Standard setting decisions made by other Western UBE states in establishing their pass score (See Exhibit 10).

The results of the May panel study provided the Task Force with keen insight into how the term "minimum competence" is actually applied in the current Oregon legal marketplace, and the bar exam scores that are considered a good reflection of this definition. While the current 274 pass score was viewed as a good representation of at least minimum competence, many lower scores were viewed as substantially achieving this standard as well. Due to concerns about the previously mentioned policy issues, and a lower tolerance for false negatives brought on by the panel study, the Task Force reached a consensus that the pass score should be lowered.

Given the number of issues and factors that may influence the Court's decision on the pass score, the Task Force unanimously agreed that a range of pass scores should be recommended to the Court. Additionally, it was agreed that a singular pass score should, if possible, be selected within that range to represent a score that all Task Force members believe adequately addresses the concerns of risk tolerance that are embedded within the "minimum competence" definition while

simultaneously making important progress on issues of equity and access to justice that are addressed, in part, by a reduced score.

While two-thirds of the panelists in the psychometrician's study supported an argument that 268 should be viewed within the minimum competence standard, some Task Force members thought that this number would present too high of a tolerance for false positives on the bar exam; thus, no consensus could be reached in recommending this as a singular score. However, it was unanimously agreed that 268 would be a good representation for the bottom of the recommended range, because the Court may not have the same risk concerns or may place more emphasis on competing considerations. As there was consensus that the pass score should be lowered, 273 is recommended as the peak of the range because 273 was the top result of two-thirds of the "pass" results within the psychometrician's study. With consensus achieved on the range of 268 to 273, all members of the Task Force agreed that 270 is the score that may come closest to balancing the ideals of the minimum standard, as well as achieving consistency with other Western states, while still adequately addressing the concerns of members on various policy issues and risk tolerances.

Task Force Members

The Task Force included the following voting members:
Chair, Caroline Wong (BBX Member)
Cassandra McLeod-Skinner (BBX Member)
Michael Slauson (BBX Member)
Helen Hierschbiel (CEO of Oregon State Bar)
Dean John Parry (Lewis and Clark Law School)
Dean Megan McAlpin (University of Oregon Law School)
Dean Jeffrey Dobbins (Willamette University School of Law)
Justice Meagan Flynn (Oregon Supreme Court)

Liaisons to the Task Force included the following:
Lisa Norris-Lampe (Appellate Legal Counsel, Oregon Supreme Court)
J.B. Kim (Diversity & Inclusion Director, Oregon State Bar)
Troy Wood (Regulatory Counsel, Oregon State Bar)

Conclusion

Based on the foregoing, the voting members of the Task Force make the following three recommendation to the Oregon Supreme Court related to the Oregon Bar Exam pass score: 1) that the pass score should be lowered from its current level of 274; 2) that the pass score should probably be lowered to a score within the range of 268 to 273; and 3) that the score that possibly best reflects the competing interests that go into a pass score decision is 270.

Sincerely,

Caroline Wong
Chair, 2021 Standard Setting Task Force
Oregon State Board of Bar Examiners

ACS VENTURES
BRIDGING THEORY & PRACTICE

Cut Score Review for the Oregon Bar Exam

Final Report

June 11, 2021

Submitted By:
Andrew Wiley, Ph.D.
Office: 917.885.0858
awiley@acsventures.com

Contents

Executive Summary

As part of the review and maintenance of the Oregon Bar Exam, the Oregon State Bar contracted with ACS Ventures, LLC (ACS) to complete a review and evaluation of the current exam cut score. To conduct this review, ACS managed and facilitated a 2-day workshop with a committee of current Oregon based lawyers that focused on a review of exam candidates' performance on the Bar Exam and whether their performance was consistent with the professional experience and expectations of the committee members.

The purpose of this study was to complete a review of the current cut score used to make Pass/Fail judgment for candidates on the Bar exam. The meeting was designed to enlist a committee subject matter experts (SMEs) to serve as reviewers of candidate performance on the Oregon Bar Exam and to evaluate whether candidate performance was consistent with their professional expectations and experience with lawyers practicing in the state of Oregon.

Prior to the workshop, the Oregon Bar shared de-identified candidate score data for the June 2019 test administration of the Oregon Bar. Using this data, ACS identified a sample of candidate responses that would be used during the workshop. The candidate responses were classified into two waves of eight candidate samples each. The first wave was designed to provide a somewhat broad range of scores, while the second was designed to be slightly more focused, based upon the feedback provided during the first wave.

During the workshop, the committee members discussed their knowledge and expectations for candidates who should pass the bar examination. They also reviewed the general scoring rules for all sections of the Oregon Bar exam. The committee members then proceeded with a review of a single candidate in which they reviewed the candidates' response to all MEE and all MPT prompts on the test. After reviewing the responses, each committee members provided a holistic Pass/Fail judgment for the candidate. After all committee members completed this work, the group held a group discussion focused on the specific characteristics of the response and how they arrived at their specific rating. Once this was completed, the committee members independently completed their ratings of the remaining seven candidates in the Wave 1 sample.

After completing their initial ratings, the committee members discussed each of the sample candidates. After the discussion, the committee members were given the opportunity to update their ratings for each of the candidates. After completing this second round of ratings, panelists followed an identical process in reviewing the second set of eight candidate samples. As with the first wave, panelists completed an independent review, followed by group discussion, followed by an opportunity to update their ratings. Based upon the review of both the first and second wave of candidate samples, a few key findings can be observed:

- There was consistent agreement among the committee members that candidate that were reviewed and had scored at or above the current cut score demonstrated performance that was consistent with a Pass designation.
- There was consistent agreement that candidates dramatically lower than the current cut score (i.e. < 265) did *not* demonstrate performance that was consistent with a Pass designation.
- Candidates with test scores slightly lower than the current cut score (i.e. 268-273) had fairly strong support that their performance was consistent with performance that was consistent with a Pass designation. However, the support was not as consistent with those at higher score points.

Overall, the committee members appeared to feel comfortable with the process that was followed and seem to fully understand the task and data that was provided to them. This information can be utilized by the Oregon Bar as it considers options for modifying or keeping the current cut score as is.

Introduction and Overview

As part of the review and maintenance of the Oregon Bar Exam, the Oregon State Bar contracted with ACS Ventures, LLC (ACS) to complete a review and evaluation of the current exam cut score. To conduct this review, ACS managed and facilitated a 2-day workshop with a committee of current Oregon based lawyers that focused on a review of exam candidates' performance on the Bar exam and whether their performance was consistent with the professional experience and expectations of the committee members.

The purpose of licensure examinations like the Oregon Bar exam 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 cut score review was designed to focus expert judgments about the level of performance that aligns with minimal competence.

Assessment Design

The Oregon 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. All candidates complete all components of the Multistate Bar Exam (MBE), which includes 200 multiple-choice questions (MBE), six essay questions (MEE), and two performance tasks (MPT)[1]. The combined score for the examination weights the MBE at 50% and the constructed response components at 50% with the MEE weighted at 30% and the MPT weighted at 20%. A decision about passing or failing is based on the performance of applicants on the entire 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.

Study Purpose and Validity Framework

The purpose of this study was to complete a review of the current cut score that distinguished the performance characteristics of someone who was minimally competent from someone who was not competent. To complete this review, Dr. Andrew Wiley and Ms. Kelley Wheeler from ACS facilitated a virtual workshop on May 17-18, 2021. The meeting was designed to enlist subject matter experts (SMEs) to serve as reviewers of candidate performance on the Oregon Bar exam and to evaluate whether candidate performance was consistent with their professional expectations and experience with lawyers practicing in the state of Oregon.

To evaluate the cut score review, 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

[1] 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.

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 results of the review. The State Bar is receiving this report as part of their initial review of the passing score to assist in the overall review of the Oregon Bar Exam. Based upon this report as well as additional information, the Oregon State Bar may elect to adopt a policy recommendation for an updated cut score, that will then be provided to the Oregon 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 Oregon Bar Exam.

Procedures

The cut score review followed a process consistent with the Analytic Judgment Method for setting a cut score (AJM; Plake & Hambleton, 2001)[2]. 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 Oregon Bar Exam (see Buckendahl & Davis-Becker, 2012).

Panelists

A total of 15 panelists participated in the workshop[3]. The panelists were licensed attorneys with an average of 13 years of experience in the field. Panelists were recruited by task force members to represent a range of stakeholder groups. These groups were defined as Recently Licensed Professionals (panelists with less than five years of experience), Management Level Professionals (panelists with ten or more years of experience who review associate work as a regular part of their practice), and Small Firm or Solo Practitioners (panelists with six or more years of experience who work in firms of three or fewer licensed attorneys). Note that some panelists were associated with multiple roles. Some of the experienced attorneys also served as part-time adjunct faculty members at law schools, but all maintained a full-time law practice. 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, a representative from the Oregon State Bar also attended portions of the workshop. The representative did provide clarification and further explanation on some Bar exam related inquiries from panelists, but the representative did not participate in discussions regarding candidate performance and did not provide any recommendations as part of the process. All panelists signed

[2] Within the psychometric literature, there are multiple methods for a formal process of evaluating a test and developing a cut score recommendation. The purpose of this workshop was to review the cut score; but *not* to provide a formal recommendation; therefore the methodology did not follow the exact processes, but was consistent with the overall AJM methods.

[3] Members of the Standard Setting Task Force were assigned the task of nominating licensed Oregon lawyers for the standard setting panel. Members were asked to nominate practitioners who hold themselves to the highest professional standards, care about the public image of the profession, and have a genuine concern for protecting the legal consumer from incompetent representation. The panelist recommendations were then reviewed on their merits by the entire Standard Setting Task Force, who selected participants to represent diverse backgrounds with respect to experience, practice areas, size of firms, geographic location, gender, and race/ethnicity.

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.

Table 1. Summary of panelist demographic characteristics.

Years of Practice	
5 years or less	40%
6 to 11 years	27%
More than 10 years	33%
Primary Employment Type	
Large-firm	40%
Small-firm	20%
Solo practice	27%
District Attorney	7%
Other Government	7%

Practice Areas	
Business Formation/Advisement	33%
Civil Litigation	53%
Contracts/Transactional	60%
Criminal Law	7%
Domestic Relations	7%
Employment Law	20%
General Practitioner	13%
Personal Injury	13%
Probate/Estate Planning	13%
Real Estate Transactions/Litigation	20%
Regulatory Compliance	7%
Workers Compensation	7%

Demographic Data:

Gender:	
Female	53%
Male	47%
Non-binary	0%
Trans	0%
Race/Ethnicity:	
White/Non-hispanic	73%
BIPOC	27%

Location of Practice:	
Portland Metro	60%
Oregon Coast	13%
Southern Valley	13%
Central Oregon	7%
Eastern Oregon	7%

Candidate selection

Prior to the workshop, the Oregon Bar shared de-identified candidate score data for the June 2019 test administration of the Oregon Bar. Using this data, ACS identified a sample of candidate responses that would be used during the workshop. To focus the activities of the committee, candidate responses were selected that were reasonably close to the current cut score. When candidate samples were identified, the entire set of candidate responses were selected and reviewed.

The candidate responses were classified into two waves of eight candidate samples each. The first wave was designed to provide a somewhat broad range of scores, while the second was designed to be slightly more focused, based upon the feedback provided during the first wave. Prior to the workshop, a greater number of candidate samples were identified and prepared. Because of this, the second wave could include more candidates who had scored somewhat close to where the committee felt comfortable and could avoid candidate responses that were dramatically different than where the committee was focused.
All candidate responses were prepared with the entire candidate response included in a single PDF file. The PDF file was secured on the ACS SharePoint site with access provided to the panelists on the first morning of the workshop. Access to the SharePoint was discontinued at the conclusion of the workshop. All candidate responses were completely de-identified so that the name and information regarding the candidate could not be determined.

Workshop Activities

The Oregon Bar Exam cut score review meeting was conducted virtually on May 17-18, 2021, using the ACS Zoom platform. Prior to the meeting, participants were informed that they would be engaging in tasks that would result in a review of the current cut score and that they would be reviewing samples of candidate performance from a previous Oregon Bar test administration. They were also provided descriptions of the six MEE prompts and the two MPT prompts that were administered in June of 2019 and that would be reviewed during the workshop. Along with the prompts, panelists were also provided information on the scoring of each of the MEE and MPT prompts. The cut score review consisted of orientation and training, discussion of the minimally competent candidate, the review of samples of candidate performance, and group discussion. Andrew Wiley, Ph.D., served as the facilitator for the meeting along with Ms. Kelley Wheeler. Workshop orientation materials are provided in Appendix A.

Orientation

The meeting commenced on May 17th with Dr. Wiley providing a general orientation for all panelists that included the goals of the meeting. Additionally, Mr. Troy Wood, Regulatory Counsel for the Oregon State Bar provided additional information for how the information gathered during this workshop would be used by the Oregon State Bar in their review of the current cut score. 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 that was based upon 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. An initial proposal for a definition of minimal

competence was presented to the committee. The initial definition focused on four essential traits for a minimally competent candidate (MCC), more specifically:

> 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 B.

Initial rating

After the review of the MCC, the panelists proceeded to complete their review of the first sample candidate. The panelists were instructed that they were to review the first candidate's response to all six MEE prompts and both MPT prompts and make an overall holistic judgment on whether they believed the candidate had demonstrated sufficient knowledge and skills to be considered a passing candidate. Each panelist independently reviewed the performance of the first candidate and completed the holistic Pass/Fail judgment. Once all panelists completed their review, the group was reconvened to allow for a group discussion of the panelists' judgments. This discussion included what features from specific prompts they believed supported their Pass/Fail decision, the candidate's consistency of performance across the prompts, and the particular strengths and weaknesses of the candidate response.

After completing the group discussion of the first candidate, the committee members proceeded to independently review the MEE and MPT responses for an additional seven candidates. This work was completed independently and offline. Upon completing their review and ratings, the ratings were submitted

to Dr. Wiley and Ms. Wheeler so they could be summarized. The independent review of the candidates completed the activities for Day 1 of the workshop.

Discussion of the Wave 1 candidate samples

At the beginning of Day 2, Dr. Wiley facilitated a discussion of the candidate samples from Wave 1. During the discussion, the committee members discussed the strengths and weaknesses of each candidate, the features of the responses that impacted their holistic Pass/Fail judgments, and consistency of the response across the six MEE prompts and the two MPT prompts.

After this discussion, the committee members were provided information on the overall score of each the candidates they had reviewed. Using this information, further discussion was facilitated across the panel focused on whether the assigned scores were consistent with the committee members ratings and whether any panelists appeared to have a notably different score than what was expected. Finally, after all group review was completed, the committee members were provided an opportunity to update their initial Pass/Fail judgments. These updated ratings were submitted to Dr. Wiley and Ms. Wheeler via email.

Review of Wave 2 candidate sample and discussion

Based upon the initial recommendation of the committee members, a second wave of candidate responses was identified by Dr. Wiley and Ms. Wheeler. This second wave was designed to be somewhat more focused on the score points that the committee members appeared to be most interested in discussing. As with the 1st wave of candidate responses, all responses were grouped into PDF files and placed on the ACS SharePoint site. Access to the site was sent to the panelists as the discussion of the first round was reaching its conclusion.

The committee members completed an independent review of the eight candidate samples and submitted their judgments to Dr. Wiley and Ms. Wheeler. After all committee members had completed their independent judgments, the group reconvened on Zoom and Dr. Wiley facilitated a discussion of their judgments and the candidate samples. As with the Wave 1 sample, the discussion focused on the strengths and weaknesses of each candidate, the features of the responses that impacted their holistic Pass/Fail judgments, and consistency of the response across the six MEE prompts and the two MPT prompts. After this discussion the score information for each candidate was also provided to support additional discussion.

Analysis and Results

For each of the candidate samples that were reviewed, the percentage of committee members who supplied a judgment of Pass was calculated. The percentage of committee members that assigned a Pass judgment to each of the Wave 1 candidate responses is reported in Table 2 below. As can be seen in the table, there was 100% agreement in their ratings for both the highest and the lowest scored sample, with all committee members assigning a Pass to the candidate with a score of 278, and no committee members assigning a Pass to the candidate who scored at 264. There also appeared to be strong agreement for the candidate who scored at 267, with only 13.3% of the committee members assigning a Pass. Between the scores of 268 and 275, there were less agreement amongst the committee members, with as few as 46.7% of committee members indicating the candidate should pass and up to 86.7% of committee members indicating another candidate should pass. Most interestingly, the committee members did not necessarily follow the same pattern as the scores assigned to the candidate, with some of the lower scored responses (score of 270 passed by 86.7% of committee members) receiving higher overall judgments from the committee members.

Table 2: Percentage of committee members assigning a Pass judgment to each candidate response in Wave 1[4]

Cand. ID	Total Score	% Pass
1.1	264	0.0%
1.2	267	13.3%
1.3	268	46.7%
1.4	270	86.7%
1.5	271	53.3%
1.6	274	86.7%
1.7	275	66.7%
1.8	278	100.0%

Table 3 provides the same information but for the second wave of candidate samples that the committee members reviewed. Interestingly, the ratings provided by the committee members appear to present a pattern that is more consistent with the scores assigned to the candidates. For the Wave 2 candidates, there was strong agreement amongst the committee members that the lower scoring candidate responses should not receive a Pass rating. Alternatively, for those at the higher end, starting at a score of 268, a high percentage of committee members indicated that they believed the candidate response should pass, and that number increased as responses close to the current cut score were reviewed (i.e., 86.7% of committee members said the response at a score of 274 should pass).

Table 3: Percentage of committee members assigning a Pass judgment to each candidate response in Wave 2[5]

ID	Score	%
2.1	260	20.0%
2.2	264	6.7%
2.3	265	26.7%
2.4	268	66.7%
2.5	271	66.7%
2.6	273	66.7%
2.7	274	86.7%
2.8	275	80.0%

[4] The data in Table 2 is based upon the 2nd round of ratings provided by the committee members. For reference, the initial first round is summarized and provided in Appendix C.
[5] The data in Table 3 is based upon the 2nd round of ratings provided by the committee members. For reference, the initial first round is summarized and provided in Appendix C.

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 4 and the comments provided are included in Appendix C. For all questions, higher ratings indicate the panelists had more comfort or confidence in the process and/or outcomes of the workshop.

Table 4: Evaluation results for the cut score review workshop

	Median	1 - Lower	2	3	4 - Higher
1. Success of Training					
A. Orientation to the workshop	3	0	2	7	4
B. Overview of the exam	3	0	0	8	5
C. Discussion of the PLD	3	0	0	10	3
D. Training with the MEE	3	0	1	8	4
E. Training with the MPT	3	0	0	9	4
2. Confidence defining the Minimally Competent Candidate	3	0	0	11	2
3. Time allocated to Minimally Competent Candidate	3	0	0	12	1
4. Confidence discussing the MEE	3	0	0	8	5
5. Time allocated to the MEE	3	1	1	8	3
6. Confidence discussing the MPT	3	0	0	8	5
7. Time allocated to the MPT	3	0	1	9	3
8. Overall success of the workshop	3	0	0	9	4
9. Overall organization of the workshop	3	0	0	9	4

Collectively, the results of the panelists' evaluation suggest generally positive perception of the activities for the workshop, their ratings, and the outcomes. The ratings were slightly lower for some of the questions related the time allocated for the tasks, which are likely a reflection of the challenge of the task and the requirements to complete a review of multiple candidates in a tight time window.

Evaluating the Cut Score Recommendations

To evaluate the workshop, we applied Kane's (1994; 2001) framework for validating standard setting activities. Because Kane's framework is focused on standard setting, or determining a cut score, the framework is not directly applicable, but the framework can provide some useful information to consider. 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. Traditionally, this would be evaluated using the consistency of the recommended cut scores. In our workshop, we did see more consistency with the ratings as we moved from Wave 1 to Wave 2. In Wave 2, we observed more agreements amongst the committee members, indicating the group started to reach a consensus on behavior indicative of passing performance.

External

Traditionally, external evidence is some of the most difficult evidence to collect. In some scenarios, the passing rate that would be observed given a recommended cut score can be compared to other measures to determine if they are consistent with these other observed measures. For this workshop, the current passing score and pass rate in Oregon can be reviewed and compared with other neighboring states.

Table 5 presents the historic passing rates for Oregon along with a comparison to the neighboring states of Oregon. As can be seen, the pass rate of first-time test takers as well as the total group of test takers was the highest for the state of Oregon in 2017 to 2019.

Table 5: Historical pass rates for the July test administration of multiple state bar exams

	2015		2016		2017		2018		2019	
	1st-time	All	1st-time	All	1st-time	All	1st-time	All	1st-time	All
OR	68%	(60%)	62%	(58%)	84%	(79%)	78%	(73%)	84%	(75%)
CA	60%	(47%)	56%	(43%)	62%	(50%)	55%	(41%)	64%	(50%)
ID	NA	(69%)	NA	(73%)	NA	(76%)	NA	(66%)	NA	(64%)
NV	NA	NA	NA	(51%)	NA	(66%)	NA	(57%)	NA	(61%)
WA	81%	(76%)	76%	(70%)	76%	(72%)	75%	(69%)	76%	(68%)

Cut score considerations

As the Oregon Bar reviews the results of this workshop, a number of critical considerations can be factored into this process. First, it is important to note that the workshop conducted in May was *not* designed to provide a single cut score recommendation for the Oregon bar. Bar Exam pass scores typically represent an outcome that was derived from reviews, studies, research, analysis and debate on at least some of the subject matters included in the following non-exclusive list: 1) the minimum professional standard; 2) psychometric studies; 3) the consumer protection role that bar admissions plays for the profession; 4) statistics related to malpractice and attorney discipline for newer lawyers vs. established practitioners; 5) regulatory requirements that assist newer lawyers in the practice of law and offer consumer protection; 6) causality between current lawyer populations and access to justice for underserved populations; 7) expanding diversity and inclusion within the profession; and 8) the strengths and weaknesses of the current bar exam in assessing whether a person is minimally competent. Given the complexity that these issues can have on the ultimate pass score decision, our review was designed to evaluate if the panelists generally agreed with the current Pass/Fail designation assigned to current candidates of the Oregon Bar. After the panel found the current pass score represented competent candidate performance, the committee was then asked to review candidate samples below the cut score to identify the lowest score that a super-majority of panel members believed represented minimum professional competence.

In general, there was strong support from the committee members for all of the candidates reviewed that received a Pass designation on the current Bar exam. Every passing candidate was considered by a minimum of 66.7% of the committee members as an appropriate Pass, with most of the candidates supported by approximately 80 to 86% percent of committee members. In addition, candidates who scored slightly lower (e.g., scores between 268 and 273) had fairly strong support as well. However, it was not as consistent with one candidate only supported by 47% of the committee members and another only supported by 53% of committee members. Candidates who scored at the lower end of the observed scores (e.g., scores between 260 and 263) generally had a fairly small percentage of committee members who thought they should pass the exam.

To further aid in the considerations of the Oregon Bar, the performance of candidates on the July 2019 exam was reviewed and the hypothetical pass rate that would be observed across multiple score points was calculated. Table 6 shows the pass rate that would have been observed across multiple hypothetical pass rates. It should be noted that the pass rates are based on the total group, not first-time takers.

Table 6: Hypothetical pass rates for the July 2019 on multiple possible cut points

Score	264	265	266	267	268	269	270	271	272	273	274
Projected 2019 Pass Rate	82.6%	81.2%	80.4%	79.6%	78.7%	77.7%	77.1%	76.6%	75.2%	74.7%	73.3%

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).

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 – Standard Setting Materials

Training Slides

OR Bar Workshop Orientation 12May20

Workshop evaluation

OR Bar Cut Score Review Evaluation 12

Appendix B – Standard Setting Information and Data

Definition of Minimally Competent Candidate (MCC)

Excel file with Raw Wave 1 committee ratings

OR Bar Analysis
Wave 1 Raw 11June2

Excel file with Raw Wave 1 committee ratings

OR Bar Analysis
Wave 2 Raw 11June2

Appendix C Candidate classification summary for Round 1 ratings

Table C.1: Percentage of committee members assigning a Pass judgment to each candidate response in Wave 1, Round 1

Cand. ID	Total Score	% Pass
1.1	264	13.3%
1.2	267	33.3%
1.3	268	60.0%
1.4	270	66.7%
1.5	271	53.3%
1.6	274	86.7%
1.7	275	46.7%
1.8	278	100.0%

Table C.2: Percentage of committee members assigning a Pass judgment to each candidate response in Wave 2, Round 1

Cand. ID	Total Score	% Pass
2.1	260	40.0%
2.2	264	0.0%
2.3	265	26.7%
2.4	268	66.7%
2.5	271	66.7%
2.6	273	53.3%
2.7	274	80.0%
2.8	275	80.0%

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.

Please provide any comments about the discussion of the Minimally Competent Candidate (MCC) that would be helpful in planning future workshops:

- Thought what we did was just about right.
- I think I just needed a little more training on how to grade the exams. Other than that it was great!
- A lot more work than I had originally thought there would be, but if I had known I don't know if I would have agreed to participate. Overall, I appreciated that the presenters did value our time and input and acknowledged it, which helped a lot.
- I truly don't think you should give us the score that each applicant actually received - it absolutely skewed my thinking. I think maybe either rank them or just say whether they passed or not.
- This is not exactly responsive, but part of the problem is that I do not believe that's what the intent of the Bar Exam is, and therefore it becomes difficult to judge exams based on that as the premise. I believe the purpose of the Bar Exam may have minimal competence as one aspect, but also is in part an exercise in forcing people to work hard at memorizing a bunch of stuff, to be pain in the butt to study for because we don't want it to be easy to become an attorney. And so failing to memorize a bunch of stuff is a detriment to one's performance on the exam, and yet through the lens of an MCC, it's hard to argue that memorizing hundreds of pages of blank letter law is part of minimal competence. So there's a disconnect in the work of reviewing essays that test memorization and then evaluating whether the person has minimal competence.

Please provide any comments about the discussion of the MEE that would be helpful in planning future workshops:

- I think it would have been helpful to instruct us to read and review the scoring materials before the first meeting so I could be more prepared.
- See answer 4 above. ("I truly don't think you should give us the score that each applicant actually received - it absolutely skewed my thinking. I think maybe either rank them or just say whether they passed or not.")
- To echo earlier comments, it seemed that across the board reviewers gave greater weight to the MPT and less to the MEE, because the MEE was about memorization, while the MPT was more of a real world test. But if the MEE is part of the exam, then we need to accept that the purpose of the exam is not just about seeing if someone is ready to work on a "real world" project. We need to acknowledge that knowing black letter law is part of the requirement, and the MEE is given roughly the same (actually greater) value than the MPT and so reviewers should weigh it roughly equally.

Please provide any comments about the discussion of the MPT that would be helpful in planning future workshops:

- I think it would have been helpful to instruct us to read and review the scoring materials before the first meeting so I could be more prepared. It was provided to us, but I didn't realize it would benefit me to read it in advance. I didn't know that I would have to actually apply it like we did.
- See answer 4 above ("I truly don't think you should give us the score that each applicant actually received - it absolutely skewed my thinking. I think maybe either rank them or just say whether they passed or not.")

Please provide any comments about the workshop activities that would be helpful in planning future workshops:

- Good first round, I think it can be improved upon but good start.
- Again, sorry, didn't know this question was coming, but "I truly don't think you should give us the score that each applicant actually received - it absolutely skewed my thinking. I think maybe either rank them or just say whether they passed or not." What you guys do is FASCINATING. I wondered how this would work and when I figured it out, I felt a click in my head. Great stuff. Thanks for helping us have a strong Bar.
- I found the time to review all the exams to be way too compressed. It was every bit a time crunch as a rush project for a client, to the point where I was forced to skim over aspects of certain ones. So I would suggest that either there is more time between passing out the materials and returning feedback reviews, or fewer of them (even one fewer in each batch would have made me feel much more comfortable about the review time).
- It was as organized as it needed to be. Some informality was an welcome part of the experience. Don't take the fact that I put "Organized" instead of "Very Organized" as a criticism.

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

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.

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

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

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

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?

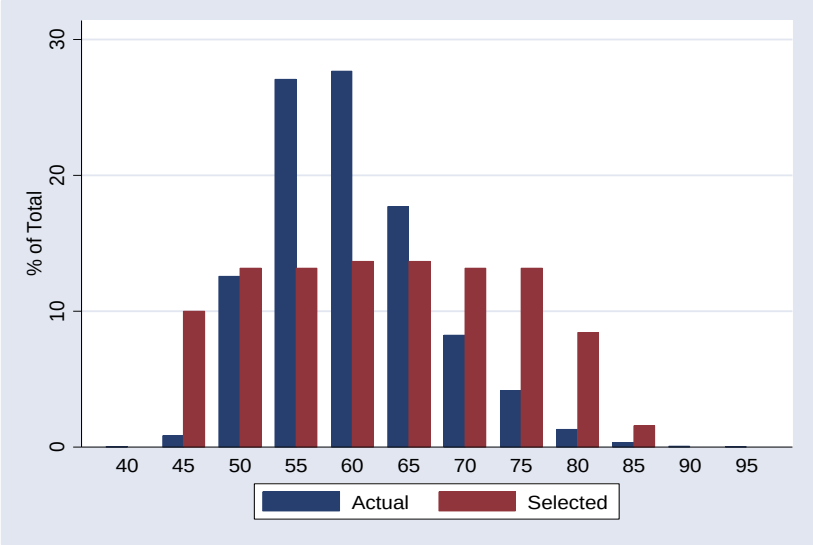
The results of this discussion and the illustrative characteristics of MCC performance for each of the subject areas that were included in this study are included as an embedded document in Appendix C.

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

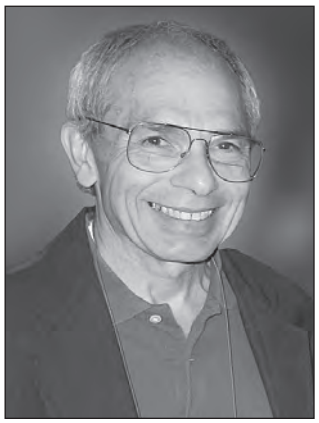
- This no doubt took a lot of work, so thank you to all staff and State Bar folks!
- The early activities and group discussion were helpful in allowing me to orient and direct what I ought to be doing for my recommendations. Perhaps a few more panelists to ease the burden would be helpful for the future!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

The Testing Column

Standards? We Don't Need No Stinking Standards!

by Mark A. Albanese, Ph.D.



Imagine . . . a world of people who all have outstanding innate academic and real-world talents and who all receive the best education possible, and where opportunity and jobs are plentiful in all occupations. Anyone who wants to go to law school can go to law school and will graduate with honors. There is no need for licensing tests, because all law school graduates are so far above average that they can't even see average, it is so far below them; and all law school graduates go on to become Supreme Court justices . . . .

The reality is that not everyone has the innate academic and real-world talents to succeed in a profession like law, even if they have a burning desire to be a lawyer. Given that the numbers of students applying to law school are at lows not seen since the 1970s, and that there are now far more law schools than there were back then, academic standards for admission and graduation at some law schools do not signify what they may have in the past. Although bar passage has been generally declining for over a decade, it started to cascade downward in July 2014. In response to the declining passing rate, law school deans in many jurisdictions have called for jurisdictions to consider re-evaluating their passing standards; and two large jurisdictions, Texas and California, have formally begun to do so. Clearly, incompetent lawyers can have disastrous consequences for their clients. In this environment, the bar admissions process is the last line of defense against incompetent lawyers being allowed to practice.

But that is not the only perspective. An opposing view comes from graduating law students who have put in years of study and have often paid massive amounts of money for obtaining their law degrees. If they cannot practice law, their dreams of becoming a lawyer will be crushed, many will have no way to make a living sufficient to repay their educational debt, and the debt load may be a lifelong millstone. The only thing that keeps them from attaining their dream and avoiding this scenario is the bar admissions process.

So, a lot is riding on the process of admission to the bar, and what makes or breaks that process is the standard set for passage of the bar examination—the minimum score that determines passage. (There is, of course, also a character and fitness component that must be satisfied, in most cases before the bar examination.) Depending on the jurisdiction, the bar exam can take two to three days, beginning with essays and performance tests and, in some cases, jurisdiction-specific multiple-choice tests, and ending with the Multistate Bar Examination (MBE) on the last Wednesday of the month (February or July).[1] The jurisdiction generates a score from the written component answers and combines the score in some form with the scaled score generated from the MBE to produce the score used to determine the pass/fail result on the bar examination. On the MBE scale,

where scores can range from 0 to 200, passing scaled scores range from 129 to 145 across jurisdictions.

With the continuing decline in performance on the bar examination for the past few years, many jurisdictions have questioned whether their passing standards are set at the appropriate level. The purpose of this article is to provide information on several approaches that have been used to set standards on high-stakes licensing examinations and to highlight some of the challenges that exist in arriving at standards that ensure the protection of the public and are fair to law school graduates.

Standard-Setting Methods

The methods of setting standards all employ judgments on the part of a group of knowledgeable experts. The selection of these individuals is crucial to the credibility of the standards that result. The experts should be respected, and anyone questioning the credibility of the standard-setting process should, upon viewing the credentials of the experts, conclude that the experts are appropriate for the assignment. Depending upon the standard-setting approach used, the standard-setting panel may be required to make judgments about exam content, examinees, or task requirements. Those on the panel should also be free from any conflicts of interest, such as setting standards for students they have taught.

The approaches that have been used can be grouped into three types: arbitrary standard setting, test-centered methods, and examinee-centered methods.[2]

Arbitrary Standard Setting

The arbitrary method of standard setting is probably the most common approach and involves setting standards without systematically examining either the task requirements or samples of examinee performance. Susan M. Case, Ph.D., cites the BOGSAT approach (the acronym standing for Bunch of Guys Sitting at a Table).[3] The problem with arbitrary standards is that they are not easily defended, which might make them acceptable for low-stakes decisions like grades in a single course, but makes them ill-suited for making high-stakes decisions like bar passage. If no one currently involved with bar admissions in a jurisdiction has any idea of how that jurisdiction's standard was set, it should be reevaluated through a process other than the arbitrary method.

Test-Centered Methods

Psychometricians have been developing methods of setting defensible standards since at least the middle of the 20th century. The early methods were primarily test-centered, and it could be said that the operating principle undergirding these methods is that a standard would be more defensible than an arbitrary standard if the experts actually examined the test in terms of expectations of how the minimally competent examinee would perform.

Nedelsky Method

The earliest method in the psychometric literature is the Nedelsky method, named after its originator, Leo Nedelsky. It derived a standard, also known as a minimum pass level or MPL, for multiple-choice items by having the experts determine which of the incorrect answers the minimally competent examinee would be able to eliminate for any given multiple-choice item, assuming that the examinee would then guess among the remaining answers. For instance, if an item had four answer options and experts estimated that a minimally competent examinee could eliminate two of them, the MPL would be 1/2 (or 50%) since the examinee would be guessing between the correct answer and the remaining incorrect answer. The Nedelsky method was obviously limited to multiple-choice items.

Angoff Method

The Angoff method was subsequently developed by William Angoff with an eye toward broader applications. It or one of its variants has been applicable to almost any type of item, score, or task. This versatility is one of the factors that probably has motivated the Angoff method's use in many professional settings, such as medical licensure.[4] The Angoff approach has the experts first define characteristics of the minimally competent examinee. With a clear picture of what this minimally competent individual can do imprinted on their minds, the experts review the task at hand and give their estimate of the likelihood of this minimally competent individual being successful on the task. Because experts sometimes find it difficult to give an estimate of the likelihood of an individual being successful on a task, it is often reframed as how many of a group of 100 minimally competent examinees the expert would expect to be successful on this task.

M. Friedman Ben-David adapted the Angoff approach to use with a performance exam where examinees rotate between different stations and where, at each, they must demonstrate a specific skill or set of skills, and are often graded by a point system as they demonstrate aspects of the skill(s).[5] The adaptation of the Angoff method was made by having the experts determine the number of scoring points an individual borderline candidate would receive in order to pass the station. If an essay question (or performance test) is substituted for the skill(s) to be demonstrated at the station, the approach has direct applicability to the bar examination.

Examinee-Centered Methods

There are a number of methods that focus on the performance of examinees. The two most commonly encountered approaches are the Contrasting Groups method and the Bookmark method (although the Bookmark method has forms that are strictly test-centered). A third method with different application is the Hofstee method.

Contrasting Groups Method

There are variations of the Contrasting Groups method, but Michael T. Kane, Ph.D., describes this method in general as having experts determine if examinees have the knowledge, skills, and judgment needed to practice, based upon a sample of their performance; categorizing examinees into two groups (i.e., those who have met the requirements and those who have not); and then selecting a passing score that differentiates between the two groups as well as possible.[6]

Bookmark Method

The Bookmark method involves making judgments about either the tasks or the performance of examinees on a task, but it requires actual data. The easiest case for illustration purposes is for multiple-choice test performance. Items would be ordered from easiest to hardest (based upon actual data), and the experts would start with the easiest item and stop when they reach the point where they think a minimally competent examinee would have a specified probability of answering the item correctly (e.g., 50%). The difficulty of the item at this point would become the standard.

An alternative for performance-based assessments would be to order a sample of performances (e.g., essays) according to the grade awarded from lowest to highest. Experts would start at the lowest-graded performance and work their way up through the higher-graded performances until reaching the point where they find the first performance that a minimally competent examinee would have the specified probability of reaching. The grade of that performance would then be the passing standard.

In practice, the ordering of items/performances is not perfect, and one needs to have experts go up for a few more items/performances to be certain that some of the higher-graded ones were consistent with the stopping point and that where they stopped initially was not at an item/performance out of order. The challenge with the Bookmark method is to get the ordering of the items/performances before the standard-setting session, since it generally requires actual data to make the ordering.

Hofstee Method

The Hofstee method of standard setting does not make assessments of performance at the individual item level but requires experts to give their impressions of what the minimum and maximum failure rates should be for the exam, as well as what the minimum and maximum percent correct scores should be. These minimum and maximum failure rates and percent correct scores are averaged across experts and projected onto the actual score distribution to derive a passing score. Because it operates at the overall test level, it can be combined with other standard-setting methods as a cross-check. In fact, having experts go through the standard-setting process with, say, the Angoff method can be a good training approach for experts before they apply the Hofstee method.

Challenges

Generally, the methods used to derive standards seem relatively straightforward conceptually. However, the devil is in the details. Ronald K. Hambleton, Ph.D., provides the following 11 steps for setting performance standards on educational assessments, which can be applied to any of the standard-setting methods discussed.

1. Choose a panel (large, and representative of the stakeholders).
2. Choose one of the standard-setting methods, and prepare training materials and finalize the meeting agenda.
3. Prepare descriptions of the performance categories (e.g., basic, proficient, and advanced [or, in the case of the bar exam, fail and pass]).
4. Train panelists to use the method (including practice in providing ratings).
5. Compile item ratings and/or other rating data from the panelists (e.g., panelists specify expected performance of examinees at the borderlines of the performance categories).
6. Conduct a panel discussion; consider actual performance data (e.g., item difficulty values, item characteristic curves, item discrimination values, distractor analysis) and descriptive statistics of the panelists' ratings. Provide feedback on interpanelist and intrapanelist consistency.
7. Compile item ratings a second time that could be followed by more discussion, feedback, and so on.
8. Compile panelist ratings and obtain the performance standards.
9. Present consequences data to the panel (e.g., passing rate).
10. Revise, if necessary, and finalize the performance standards, and conduct a panelist evaluation of the process itself and their level of confidence in the resulting standards.
11. Compile validity evidence and technical documentation.[7]

Each of these 11 steps also seems relatively straightforward. However, as I said earlier, the devil is in the details. Taking the first step, "Choose a panel (large, and representative of the stakeholders)," here are some of the devilish details to think about. Who are the stakeholders? Are there some stakeholders who must actually be on the panel as opposed to

simply being represented? How many experts are needed? (There is large and then there is LARGE.) It has been recommended in the psychometric literature that 15 to 20 experts be used in setting the standard for a high-stakes examination like the bar examination. Choosing and recruiting experts is no small challenge. Representing a stakeholder group is not enough. What background and experiences does an expert need to be credible with not only the stakeholder group but the public at large? Serving on a standard-setting panel is a major time commitment. Are the experts willing to serve, and are they available to do so when needed? Setting standards is a very important task, and it requires careful thought at each step. Enlisting assistance from someone who has experience with the standard-setting process will help avoid major problems.

So, you might have noticed that the title of this article is really a double negative that translates into "Standards? We do need standards!" Whether they stink or not will depend upon how they are set.

References

Academy of Medical Royal Colleges, Guidance for Standard Setting: A Framework for High Stakes Postgraduate Competency-Based Examinations (October 2015), *available at* https://www.aomrc.org.uk/wp-content/uploads/2016/05/Standard_setting_framework_postgrad_exams_1015.pdf (accessed May 2, 2017).

John J. Bowers, Ph.D., and Russelyn Roby Shindoll, "A Comparison of the Angoff, Beuk, and Hofstee Methods for Setting a Passing Score," ACT Research Report Series No. 89-2, May 1989, *available at* https://forms.act.org/research/researchers/reports/pdf/ACT_RR89-2.pdf (accessed May 2, 2017).

S.M. Downing, A. Tekian, and R. Yudkowsky, "Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education," 18(1) *Teaching and Learning in Medicine* (Winter 2006) 50–57.

S.L. Fowell, R. Fewtrell, and P.J. McLaughlin, "Estimating the Minimum Number of Judges Required for Test-Centred Standard Setting on Written Assessments. Do Discussion and Iteration Have an Influence?" 13(1) *Advances in Health Sciences Education: Theory and Practice* (March 2008) 11–24.

W.K.B. Hofstee, "The Case for Compromise in Educational Selection and Grading," in *On Educational Testing* 109–127 (S.B. Anderson and J.S. Helmick eds., Jossey-Bass 1983).

National Board of Osteopathic Medical Examiners, Standard Setting, The Approach to Standard Setting, http://www.nbome.org/standardsetting_app.asp.

M.R. Raymond and J.B. Reid, "Who Made Thee a Judge? Selecting and Training Participants for Standards on Complex Performance Assessments," in *Setting Performance Standards: Concepts, Methods, and Perspectives* 119–157 (G.J. Cizek ed., Lawrence Erlbaum Associates 2001).

Notes

1. For the July 2017 and February 2018 bar examinations, Massachusetts will continue to administer its 10 essay questions on the Thursday following the administration of the Multistate Bar Examination. Effective with the July 2018 bar examination, Massachusetts will administer the Uniform Bar Examination, which ends with the MBE on Wednesday.
2. Michael T. Kane, Ph.D., "Standard Setting for Licensure Examinations," 70(4) *The Bar Examiner* (November 2001) 6–9.
3. Susan M. Case, Ph.D., "The Testing Column: Sometimes BOGSAT Is Just Not Good Enough," 81(3) *The Bar Examiner* (September 2012) 31–33.
4. R.J. Nungester, G.F. Dillon, D.B. Swanson, N.A. Orr, and R.D. Powell, "Standard-Setting Plans for the NBME Comprehensive Part I and Part II Examinations," 66(8) *Academic Medicine* (August 1991) 429–33.
5. M. Friedman Ben-David, "AMEE Guide No. 18: Standard Setting in Student Assessment," 22(2) *Medical Teacher* (2000) 120–130.
6. Kane, s*upra* note 2.
7. Ronald K. Hambleton, "Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process," Laboratory of Psychometric and Evaluative Research Report No. 377, School of Education, University of Massachusetts, Amherst, MA, *available at* http://www.nciea.org/publications/SetStandards_Hambleton99.pdf (accessed May 2, 2017).

Mark A. Albanese, Ph.D., is the Director of Testing and Research for the National Conference of Bar Examiners.

The Bar Examiner
The bar admissions information source

Standard Setting 101: Background and Basics for the Bar Admissions Community

Fall 2018 (Vol. 87, No. 3)

This article originally appeared in *The Bar Examiner* print edition, Fall 2018 (Vol. 87, No. 3), pp 9–17.

By Michael T. Kane, Ph.D., and Joanne Kane, Ph.D.

Licensure examinations such as the bar exam are high-stakes tests. A high-stakes test is defined by the *Standards for Educational and Psychological Testing* as "a test used to provide results that have important,

direct consequences for individuals, programs, or institutions involved in the testing."[1] In the definition provided in the *Standards*, the focus is on the exam taker and, as perhaps in our case, the exam administrator. However, we argue that the bar exam and many other professional licensure examinations should also be considered high-stakes from the perspective of the public; professional licenses are designed to protect the public from individuals seeking to practice who lack the requisite knowledge, skills, and abilities to adequately do so. Members of the public are counting on professional licenses to help ensure that a practitioner they would employ is at least minimally ready for practice.

The process of establishing a passing score is commonly referred to as *standard setting*. Standard setting in the licensure examination context is designed to address the basic policy question of how high an examinee's score must be for the examinee to pass the examination.

Establishing a passing score on such a high-stakes test is a critical component of ensuring the testing program's public protection function. There are costs to both individuals and the public associated with setting the bar too low or too high. If the bar is set too low, members of the public may be harmed through ineffective legal representation or actual malpractice. The public will have less confidence in members of the profession, and as a result, "consumer uncertainty" will increase. On the other hand, if the bar is set too high, would-be lawyers who would be able to competently represent clients will be inappropriately prevented from doing so. Individual examinees seeking to enter the profession, serve the public, and repay their student loans will suffer, and members of the public could be harmed by having their access to justice unduly limited via increased direct costs of representation or through increased caseloads for individual lawyers.

The process of establishing a passing score is commonly referred to as *standard setting*. Standard setting in the licensure examination context is designed to address the basic policy question of how high an examinee's score must be for the examinee to pass the examination. This standard represents the basic level of competence expected for entry-level practice. This article discusses the concept of standard

setting, how standards facilitate the process of making licensure decisions, and a few of the methods used to set standards in high-stakes contexts.

Why Do We Have Standards?

The adoption of a passing score for a licensure examination such as the bar exam changes what could be subjective decisions into objective or even mechanical ones and thereby promotes fairness and transparency. The process is highly efficient, reliable, and replicable. Determining whether an examinee passes or fails based on one clear criterion—his or her scaled score in relation to the passing score—is fast, unambiguous, and automatic.

In licensure decisions, the *decision rule*—in our case, the rule that is applied to make pass/fail determinations—typically specifies that if an examinee's scaled score is at or above the passing score, the examinee passes the test, and if the examinee's scaled score is below the passing score, the examinee fails the test.[2] This simple decision rule can easily be applied across examinees more or less automatically; no human judgment need come into play in the application of the rule. Of course, plenty of human judgment comes into play in the broader decision context in terms of identifying the requisite knowledge, skills, and abilities to be measured; designing the measurement instrument itself (i.e., the exam and its components); scoring the written components of the exam; and setting the passing score in the first place. But once the passing score is set and the scoring is completed, it is a simple matter to apply the decision rule.

To the extent that the standard of performance represented by the passing score is accepted, decisions based on that standard tend to be accepted by relevant stakeholders. The application of a clear standard has been described as a way of making a decision without appearing to decide.[3] It is hard to imagine a legitimate challenge, from the perspective of fairness, to the simple process of comparing a score to a passing score.

Are Standards Arbitrary?

Passing scores do not exist until some group develops them. Standards are *set* rather than "found" or estimated. The question, therefore, is not whether a passing score is "accurate" but rather whether the passing score, as set, achieves its purpose at an acceptable cost. Setting the passing standard is, in essence, a balancing act whereby policymakers weigh the benefits and costs of choosing a particular standard; the goal is to avoid setting the standard too high or too low.

Given a set of scores on a test, an increase in the passing score will generally decrease the pass rate, and a decrease in the passing score will generally increase the pass rate.[4] Even modest changes in the passing

score can yield substantial changes in pass rates, and these changes can vary substantially across groups (e.g., race/ethnicity, gender).

> Passing scores do not exist until some group develops them. Standards are *set* rather than "found" or estimated. The question, therefore, is not whether a passing score is "accurate" but rather whether the passing score, as set, achieves its purpose at an acceptable cost.

In 1978, a prominent researcher, Gene Glass, suggested that the results of educational standard setting tend to be arbitrary.[5] In response, a number of researchers acknowledged that standards are inherently judgmental but argued that they need not be arbitrary in the sense of being unjustified or capricious.[6] Further, context matters—the extent to which arbitrariness is a problem depends on how much it interferes with the intended use of the standard and, perhaps, the kind and degree of severity of unintended consequences.

Whenever a continuous variable is cut, the position of the cut can seem arbitrary. For example, the current maximum gross monthly income limit for SNAP (Supplemental Nutrition Assistance Program) recipients is $1,307.[7] Surely someone who earns $1,308, or $1,307.01, is effectively as food insecure as someone who earns $1,306.99. A penny or two—or even a dollar or two—one way or the other will not meaningfully influence a person's ability to provide for him- or herself and/or others. And yet, for a federal program serving millions[8] of Americans to be efficiently run and for benefits to be distributed in consistent ways, clear guidelines must exist. Which is to say, policies must be set.

Adopting a passing score for a licensure examination is, essentially, adopting a policy. Changes in policies can have dramatic effects. Changes to the SNAP income limits in either direction would affect millions of people and families.

On a "lighter" note, on June 17, 1998, the National Institutes of Health adopted new cut scores for the body mass index (BMI), a measure of percentage body fat based on a person's weight and height

measurements.[9] As a result, almost 30 million Americans were "suddenly" reclassified as clinically overweight, and several million were reclassified as clinically obese.

The inherent arbitrariness associated with standard setting needs to be controlled by providing support for the particular standard chosen. The 1998 changes to the BMI cut scores were developed judgmentally by a committee, but they were supported by clinical research, and the general locations of the cut scores were, therefore, far from unjustified or capricious.

To be considered acceptable (that is, to be defensible and comport with best practices in the measurement field), standards must meet certain criteria:

> They must be developed using generally accepted procedures based on relevant data.[10]
>
> They must be at an appropriate level or, at the very least, not at an obviously inappropriate level.
>
> They must be applied consistently over individuals and occasions.

In a paper titled "Justifying the Passing Scores for Licensure and Certification Tests," this author (Michael Kane), along with the two co-authors of the paper, proposed what they called the "Goldilocks Criteria" for evaluating passing scores and the standard-setting methods used to generate them:

The ideal performance standard is one that provides the public with substantial protection from incompetent practitioners and simultaneously is fair to the candidate and does not unduly restrict the supply of practitioners. We want the passing score to be neither too high nor too low, but at least approximately, just right.[11]

The standard should be high enough to provide assurance that new practitioners have certain competencies but not so high as to have serious negative consequences.

A Look at Standard Setting in Other High-Stakes Contexts

In thinking about standard setting for test scores, it can be useful to consider how standards are set in other high-stakes contexts. The organizations that develop pharmaceutical standards and other health-related standards generally rely on empirical research relating input variables to various outcomes. To develop these standards, they may use dosage-response curves, which represent the empirical relationship between an input (e.g., the dosage of a medication) and an outcome (e.g., the response in terms of pain reduction). A variety of key stakeholders, including patients, doctors, and health organizations, would agree that the dosage should be high enough to achieve the intended outcome (e.g., control of pain) but not so high as to cause unnecessary side effects or unintended consequences.

Dosage-response curves, like those shown in Figures 1 through 3, can be used to suggest or to check on the general location for a standard dosage. As illustrated in Figure 1, for low dosages, the response may be very limited, and the response may not increase much as the dosage increases, until it gets into a critical range where the effect increases fairly quickly as a function of the dosage. For higher dosages, the response often levels off, or plateaus. In order to achieve a high response, the dosage should be at or near the high end of the critical range. Going beyond the critical range does not add much to the expected response, and using higher dosages may lead to toxic side effects or could be costly (in terms of actual dollars and/or the ability to treat as many patients as possible) if the medication is expensive to produce or in short supply. For the dosage-response curve shown in Figure 1, a dosage of about 30 or a little higher (e.g., 31 or 32) would seem to be an optimal choice in terms of achieving the intended response without the unnecessary risks that might be associated with higher dosages.

Most dosage-response curves are not as sharp as the curve in Figure 1. For the dosage-response curve shown in Figure 2, 30 may again be a reasonable candidate for the standard dosage, but the range of acceptable values—that is, the dosage values yielding a reasonable response, without unnecessarily risking significant side effects—is much wider as compared to the clear-cut case shown in Figure 1, where no response is obtained at all until the dosage approaches 30 and where the response plateaus slightly above 30. In Figure 2, the range of acceptable values for the dosage extends from about 30 to about 40, or even further. For Figure 3, a dosage of 30 could be a reasonable choice for the standard, perhaps, but the range of plausible choices is much wider than in the curves shown in Figures 1 and 2.

Figure 1: Dosage-Response Curve—an “Easy” Case

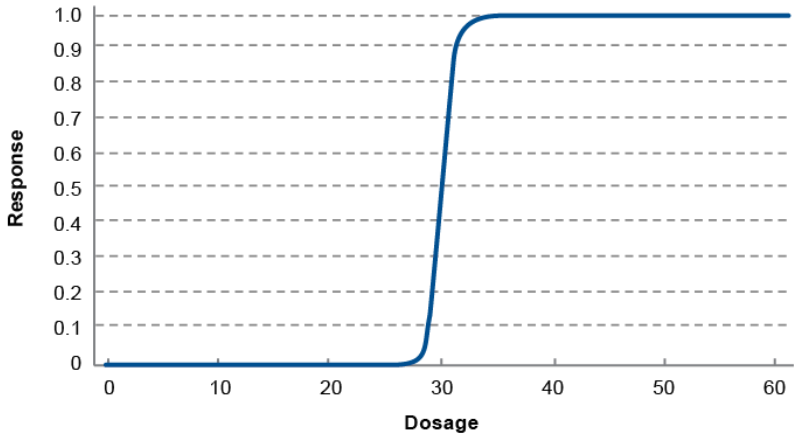


Figure 2: Dosage-Response Curve—an “Intermediate” Case

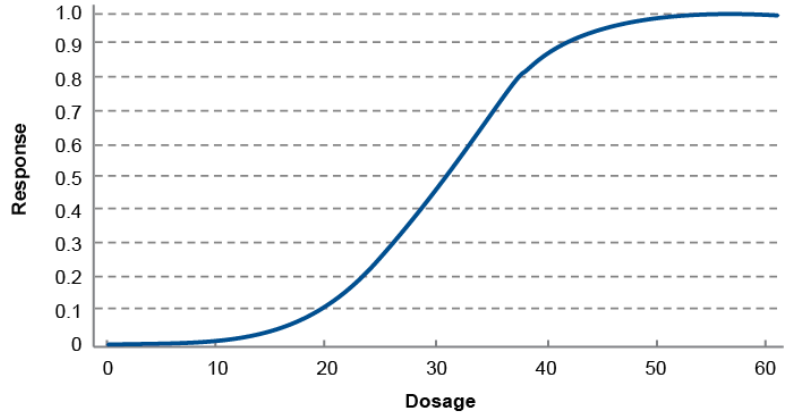
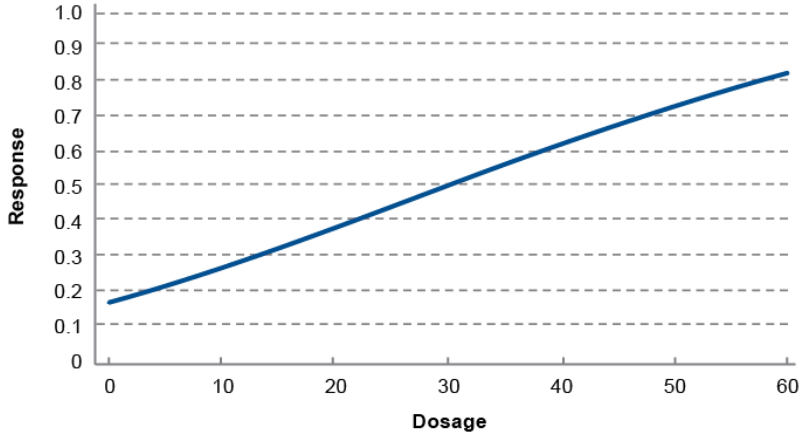


Figure 3: Dosage-Response Curve—a “Hard” Case



Dosage-response curves can be helpful in the standard-setting process, but they do not fully resolve the question of what the ideal standard would be. Note that even in the seemingly clearest of possible cases (Figure 1), it is not actually fully clear where the dosage should be set: Would the most appropriate dosage be 31 or 32, where the patient is getting most of the response? Or would a more appropriate dosage be closer to 35 or even 40, where it would be essentially certain that the patient will get the full response? The decision should depend on additional factors (e.g., potential side effects and direct and indirect costs) not reflected in the dosage-response curve.

Without additional information, the standard dosage can seem—and in fact can be—arbitrary. Figure 2 shows an intermediate case where the strength of response increases more gradually over a wider range of dosage levels, making it potentially more challenging to pinpoint an optimal dosage. And Figure 3 illustrates a case that barely hints at an optimal dosage. Additional considerations and constraints would be needed to determine the ideal dosage. The issue is one of balancing positive and negative consequences.

The use of dosage-response curves to set or evaluate standards involves the use of relevant empirical relationships to put bounds on the standard, followed by a judgment about where to put the standard within that range. The empirical results provide support for the general location of the standard (i.e., the critical range), but not for any single precise value within the range in most cases.

Unfortunately, although standards are often discussed as if there is an easy case akin to the one shown in Figure 1, high-stakes examinations usually present an ambiguous case more similar to the one shown in Figure 3. Thus, the standard-setting process will by necessity involve a substantial degree of human judgment. However, this human judgment need not be capricious; by involving individuals with relevant expertise, and by including a group of such individuals rather than relying on a single individual's opinion, a reasonable standard can be set. As with dosage decisions, the goal in setting a passing score in the licensing context should be to achieve the desired outcome without introducing serious negative consequences. This generally involves trade-offs.

Standard Setting for High-Stakes Examinations

In high-stakes examinations, standards are typically set using judgmental standard setting—that is, relying on the judgments of individuals with relevant expertise to determine the appropriate standard. As applied to testing, judgmental standard-setting procedures involve the use of a group of professionals (e.g., experienced practitioners, judges, and bar examiners) to recommend a passing score on some score scale to represent a certain level of performance: the *performance standard*. For licensure tests such as the bar exam, the performance standard is the basic level of competence expected of new practitioners. The goal

is to identify a passing score that reflects the performance standard and provides a reasonable basis for pass/fail decisions.

For licensure tests such as the bar exam, the performance standard is the basic level of competence expected of new practitioners. The goal is to identify a passing score that reflects the performance standard and provides a reasonable basis for pass/fail decisions.

A number of empirical methods have been developed for setting standards on tests.[12] Generally, the methods require panels of raters to conceptualize a minimally passing performance standard. The raters then use the performance standard to evaluate (i.e., rate) either examinee performances or test tasks (or both). That is, a group of experts could look at a sample of questions and say, "a minimally competent professional should be able to get at least half of these correct." Or a group of experts could make a judgment about a particular examinee—they could read an essay written by the examinee, for instance, and make a direct judgment about whether or not the examinee is minimally competent. There are also techniques for combining the two types of judgment—judgments about whether the questions are appropriately difficult and how many a minimally competent examinee should be able to answer, combined with judgments about the particular performances of individual examinees. The resulting data can be used to yield both a suggested passing score and/or a range of scores within which the suggested passing score would be considered reasonable.[13]

The results of a judgmental standard setting are not usually reported as curves (like Figure 2), but they could be presented and used in this way (based on the suggested passing score and the range of scores within which the suggested passing score might fall). The data available in judgmental standard setting are typically more limited than in the pharmaceutical case and are based on judgment rather than empirical clinical studies, but the data can be put into essentially the same mathematical form.

Judgmental standard-setting procedures may be evaluated according to several criteria:

One evaluative criterion might be procedural fairness or its cousin, methodological appropriateness: were the procedures used in the standard-setting exercise reasonable, thorough, and transparent?

Another criterion would be some sort of reliability measure or evaluation of internal consistency: are the data consistent across tasks, panels, and raters within panels?

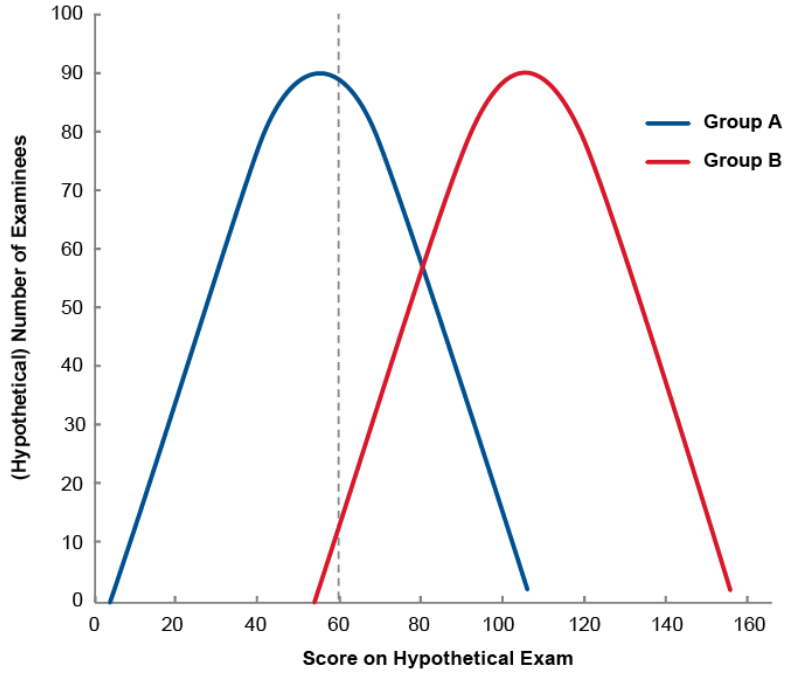
Finally, the procedure could be evaluated based on external criteria: are the results consistent with those of other studies using the same or different methods? Are the results consistent with those of historical trends and/or with a general sense of what would be reasonable? If not, is the rationale for the difference known and accepted?

Potential negative consequences can be particularly relevant in setting upper bounds for the passing score. For example, the location of a passing score can have a major impact on pass rates across demographic groups.[14] If there are two groups of test takers with different score distributions, and if the passing score is near the middle of the score distribution for the lower-scoring group (which is not uncommon) but in the lower tail of the distribution for the higher-scoring group, even a modest increase in the passing score can substantially increase the failure rate for the lower-scoring group while not having much impact on the higher-scoring group.

As a last example of the consequences and trade-offs associated with setting a particular standard, let's consider a hypothetical examination. Imagine that two groups have different score distributions on the test. Groups of interest often include race/ethnicity and gender but could include any groups significant in the social context. Further imagine that the set passing score is near the middle of the score distribution for one of the groups (i.e., about half of the examinees achieved a score at or above the set passing score) but in the lower tail of the score distribution for the other group (i.e., the majority of the examinees achieved a score at or above the passing score).

This basic hypothetical scenario is illustrated in Figure 4, where the set passing score of 60 is near the middle of the score distribution for Group A but in the lower tail of the score distribution for Group B. Moving the passing score down from 60 to 50 would have a larger effect on Group A—which has a very high number of examinees whose scores fell between 50 and 60 and who would now pass the exam—than on Group B—which has very few examinees whose scores fell between 50 and 60. In this example, the impact of the change in passing score would not be equal across groups. There are direct and indirect consequences of any given passing score for an array of stakeholder groups, including the public, and for individuals.

Figure 4: Hypothetical Examination Example



Concluding Remarks

There is no generally agreed-upon single best method for conducting a standard-setting study for a high-stakes licensure examination. That said, it can be useful to explore what other licensure organizations have done in setting their standards, what other jurisdictions have done, and what the *Standards for Educational and Psychological Testing* recommend. In addition, the methods typically used in health-care standard setting (e.g., dosage-response curves) may provide a useful model for talking about standard setting in general. The health-care approach makes extensive use of empirical research and also tries to strike a balance between competing goals.

As mentioned at the beginning of this article, standard setting in the licensure examination context is designed to address the basic policy question of how high an examinee's score must be for the examinee to pass the examination. Although empirical data should play a central role in standard setting, ultimately standards are *set*, not "found" or estimated. Thus, standards are not evaluated in terms of their accuracy per se but rather in terms of whether they support the goals of the program without introducing unacceptable and unintended consequences. In the context of the bar examination, the passing score should be high enough to protect the public, but not so high as to be unduly limiting to those seeking to enter the profession.

Editor's Note: This article is partially based on Dr. Michael T. Kane's presentation, "Standard Setting for Licensure Examinations," at the 2018 NCBE Annual Bar Admissions Conference held on April 19–22, 2018, in Philadelphia, Pennsylvania.

Notes

1. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* 214 (American Educational Research Association 2014).*(Go back)*
2. M.T. Kane, B.E. Clauser & J. Kane, "A Validation Framework for Credentialing Tests," in *Testing in the Professions: Credentialing Policies and Practice* (Eds. S. Davis-Becker and C.W. Buckendahl, National Council on Measurement and Education 2017).*(Go back)*
3. T. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* 8 (Princeton University Press 1995).*(Go back)*
4. This might seem like an obvious point, but we note that the impact in terms of pass rate will depend on both what the underlying distributions of scores look like and where the passing score falls within the distribution.*(Go back)*
5. G.V. Glass, "Standards and Criteria," 15(4) *Journal of Educational Measurement* (Winter 1978) 237–261.*(Go back)*
6. R. Hambleton & M. Pitoniak, "Setting Performance Standards," in *Educational Measurement* 433–470 (Ed. R.L. Brennan, American Council on Education and Praeger Publishers 4th ed. 2006).*(Go back)*
7. United States Department of Agriculture Food and Nutrition Service, Supplemental Nutrition Assistance Program (SNAP), https://www.fns.usda.gov/snap/eligibility#What%20are%20the%20SNAP%20income%20limits? (last visited Sep. 14, 2018).*(Go back)*
8. United States Department of Agriculture Food and Nutrition Service, Supplemental Nutrition Assistance Program (SNAP), https://www.fns.usda.gov/pd/supplemental-nutrition-assistance-program-snap (last visited Sep. 14, 2018). *(Go back)*
9. L. Shapiro, "Fat, Fatter: But Who's Counting?," *Newsweek*, June 15, 1998, at 55.*(Go back)*
10. For a recent review of some such standard-setting procedures, *see* M. Albanese, "The Testing Column: Standards? We Don't Need No Stinking Standards!," 86(2) *The Bar Examiner* (June 2017) 36–40.*(Go back)*
11. M. Kane, T. Crooks & A. Cohen, "Justifying the Passing Scores for Licensure and Certification Tests." Paper presented at the annual meeting of the American Educational Research Association, Chicago, March 1997, p. 8.*(Go back)*
12. *See* Albanese, *supra* note 10, which goes into some detail on a few of these empirical methods that have been developed for setting standards on tests.*(Go back)*
13. For additional detail on standard-setting methods, *see* B.E. Clauser, M.J. Margolis & S. Case, "Testing for Licensure and Certification in the Professions," in *Educational Measurement* 701–731 (Ed. R.L. Brennan, American Council on Education and Praeger Publishers 4th ed. 2006); Hambleton & Pitoniak, *supra* note 6; M. Kane, "Standard Setting for Licensure Examinations," 70(4) *The Bar Examiner* (November 2001) 6–9; M. Kane, "Practice-Based Standard Setting," 71(3) *The Bar Examiner* (August 2002) 14–24; M. Kane, "Conducting Examinee-Centered Standard-Setting Studies Based on Standards of Practice," 71(4) *The Bar Examiner* (November 2002) 6–13; or M.J. Zieky, M. Perie & S.A. Livingston, *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* (Educational Testing Service 2008).*(Go back)*
14. A. Mroch, M. Kane, D. Ripkey & S. Case, "Impact of the Increase in the Passing Score on the New York Bar Examination: February 2006 Bar Examination." Report prepared for the New York Board of Law Examiners by the National Conference of Bar Examiners, June 19, 2007.*(Go back)*

Michael T. Kane, Ph.D., is the holder of the Samuel J. Messick Chair in Test Validity at Educational Testing Service (ETS). He was Director of Research for the National Conference of Bar Examiners from 2001 to 2009. From 1991 to 2001, he was a professor in the School of Education at the University of Wisconsin–Madison, where he taught measurement theory and practice. Prior to that, Kane served as vice president for research and development and as a senior research scientist at American College Testing (ACT), where he supervised large-scale validity studies of licensure examinations. Kane holds an M.S. in statistics and a Ph.D. in education from Stanford University.

Joanne Kane, Ph.D., is the Associate Director of Testing for the National Conference of Bar Examiners.

Contact us to request a pdf file of the original article as it appeared in the print edition.

Supreme Court of California

JORGE E. NAVARRETE
CLERK AND EXECUTIVE OFFICER
OF THE SUPREME COURT

EARL WARREN BUILDING
350 McALLISTER STREET
SAN FRANCISCO, CA 94102
(415) 865-7000

July 16, 2020

SENT VIA USPS AND EMAIL

Alan K. Steinbrecher, Chair
State Bar of California, Board of Trustees
180 Howard Street
San Francisco, CA 94105
asteinbrecher@steinbrecherspan.com

RE: California Bar Exam

Dear Mr. Steinbrecher,

The changing circumstances surrounding the ongoing COVID-19 pandemic in California, and throughout the country, have had an unprecedented impact on professional licensure testing for graduates seeking admission to many professions, including not only law, but medicine, nursing, architecture, and engineering. The court understands that many law school graduates are being substantially affected by the resulting disruption. Some graduates have lost job offers. Many are about to lose health insurance, cannot find a job to pay bills, or are in fear of deportation if they cannot enter the bar in time to retain job offers. Many more have student loan payments that become due in mid-November, but without a law license and the ability to work, they fear going into default.

With these considerations in mind, the court has sought the safest, most humane and practical options for licensing law graduates by encouraging and working with the State Bar to pursue the option of administering the California Bar Examination online as a remote test, to avoid the need for, and dangers posed by, mass in-person testing. The court also directed the State Bar to engage in focused conversations with the National Conference of Bar Examiners (NCBE) to address the ability to administer an online version of the multiple-choice Multistate Bar Examination.

Our sister states also struggle with similar issues. Many have recently canceled in-person testing plans and have increasingly turned to online solutions. Although a few less populous states have been able to accommodate a diploma privilege that grants entry for all of the graduates of their states' constituent American Bar Association (ABA)-accredited law schools, the law schools in California, unlike in other states, represent a diverse array of ABA-accredited, California-accredited, and California-registered schools. If California were to adopt diploma-privilege criteria used by other states, graduates of nearly four dozen California law schools would not meet those criteria and would be excluded.

With these considerations in mind, the court seeks a path that ensures the fair and equal treatment of all graduates, regardless of law school accreditation status, while also ensuring that protections remain in place for consumers of legal services.

After considering all letters, comments, the actions of other states, discussions with the NCBE, consultations with the informal state bar workgroup on the status of the bar exam, and having given careful thought to the expressed needs of bar applicants, the court directs the State Bar as follows:

The September 9-10 administration is cancelled. Joining at least 15 other jurisdictions that have, to date, taken similar measures, the State Bar is directed to make the necessary arrangements for the online remote administration of the bar examination on October 5-6, 2020, and extend registration for this exam through July 24, 2020. The State Bar has worked diligently on measures for the successful deployment of the exam online. Based on that work and current information, the court has determined that an online exam can be administered and delivered without the need for an examinee to have a high-speed or constant internet connection. The court asks that the State Bar clearly explain the necessary system requirements and other details concerning the circumstances of an online exam in a "Frequently Asked Questions" guide.

The court strongly encourages law schools to assist those graduates who lack internet access at home, or who have home environments not amenable to two days of uninterrupted examination, by employing the same and similar measures, including the use of school facilities and equipment, that schools have utilized to allow students to complete the Spring 2020 semester.

In consideration of the fact that California is one of two states with the highest pass score for its minimum competency exam, and based on findings from recently completed bar examination studies as well as data from ongoing studies, the court directs the State Bar to modify the pass score for the California Bar Examination to allow for a minimum passing score of 1390, which is approximately two standard errors below the median recommended cut score of 1439 from the 2017 Standard Setting Study. This modified minimum passing score is effective for the administration of the bar examination on October 5-6, 2020, and will be applied prospectively to future administrations of the California Bar Examination (irrespective of whether the exam is administered online in the future). The court will consider any further changes pending recommendations offered by the forthcoming Blue-Ribbon Commission on the Future of the California Bar Examination.

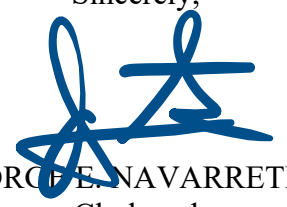
The court recognizes that postponement of the bar examination may impact employment prospects, delay incomes, and otherwise impair the livelihoods of persons who recently have graduated from law school. Moreover, the court recognizes 2020 graduates may not be in a position to study and prepare for a fall bar 2020 examination. Therefore, in order to mitigate these hardships faced by graduates while fulfilling the responsibility to protect the public by ensuring that persons engaged in the practice of law are minimally competent to do so, the court directs the State Bar to implement, as soon as possible, a temporary supervised provisional licensure program — a limited license to practice specified areas of law under the supervision of a licensed attorney.

This program will be made available for all 2020 graduates of law schools based in California or those 2020 graduates of law schools outside California who are permitted to sit for the California Bar Examination under Business and Professions Code sections 6060 and 6061. More information will be forthcoming regarding this program, and the State Bar will issue a

"Frequently Asked Questions" guide concerning the details. At a minimum, this provisional licensure program shall remain in effect until at least June 1, 2022 to permit 2020 graduates maximum flexibility. This timeframe will afford the 2020 graduates several opportunities to take the exam of their choosing through February 2022 and await the exam results. In addition, in order to expedite relief and pursuant to the court's inherent authority over the admission of attorneys into the practice of law, the State Bar should afford a public comment period of at least 15 days for any proposed supervised provisional licensure program rules. (*In re Attorney Discipline System* (1998) 19 Cal.4th 582; Cal. Rules of Court, Rule 9.3.)

With the exception of postponing the October 2020 First-Year Law Students' Examination to November 2020 or any amendments to the rules governing the number of times an examinee can sit for that exam, this letter supersedes the court's prior April 27, 2020 letter.

Sincerely,



JORGE E. NAVARRETE
Clerk and
Executive Officer of the Supreme Court

cc: Donna Hershkowitz

The State Bar *of California*

Simulation of the Impact of Different Bar Exam Cut Scores on Bar Passage, by Gender, Race/Ethnicity, and Law School Type

Office of Research and Institutional Accountability

March 18, 2020

The State Bar of California compiled historical California Bar Exam data and conducted a simulation analysis following discussions about the potential impact of different cut scores on test takers' pass rates by gender, race/ethnicity, and law school type. The tables presented in this report show the number of exam takers who would have passed the bar exam if the cut score had been 1300, 1330, 1350, and 1390.[1] The tables also show how the pass rate changes and the difference in the number of exams taken as the simulated cut score changes. The simulations presented in this report should not be construed to imply any position of the State Bar regarding the propriety of the current cut score, or any of the hypothetical cut scores evaluated. The issue was previously addressed in the standard setting study conducted in 2017.[2]

DATA AND METHODS

The simulations are based on archival data on results from 21 bar exams administered over a span of 11 years, from February 2009 to February 2019. The data allows tracking of the bar exam results for more than 85,000 examinees, who collectively took more than 140,000 exams. Table 1 shows summary statistics on the total number of exams under consideration by gender, race/ethnicity, and law school type.

[1] The specific hypothetical cut scores included were selected by a law school dean who spearheaded the production of this simulation.

[2] See Final Report on the 2017 California Bar Exam Studies: https://www.calbar.ca.gov/Portals/0/documents/reports/2017-Final-Bar-Exam-Report.pdf

Table 1. Number of Bar Examinees and Exams Taken from February 2009 to February 2019, by Race/Ethnicity, Gender, and Law School Type

	Unique Examinees	Exams Taken
Total	85,727	143,198
Gender		
Male	43,787	73,289
Female	41,386	69,082
No Response	554	827
Race Ethnicity		
Asian	18,510	32,728
Latino	9,166	17,944
African American	4,417	9,841
White	48,917	75,633
Other	1,384	2,558
No Response	3,333	4,494
Law School Type		
CA ABA Approved	42,922	63,912
Out-of-State ABA	15,587	24,732
CA Accredited	5,173	14,714
CA Unaccredited	2,221	6,447
US Attorneys	13,849	19,739
Other	5,975	13,654

To illustrate how the simulation results are calculated, Table 2 presents the experience of a hypothetical examinee. The examinee took the exam four times over three years, achieving a range of scores from 1289 at the lowest to 1395 at the highest. With the current cut score of 1440, the examinee did not pass the exam and stopped trying after the fourth attempt. Under all four hypothetical cut scores, however, they were able to pass the exam, although they would not have passed until the fourth attempt if the cut score were 1390. When they passed the exam under a hypothetical cut score, *subsequent attempts in the data are removed from the calculation* for that particular scenario. Thus, under the scenarios of 1300 and 1330, the number of exams taken for this person would be calculated as two; under the scenario of 1350, the number of exams taken would be counted as three.

Table 2. An Example of Simulated Exam Outcomes for a Repeat Examinee

		Hypothetical Cut Scores				*Current Cut Score*
Exams Taken	Total Score	1300	1330	1350	1390	1440
February 2009	1289	F	F	F	F	F
July 2009	1335	P	P	F	F	F
July 2010	1360	-	-	P	F	F
February 2011	1395	-	-	-	P	F

Note: P = pass, F = fail

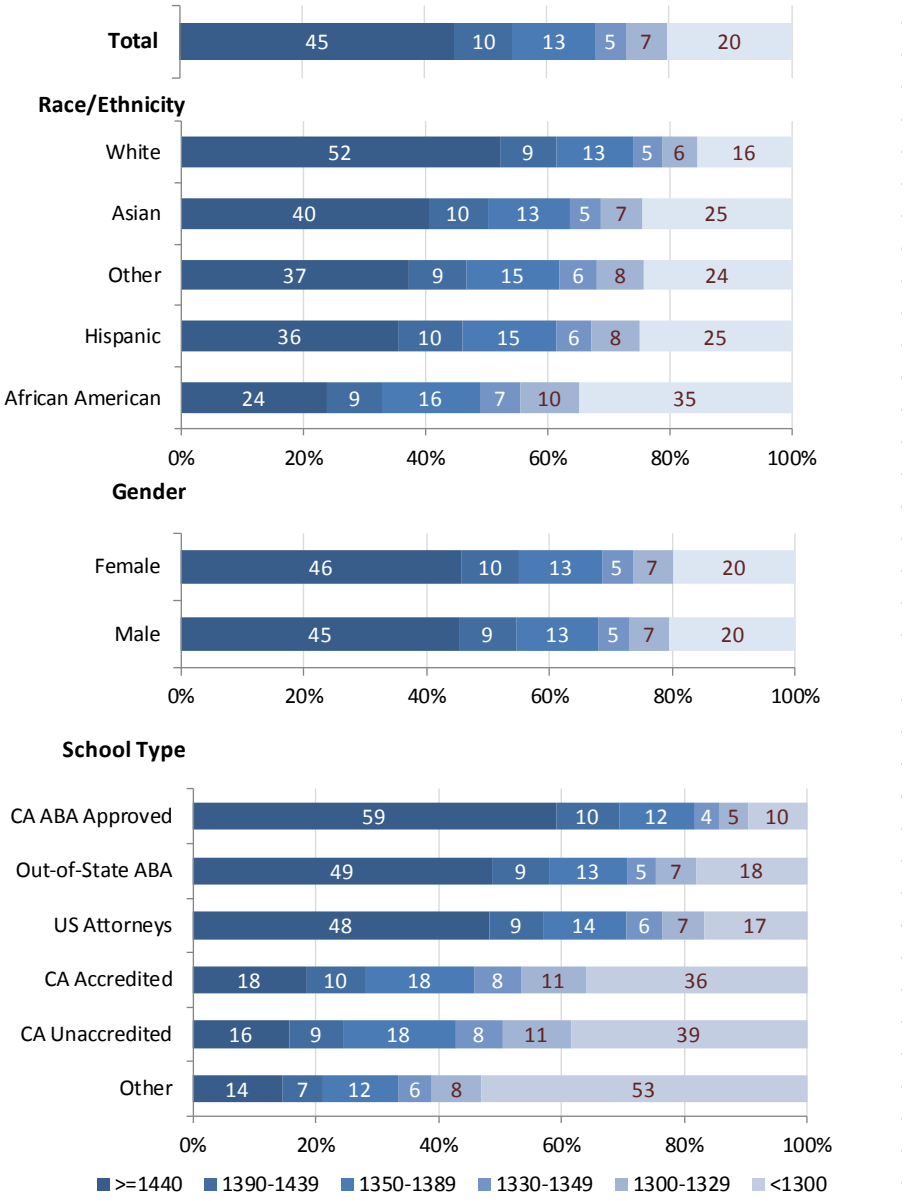
SIMULATION RESULTS

The simulation exercise produces three sets of results under each hypothetical cut score: (1) impact on the number of examinees passing the exam; (2) impact on the pass rate; and (3) impact on the number of exams taken in aggregate.

The impacts of the several cut score hypotheticals are different for various subgroups, reflecting the variation of actual bar exam performance for each subgroup. In general, a group with a higher bar pass rate under the current cut score of 1440 would see a smaller impact from a lower cut score, compared to a group with a lower bar pass rate.

As background information to assist in interpreting the simulations, Figure 1 shows the distribution of bar exam scores at different ranges for all exams included in this exercise. Note that, under the current cut score of 1440, 59 percent of examinees graduating from a California ABA law school passed the bar exam, compared to 18 percent of examinees who graduated from a California-accredited law school. In simulating the impact from lowering the cut score to 1390, Table 3 shows that examinees graduating from ABA law schools would see a 4 percent increase in the number of examinees passing the bar exam, compared to a 14 percent increase for examinees graduating from California-accredited law schools. The difference reflects the larger base of ABA law school graduates passing the bar exam under the current cut score compared to California-accredited law schools. More subgroup comparisons of the simulation results can be viewed in a separate Excel file.

Figure 1. Distribution of Bar Exam Performance: All Exams from February 2009 to February 2019, by Race/Ethnicity, Gender, and Law School Type



Impact on the Number of Examinees Passing the Exam

Table 3 contains the results of the simulation's impact on the number of examinees passing the bar exam. The first column, labeled "Current at 1440," shows the number of examinees covered in this analysis who actually passed the bar exam. The next four columns calculate the number of examinees who would have passed the exam under each of the hypothetical lower cut scores. The additional examinees passing the exam are shown in the following four columns. Increases in examinees passing the exam, in percentage terms relative to the actual passers, are

shown in the last four columns. At 1390, for example, the table shows that 3,741 more examinees would have passed the exam, a total gain of 5.8 percent.

Impact on Pass Rate

Table 4 shows the pass rate and how it changes under the different scenarios. As with Table 3, the column "Current at 1440" shows actual data to use as the point of comparison. The 45 percent shown in the first row of the first column of Table 4 represents the 65,025 examinees who passed the bar exam out of all exams taken (143,198, shown in Table 5). At the hypothetical cut score of 1390, the pass rate would have been 53 percent, which represents an increase of 8 percentage points, shown in the last column.

Impact on the Number of Exams Taken

The first column in Table 5 shows the actual number of exams taken over the period covered in this analysis. As shown in Table 2, the number of exams taken would decrease as examinees passed the exam with fewer attempts. At a hypothetical cut score of 1390, the table shows that a total of 128,702 exams would have been taken. Compared to the actual count of 143,198, it represents a decrease of nearly 15,000 exams taken, a reduction of approximately 10 percent, as shown in the last column.

Data for Intersectional Analyses

The simulation results presented in Tables 3 to 5 are limited by the three subgroup categories in which comparisons cannot be made across subgroup intersections, such as between white and African American females from ABA or California-accredited law schools. An Excel file is provided with this report to make possible these types of dynamic, multidimensional comparisons.

Table 3. Simulation of Bar Exam Outcomes: Impact on Number of Examinees Passing the Exam

		Simulated Number of Examinees Passing Exam				Number of Additional Examinees Passing Exam				Percent Increase in Examinees Passing Exam			
	Current at 1440	1300	1330	1350	1390	1300	1330	1350	1390	1300	1330	1350	1390
Total	65,025	77,913	75,570	73,740	68,766	12,888	10,545	8,715	3,741	20%	16%	13%	6%
Race/Ethnicity													
Asian	13,229	16,159	15,603	15,190	14,109	2,930	2,374	1,961	880	22%	18%	15%	7%
African American	2,345	3,499	3,257	3,098	2,639	1,154	912	753	294	49%	39%	32%	13%
Latino	6,377	8,086	7,756	7,539	6,878	1,709	1,379	1,162	501	27%	22%	18%	8%
White	39,400	45,839	44,734	43,786	41,283	6,439	5,334	4,386	1,883	16%	14%	11%	5%
Other	954	1,219	1,178	1,132	1,027	265	224	178	73	28%	23%	19%	8%
No Response	2,720	3,111	3,042	2,995	2,830	391	322	275	110	14%	12%	10%	4%
Gender													
Male	33,214	39,797	38,635	37,675	35,134	6,583	5,421	4,461	1,920	20%	16%	13%	6%
Female	31,464	37,666	36,502	35,648	33,263	6,202	5,038	4,184	1,799	20%	16%	13%	6%
No Response	347	450	433	417	369	103	86	70	22	30%	25%	20%	6%
Law School Type													
CA ABA Approved	37,729	41,680	41,105	40,630	39,097	3,951	3,376	2,901	1,368	10%	9%	8%	4%
Out-of-State ABA	12,033	14,419	13,970	13,652	12,735	2,386	1,937	1,619	702	20%	16%	13%	6%
CA Accredited	2,705	4,157	3,869	3,628	3,082	1,452	1,164	923	377	54%	43%	34%	14%
CA Unaccredited	1,008	1,711	1,561	1,467	1,191	703	553	459	183	70%	55%	46%	18%
US Attorneys	9,573	12,498	11,954	11,490	10,334	2,925	2,381	1,917	761	31%	25%	20%	8%
Other	1,977	3,448	3,111	2,873	2,327	1,471	1,134	896	350	74%	57%	45%	18%

Table 4. Simulation of Bar Exam Outcomes: Impact on Pass Rate

		Simulated Exam Pass Rate				Percentage Point Increase in Pass Rate			
	Current at 1440	1300	1330	1350	1390	1300	1330	1350	1390
Total	45%	77%	70%	66%	53%	31	25	20	8
Race/Ethnicity									
Asian	40%	70%	64%	59%	48%	30	24	19	8
African American	24%	56%	47%	42%	30%	33	24	18	6
Latino	36%	69%	61%	56%	43%	34	26	21	8
White	52%	83%	77%	73%	60%	31	25	21	8
Other	37%	71%	64%	58%	45%	34	26	20	7
No Response	61%	86%	81%	78%	68%	25	20	17	7
Gender									
Male	45%	77%	70%	65%	53%	31	25	20	8
Female	46%	77%	70%	66%	54%	31	25	20	8
No Response	42%	67%	63%	59%	47%	25	21	17	5
Total	45%	77%	70%	66%	53%	31	25	20	8
Law School Type									
CA ABA Approved	59%	90%	85%	81%	69%	31	26	22	10
Out-of-State ABA	49%	80%	73%	68%	57%	31	24	20	8
CA Accredited	18%	51%	42%	36%	24%	32	23	17	6
CA Unaccredited	16%	47%	38%	33%	21%	32	22	17	6
US Attorneys	48%	82%	76%	70%	56%	34	27	21	8
Other	14%	34%	29%	25%	18%	20	14	11	4

Table 5. Simulation of Bar Exam Outcomes: Impact on Exams Taken

		Simulated Number of Exam Takers				Reduction in the Number of Exams Taken				Percent Decrease of Exams Taken			
	Current at 1440	1300	1330	1350	1390	1300	1330	1350	1390	1300	1330	1350	1390
Total	143,198	101,374	107,414	112,429	128,702	41,824	35,784	30,769	14,496	29%	25%	21%	10%
Race/Ethnicity													
Asian	32,728	22,974	24,377	25,572	29,281	9,754	8,351	7,156	3,447	30%	26%	22%	11%
African American	9,841	6,209	6,859	7,356	8,781	3,632	2,982	2,485	1,060	37%	30%	25%	11%
Latino	17,944	11,705	12,657	13,403	15,874	6,239	5,287	4,541	2,070	35%	29%	25%	12%
White	75,633	55,133	57,915	60,293	68,273	20,500	17,718	15,340	7,360	27%	23%	20%	10%
Other	2,558	1,721	1,849	1,963	2,302	837	709	595	256	33%	28%	23%	10%
No Response	4,494	3,632	3,757	3,842	4,191	862	737	652	303	19%	16%	15%	7%
Gender													
Male	73,289	51,810	54,915	57,535	65,866	21,479	18,374	15,754	7,423	29%	25%	21%	10%
Female	69,082	48,891	51,810	54,186	62,050	20,191	17,272	14,896	7,032	29%	25%	22%	10%
No Response	827	673	689	708	786	154	138	119	41	19%	17%	14%	5%
Law School Type													
CA ABA Approved	63,912	46,237	48,272	50,099	56,796	17,675	15,640	13,813	7,116	28%	24%	22%	11%
Out-of-State ABA	24,732	18,074	19,154	19,946	22,463	6,658	5,578	4,786	2,269	27%	23%	19%	9%
CA Accredited	14,714	8,197	9,300	10,143	12,792	6,517	5,414	4,571	1,922	44%	37%	31%	13%
CA Unaccredited	6,447	3,620	4,097	4,436	5,599	2,827	2,350	2,011	848	44%	36%	31%	13%
US Attorneys	19,739	15,185	15,833	16,494	18,338	4,554	3,906	3,245	1,401	23%	20%	16%	7%
Other	13,654	10,061	10,758	11,311	12,714	3,593	2,896	2,343	940	26%	21%	17%	7%

The Bar Examiner
The bar admissions information source

The Testing Column: Did UBE Adoption in New York Have an Impact on Bar Exam Performance?

Winter 2019-2020 (Vol. 88, No. 4)

This article originally appeared in *The Bar Examiner* print edition, Winter 2019-2020 (Vol. 88, No. 4), pp 34–42.

Highlights of a Study Directed by the New York State Court of Appeals

By Andrew A. Mroch, PhD, and Mark A. Albanese, PhD

In adopting the Uniform Bar Examination (UBE), the New York State Court of Appeals directed the New York State Board of Law Examiners (NYSBLE) to study the impact of the change to the UBE on candidate bar exam performance. The NYSBLE requested assistance from NCBE in conducting the study, which NCBE provided as part of its service mission as a not-for-profit corporation.

The study covered the two bar exam administrations immediately before UBE adoption (July 2015 and February 2016) and continued through the July 2017 administration, resulting in one February administration post-UBE adoption (February 2017) and two July administrations post-UBE adoption (July 2016 and July 2017). In addition to the overall impact of UBE adoption, the study addressed potential differential effects by gender and race/ethnicity.

The New York State Court of Appeals released the study results in a publicly available report on August 20, 2019, providing a rich trove of information on the background characteristics and performance of candidates taking the bar exam in New York between July 2015 and July 2017.[1]

This article briefly highlights several findings of the report relative to what happened before and after UBE adoption in New York regarding three topics:

1. bar examination performance
2. candidate background characteristics: pre-law-school undergraduate grade point average (UGPA), Law School Admission Test (LSAT) score, and law school grade point average (LGPA)
3. the relationships between candidate bar exam performance and their background characteristics (UGPA, LSAT score, and LGPA)[2]

What Data Were Used for the Study?

Two samples of New York bar exam data were analyzed.

Domestic-educated NYSBLE sample: The first sample, referred to as the *domestic-educated NYSBLE sample*, included candidates who had received a JD degree from an American Bar Association–approved law school in the United States.[3]

School-based sample: The second sample, referred to as the *school-based sample*, was a subset of the domestic-educated NYSBLE sample and included candidates for whom law schools throughout the United States provided their UGPAs, LSAT scores, and LGPAs. To facilitate a meaningful analysis, only those candidates whose law schools provided such data for at least 25 candidates were included in the school-based sample.

The LGPAs provided by law schools used various systems. The most common was the 4-point system corresponding to A = 4, B = 3, and so on, but some schools used a 100-point system and an assortment of other approaches. In order to appropriately analyze LGPAs from different schools, LGPAs were scaled in two ways to ensure comparability: (1) to range from 1 to 4 (resulting in the *4-point LGPA*) and (2) to account for school-level differences in selectivity[4] (resulting in the *index-based LGPA*). All analyses that included LGPAs were conducted separately using each method of scaling LGPAs.

Table 1 shows the numbers and percentages of candidates included in the two samples at each bar exam administration. Compared to the total number of domestic-educated candidates in the NYSBLE sample, the percentages of candidates represented by the school-based sample were relatively low for the

February exams (22.8% and 30.5%) and for July 2015 (27.7%) compared to July 2016 (62.0%) and July 2017 (55.4%).

In addition to having a relatively small percentage of candidates represented in the school-based sample, the February results were sufficiently unstable that they were excluded from this summary. For July, there were differences in the percentage of candidates represented across years, but the numbers were sufficiently large that the school-based sample was still useful for studying candidates across July exams.

Bar examination scores also required adjustments in order to enable appropriate comparisons. Scores on the prior New York bar exam were on a 1,000-point scale but were converted in this study to the 400-point UBE scale to facilitate comparisons across exams. (See sidebar)

Table 1. Numbers and percentages of candidates in the New York UBE study samples

New York UBE study sample	February 2016 administration	February 2017 administration	July 2015 administration	July 2016 administration	July 2017 administratior
Domestic-educated NYSBLE sample–%	100.0%	100.0%	100.0%	100.0%	100.0%
Domestic-educated NYSBLE sample–(n)	(2,346)	(2,370)	**(7,513)**	**(7,292)**	**(6,776)**
School-based sample–%	22.8%	30.5%	27.7%	62.0%	55.4%
School-based sample–(n)	(534)	(723)	**(2,084)**	**(4,520)**	**(3,753)**

The data used for the results presented in this article are indicated in bold.

The New York Bar Exam, Pre- and Post-UBE

The UBE consists of

> the Multistate Bar Examination (MBE), weighted 50% of the total score, and
>
> a written component consisting of six Multistate Essay Examination (MEE) questions, weighted 30% of the total score, and two Multistate Performance Test (MPT) questions, weighted 20% of the total score.

Scores on the UBE are on a 400-point scale. A passing score in New York on the 400-point UBE scale is a score of at least 266.

The New York bar exam prior to UBE adoption consisted of

> the MBE, weighted 40% of the total score,
>
> a written component consisting of five New York–developed essay questions, weighted 40% of the total score, and one MPT question, weighted 10% of the total score, and
>
> 50 New York–developed multiple-choice questions, weighted 10% of the total score.

Prior to adoption of the UBE, the passing score was 665 on a 1,000-point scale. This passing score corresponds to a 266 on the 400-point UBE scale.

What Were the Results of the Study?

Bar Exam Performance

Figures 1 and 2 show bar exam performance and pass rates across the period of the study.

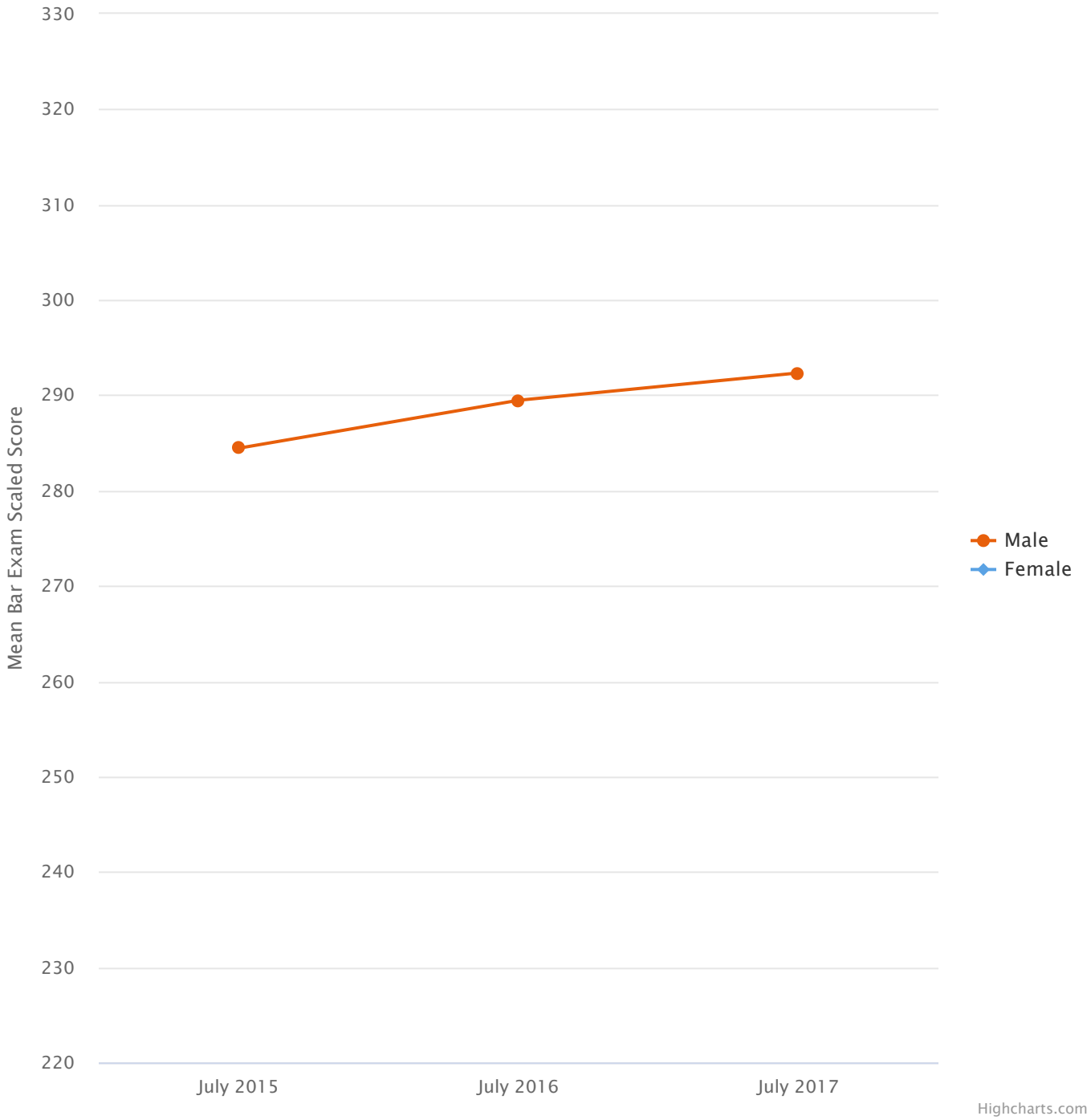
> Between July 2015 and July 2017, before and after UBE adoption in July 2016, bar exam performance and pass rates in New York increased, on average. For example, the pass rate for domestic-educated candidates in the NYSBLE sample was 72.5% in July 2015, 75.1% in July 2016, and 78.0% in July 2017 (Figure 2a).[5]
>
> Males tended to score slightly higher than females, on average, across Julys, with the difference between males and females in the domestic-educated NYSBLE sample widening slightly in July 2016 upon UBE adoption before narrowing in July 2017 (Figure 1a).
>
> Similar patterns of bar exam performance and pass rates were observed for the school-based sample (Figures 1b and 2b).

Candidates grouped by race/ethnicity showed similar differences in bar exam performance and pass rates across July exams. Bar exam performance and pass rates tended to increase for each group, particularly when comparing July 2015 to July 2017 (see Figures 1c and 1d for bar exam performance for both sample groups; see Figures 2c and 2d for pass rates for both groups). An exception was that the Black/African American group had mean bar exam scores that increased slightly and pass rates that decreased slightly between July 2015 and July 2016[6] and mean bar exam scores and pass rates that subsequently increased between July 2016 and July 2017, more than the other groups.

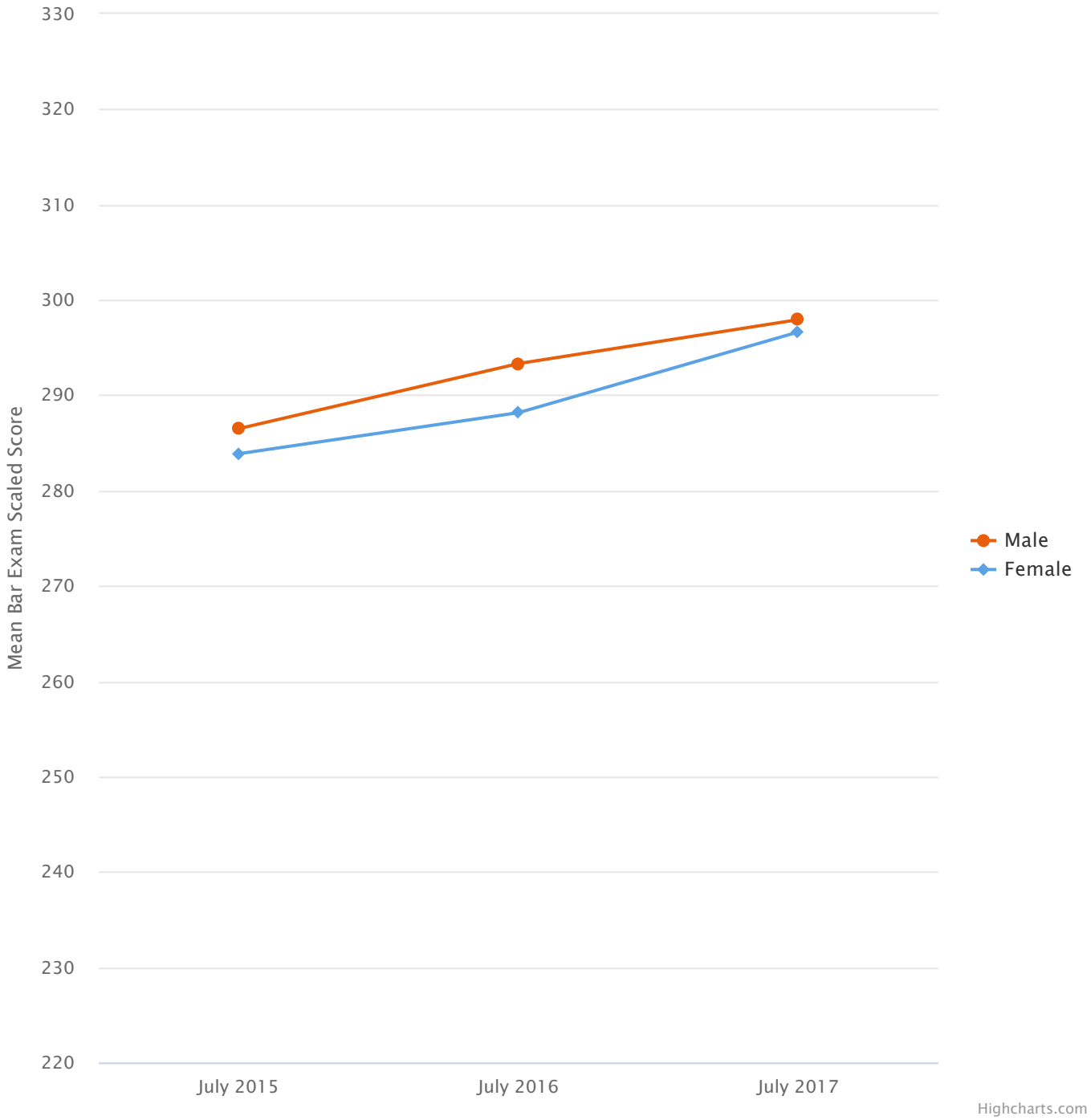
Because the composition and characteristics of candidates taking the bar exam may change across years, more information is needed to determine the extent to which the overall improvement in average performance on the bar exam in New York was due to the UBE versus other factors. This is where studying additional information, such as candidate background characteristics like UGPA, LSAT score, and LGPA, can help to better contextualize changes in bar exam performance and put them in perspective.

Figure 1. Mean bar exam scaled scores by gender and race/ethnicity for domestic-educated NYSBLE sample and school-based sample, July administrations, 2015-2017

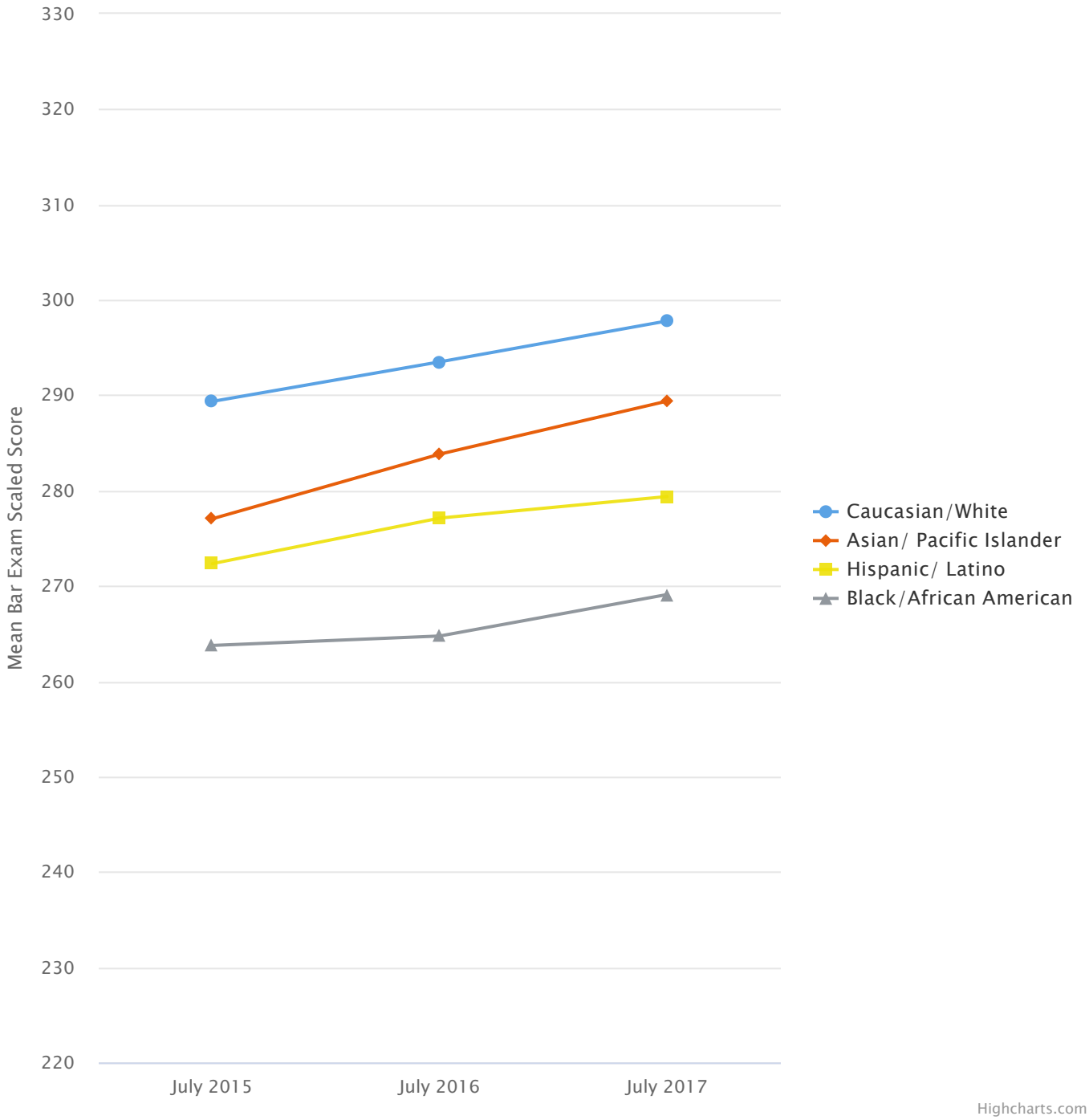
(1a) Domestic-educated NYSBLE sample



(1b) School-based sample



(1c) Domestic-educated NYSBLE sample



(1d) School-based sample

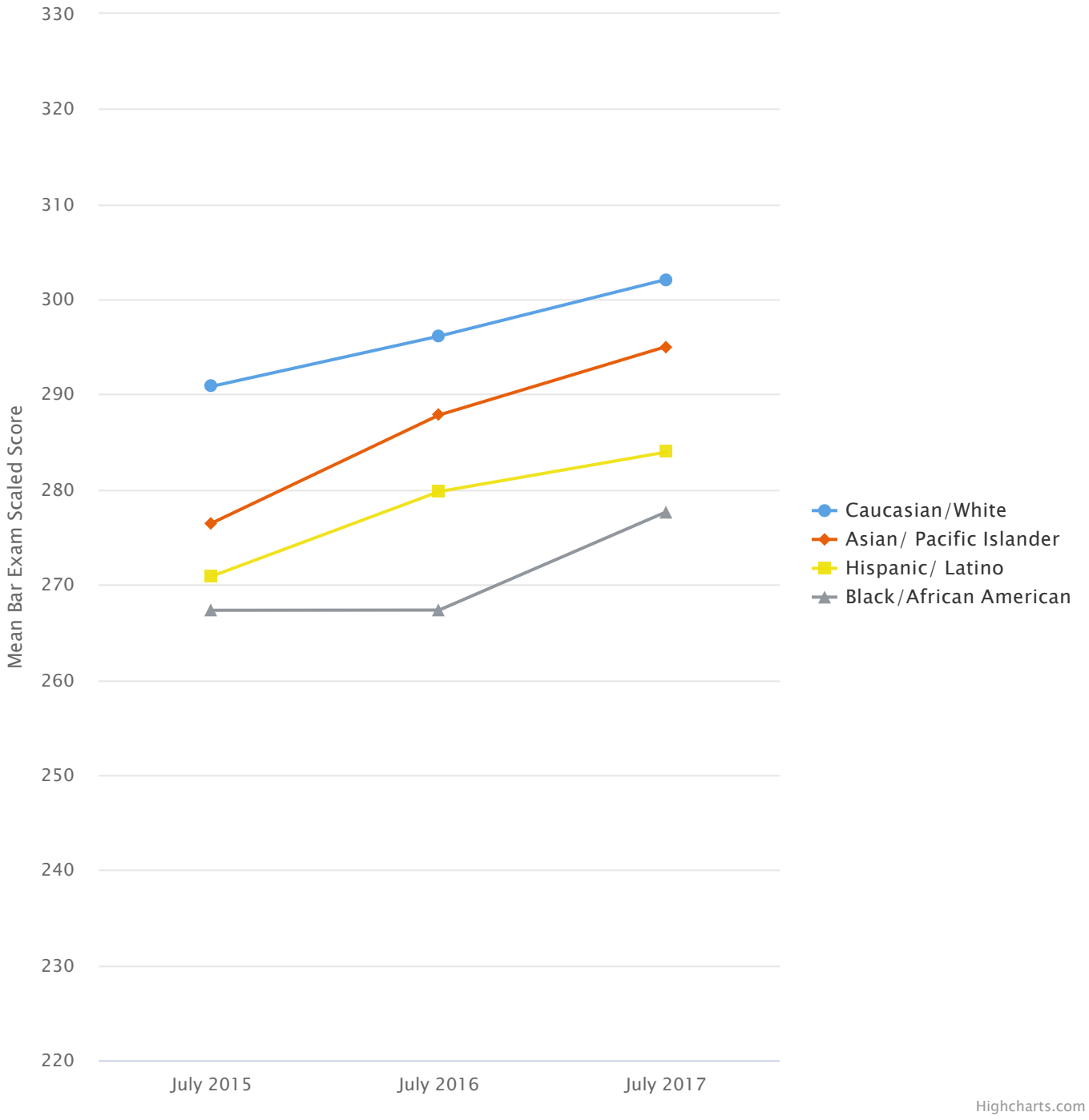
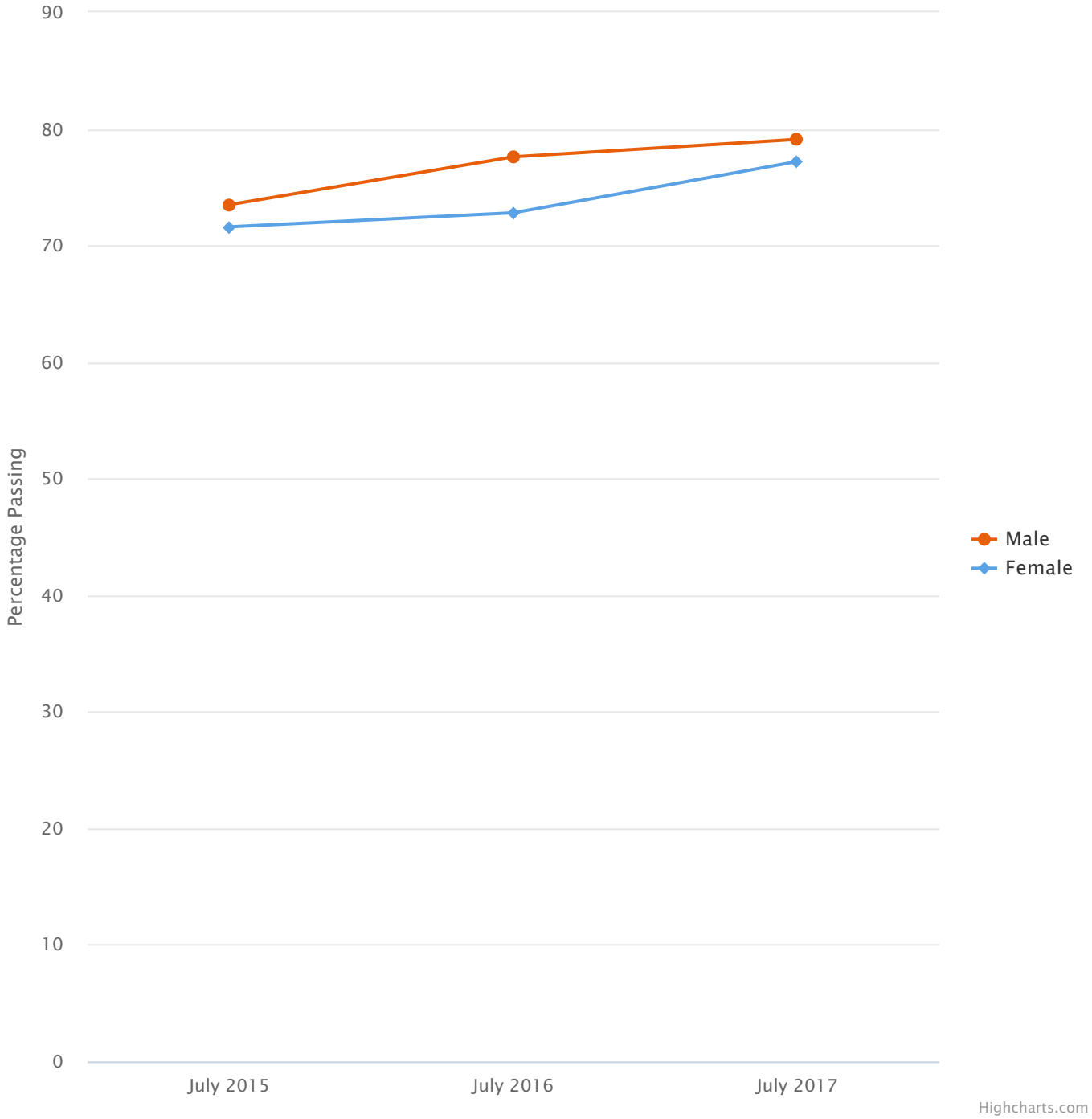
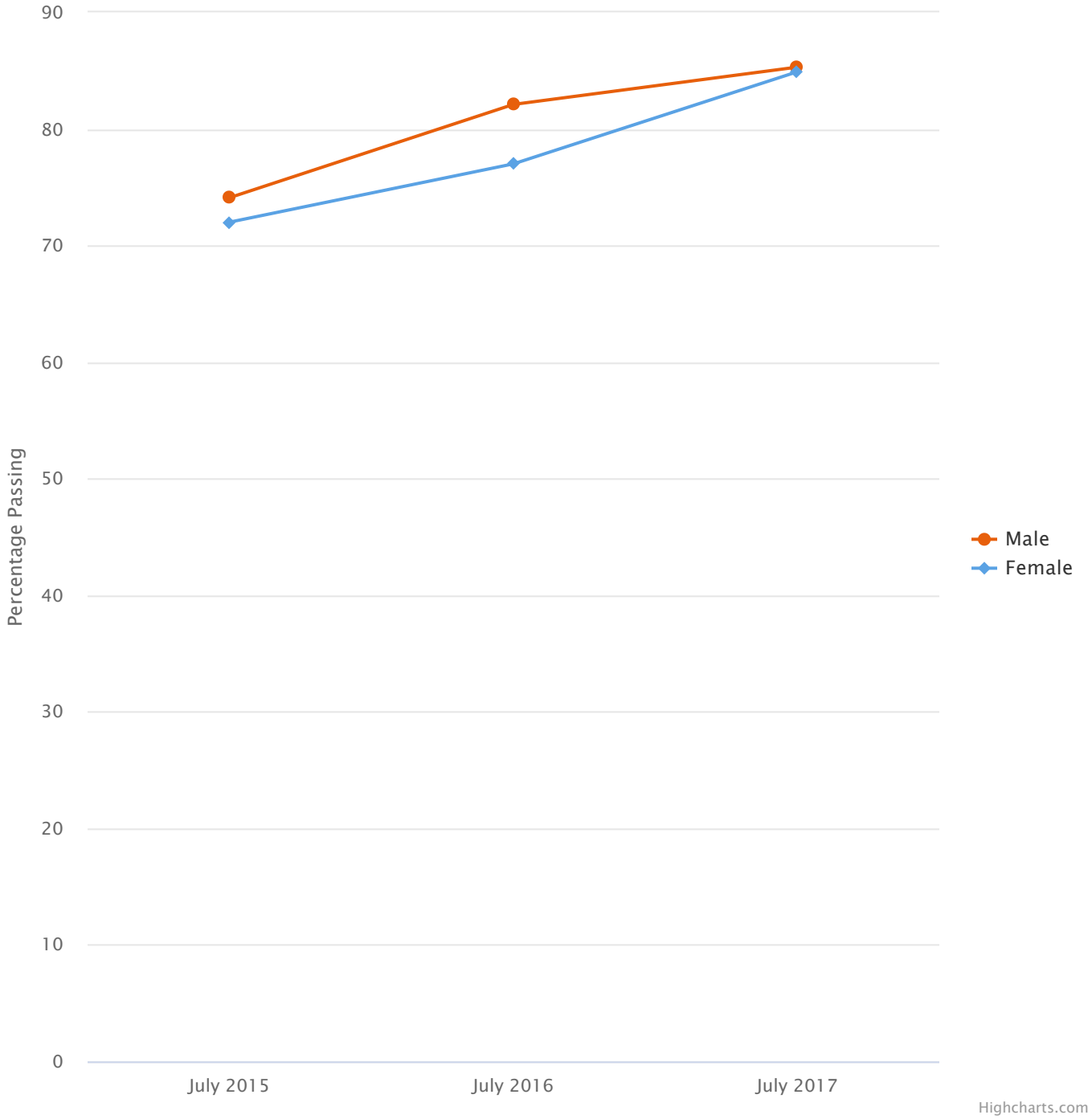


Figure 2. Pass rates by gender and race/ethnicity for domestic-educated NYSBLE sample and school-based sample, July administrations, 2015-2017

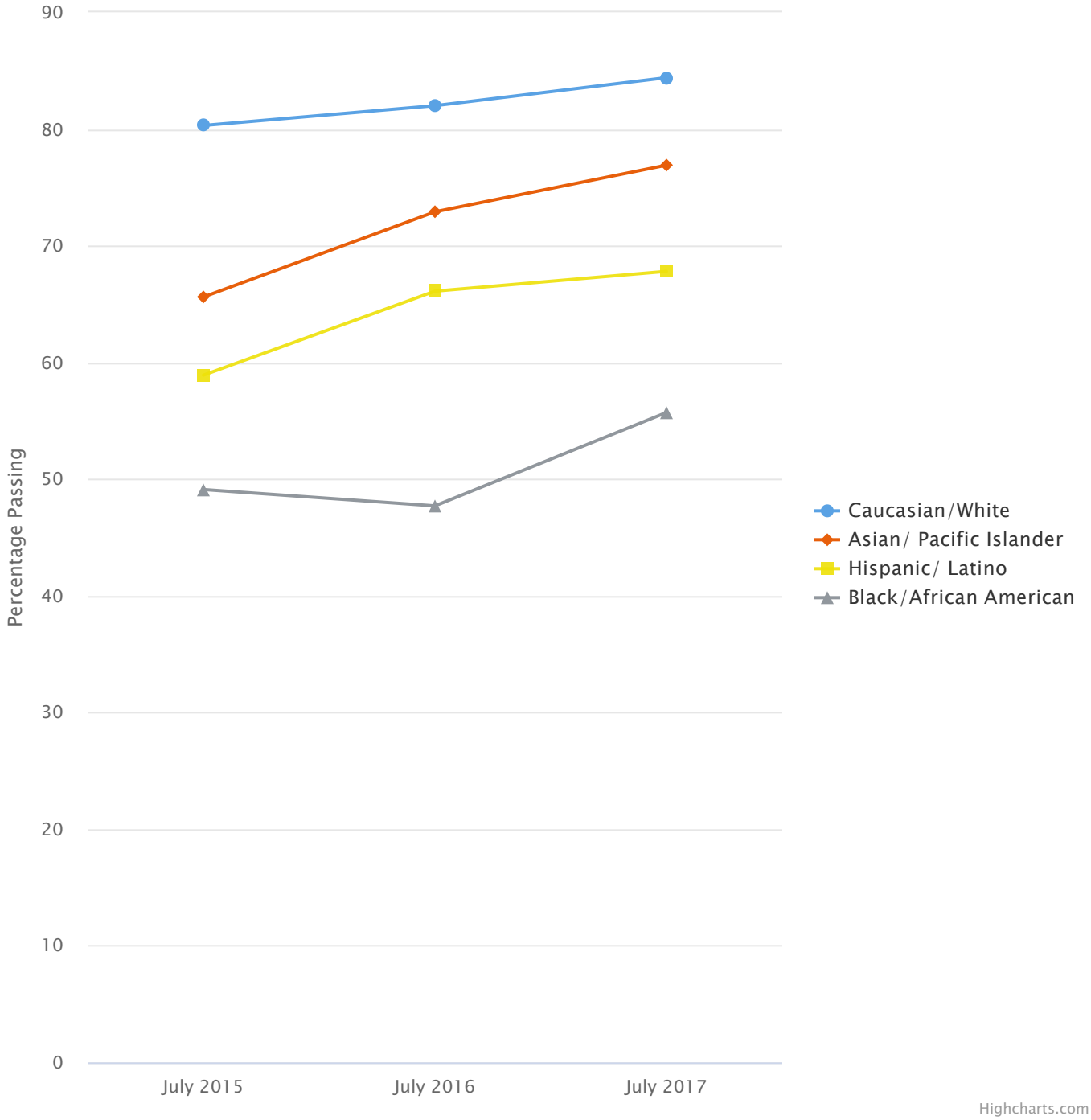
(2a) Domestic-educated NYSBLE sample



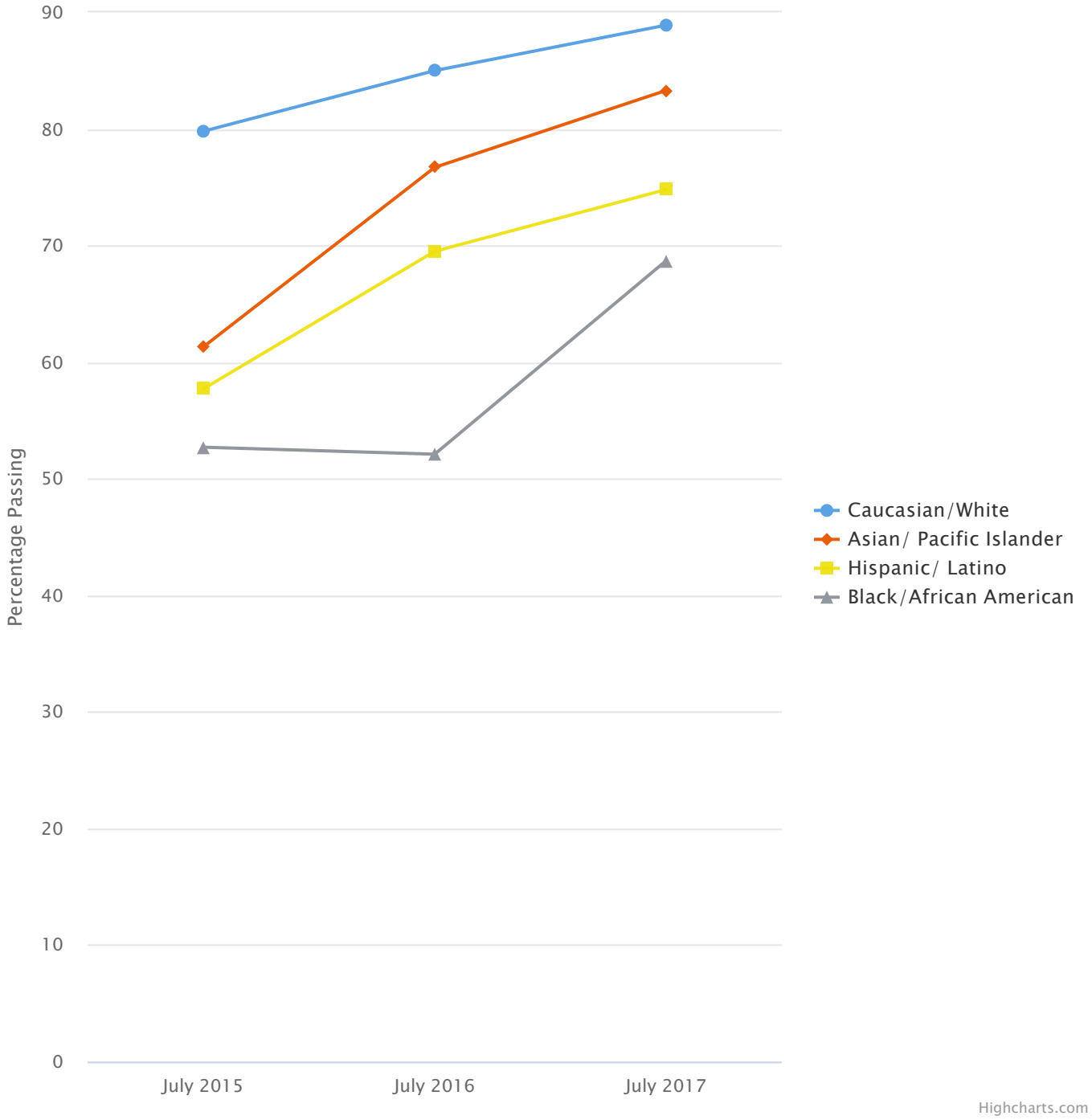
(2b) School-based sample



(2c) Domestic-educated NYSBLE sample



(2d) School-based sample



Candidate Background Characteristics

Findings of the study on candidate background characteristics by gender are shown in Figure 3:

> Of the three candidate background characteristics studied (UGPAs, LSAT scores, and LGPAs), UGPAs and LGPAs tended to remain constant or increase across the three July exams for both females and males (Figures 3a, 3e, and 3g).
>
> Mean LSAT scores decreased slightly between July 2015 and July 2016 before increasing in July 2017 (Figure 3c).

Average values for background characteristics tended to differ by gender. Females tended to have higher mean UGPAs than males for groups taking each bar exam (Figure 3a).

This pattern was reversed for LSAT scores (Figure 3c) and both the 4-point and index-based LGPAs (Figures 3e and 3g), where males tended to have higher means than females.

Differences between males and females decreased between July 2015 and July 2017 for each background characteristic (Figures 3a, 3c, 3e, and 3g).

Findings of the study on candidate background characteristics by race/ethnicity are also shown in Figure 3:

Average values for candidate background characteristics tended to differ according to candidates' race/ethnicity (Figures 3b, 3d, 3f, and 3h).

Each background characteristic between July 2015 and July 2016 tended to remain constant or increase for Asian/Pacific Islander, Black/African American, and Hispanic/Latino groups (although it can be seen in Figure 3f that the mean 4-point LGPA did dip slightly for the Black/African American group in July 2016 compared to July 2015).

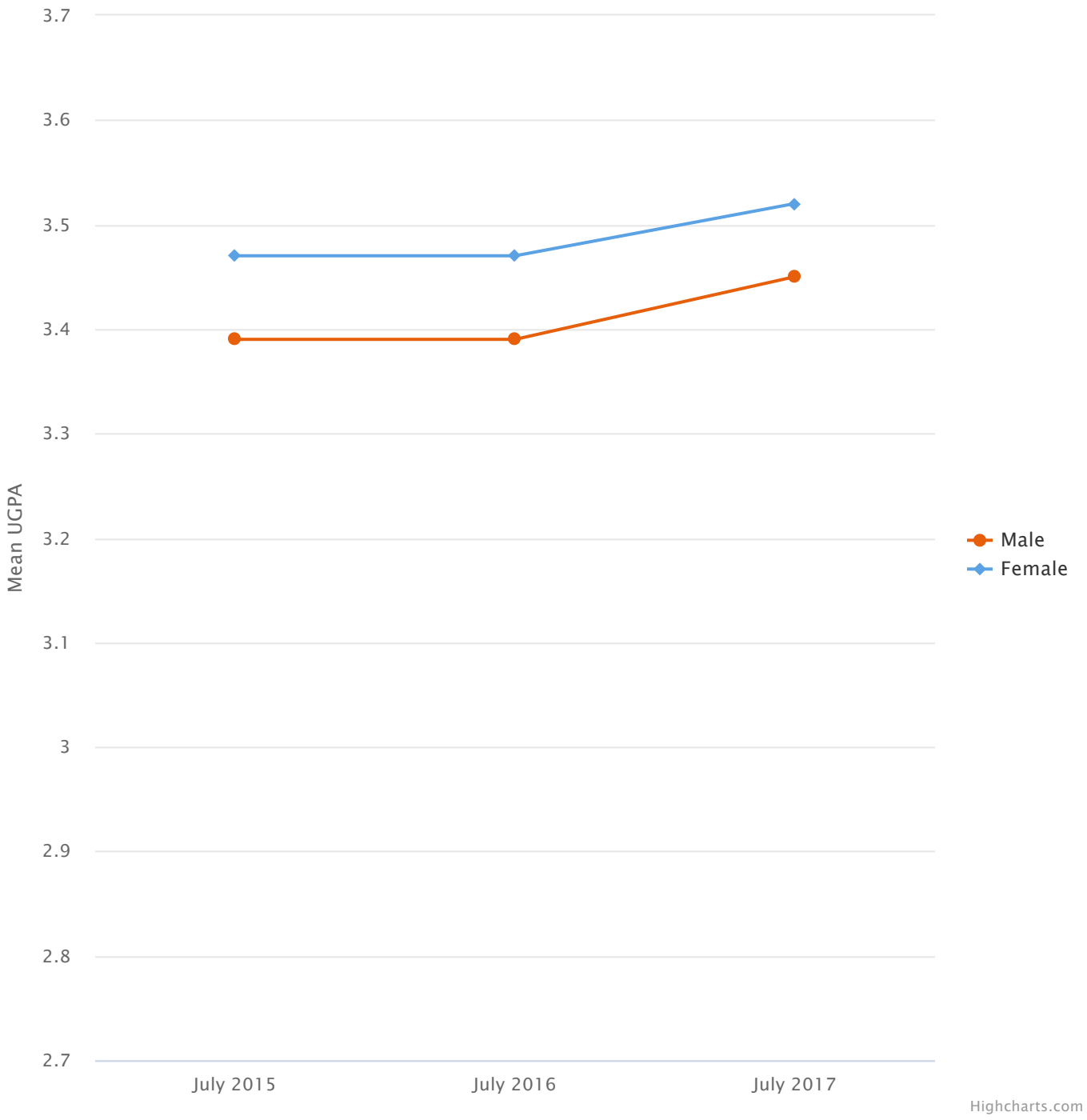
For the Caucasian/White group, each background characteristic between July 2015 and July 2016 tended to remain constant or decrease.

In July 2017, mean background characteristics tended to increase for each group, with the exception of the Hispanic/Latino group, which had similar mean UGPAs (Figure 3b) and lower mean 4-point LGPAs in July 2017 (Figure 3f) compared to July 2016.

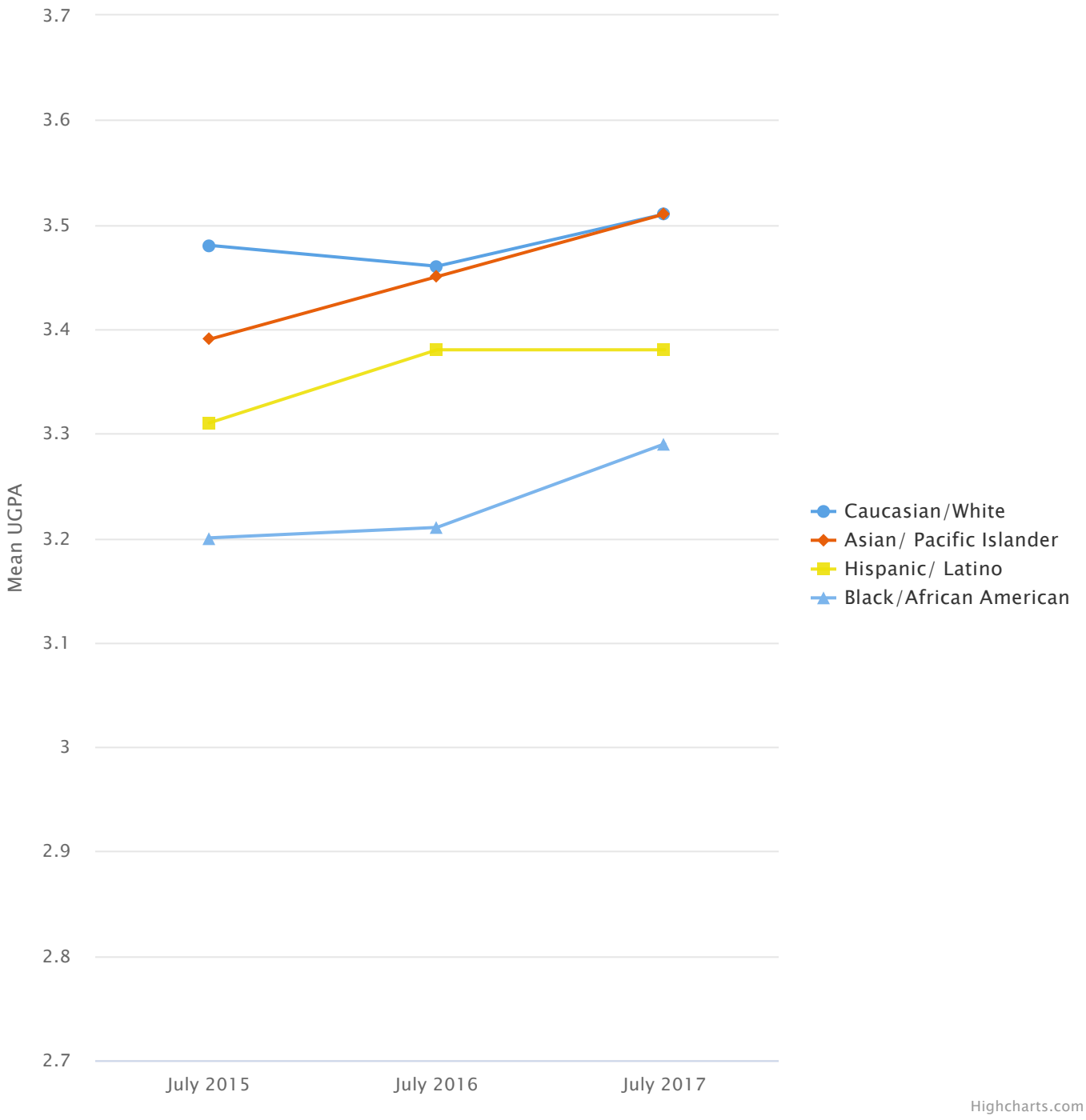
The pattern of mean UGPAs, LSAT scores, and LGPAs was generally consistent with the average performance on the bar exam between July 2015 and July 2017, where performance tended to increase.

Figure 3. Mean UGPA, LSAT score, and LGPA (4-point and index-based) by gender and race/ethnicity, July administrations, 2015-2017

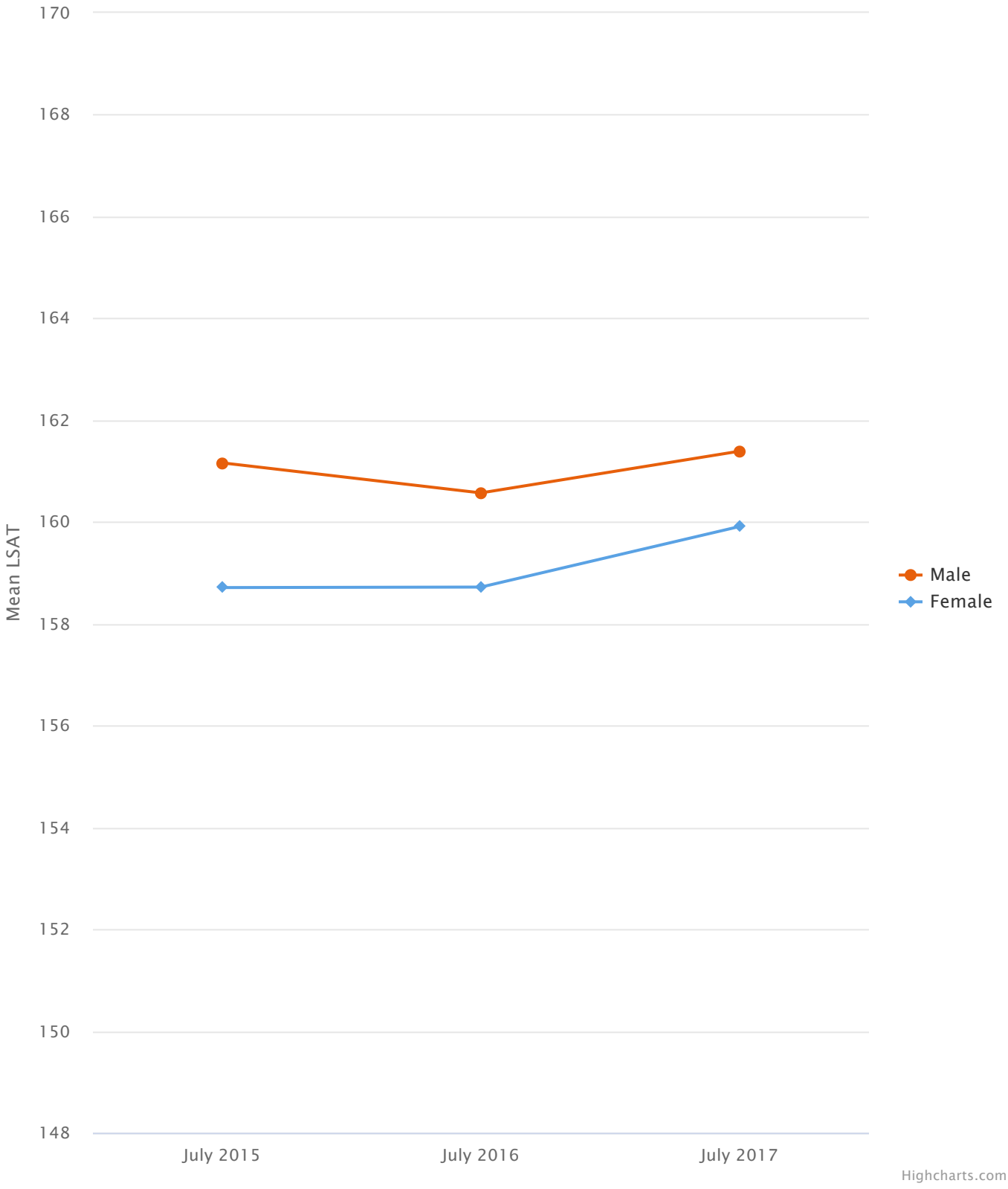
(3a)


3.7
3.6
3.5
3.4
3.3
3.2
3.1
3
2.9
2.8
2.7
Mean UGPA
July 2015
July 2016
July 2017
Male
Female
Highcharts.com

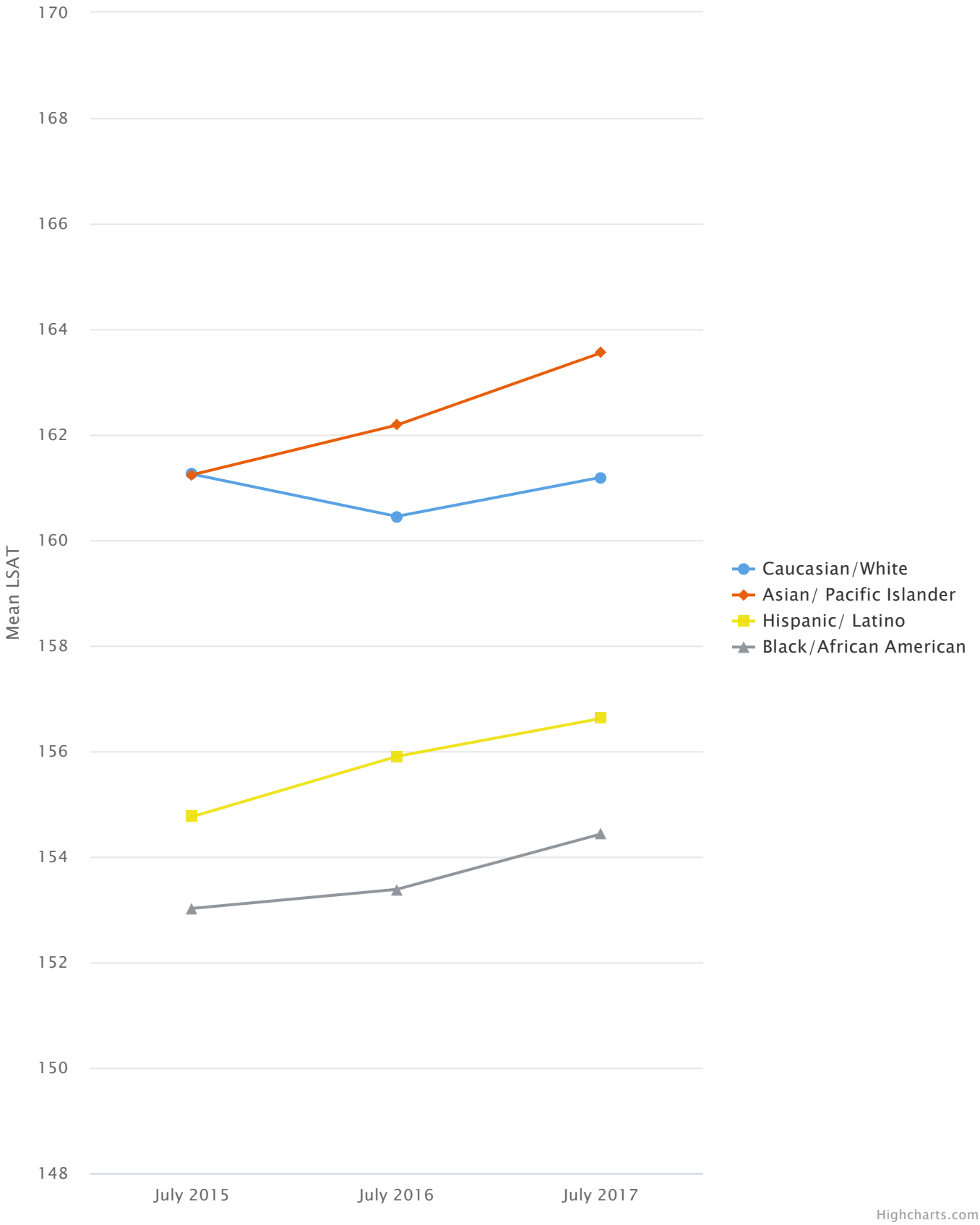
(3b)


3.7
3.6
3.5
3.4
3.3
3.2
3.1
3
2.9
2.8
2.7
Mean UGPA
July 2015
July 2016
July 2017
Caucasian/White
Asian/ Pacific Islander
Hispanic/ Latino
Black/African American
Highcharts.com

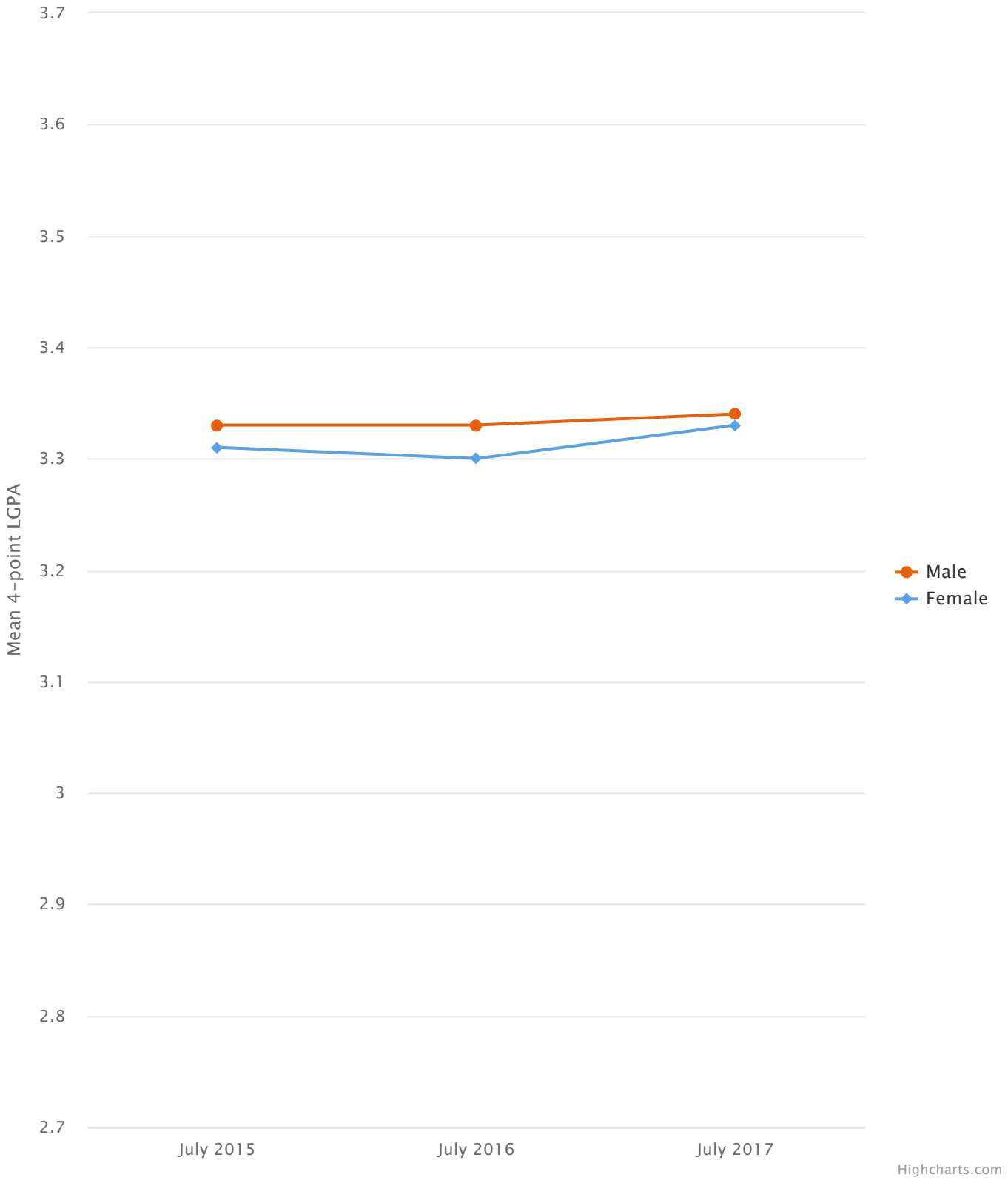
(3c)


170
168
166
164
162
160
158
156
154
152
150
148
Mean LSAT
July 2015
July 2016
July 2017
Male
Female
Highcharts.com

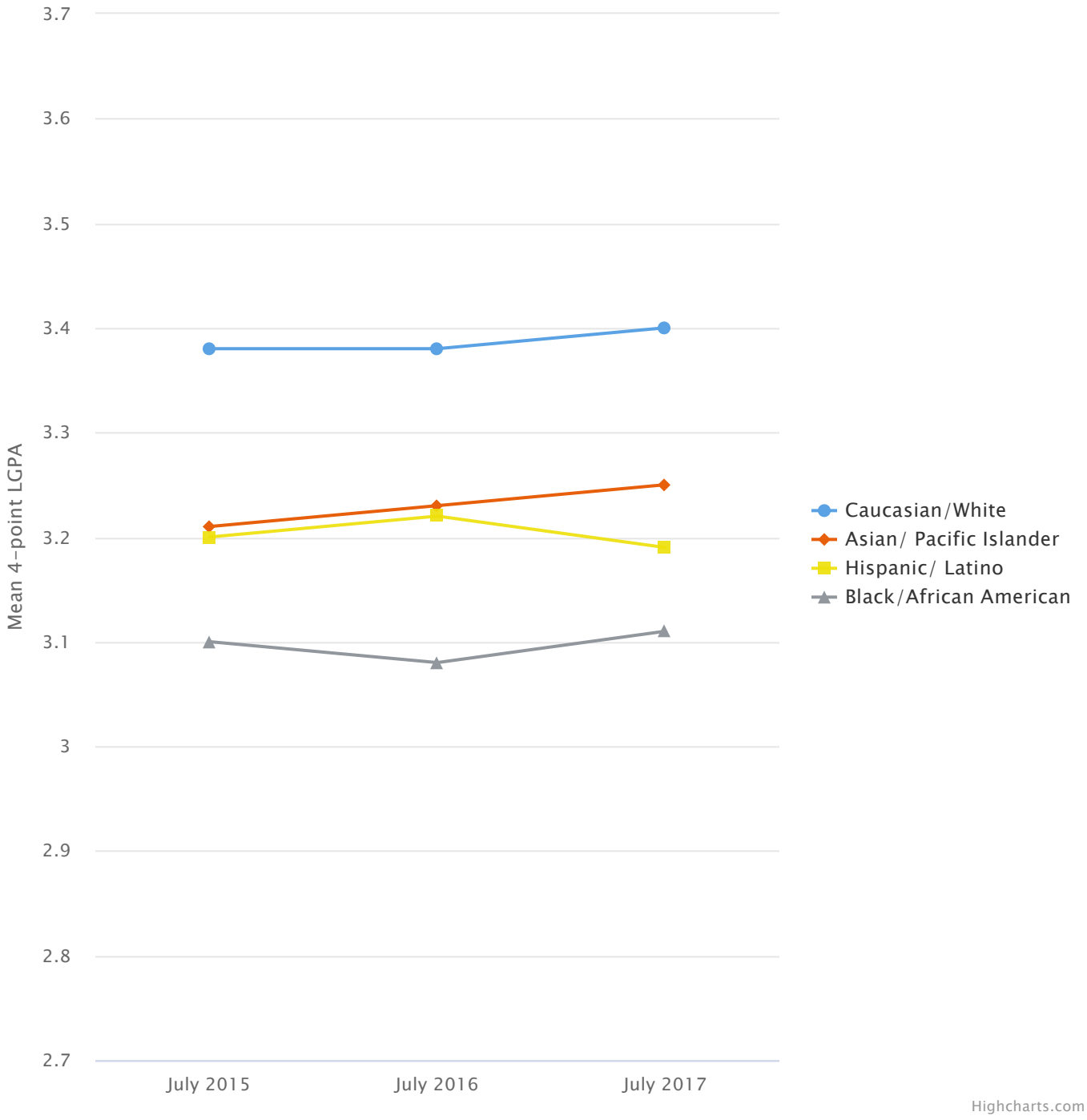
(3d)


170
168
166
164
162
160
158
156
154
152
150
148
Mean LSAT
July 2015
July 2016
July 2017
Caucasian/White
Asian/ Pacific Islander
Hispanic/ Latino
Black/African American
Highcharts.com

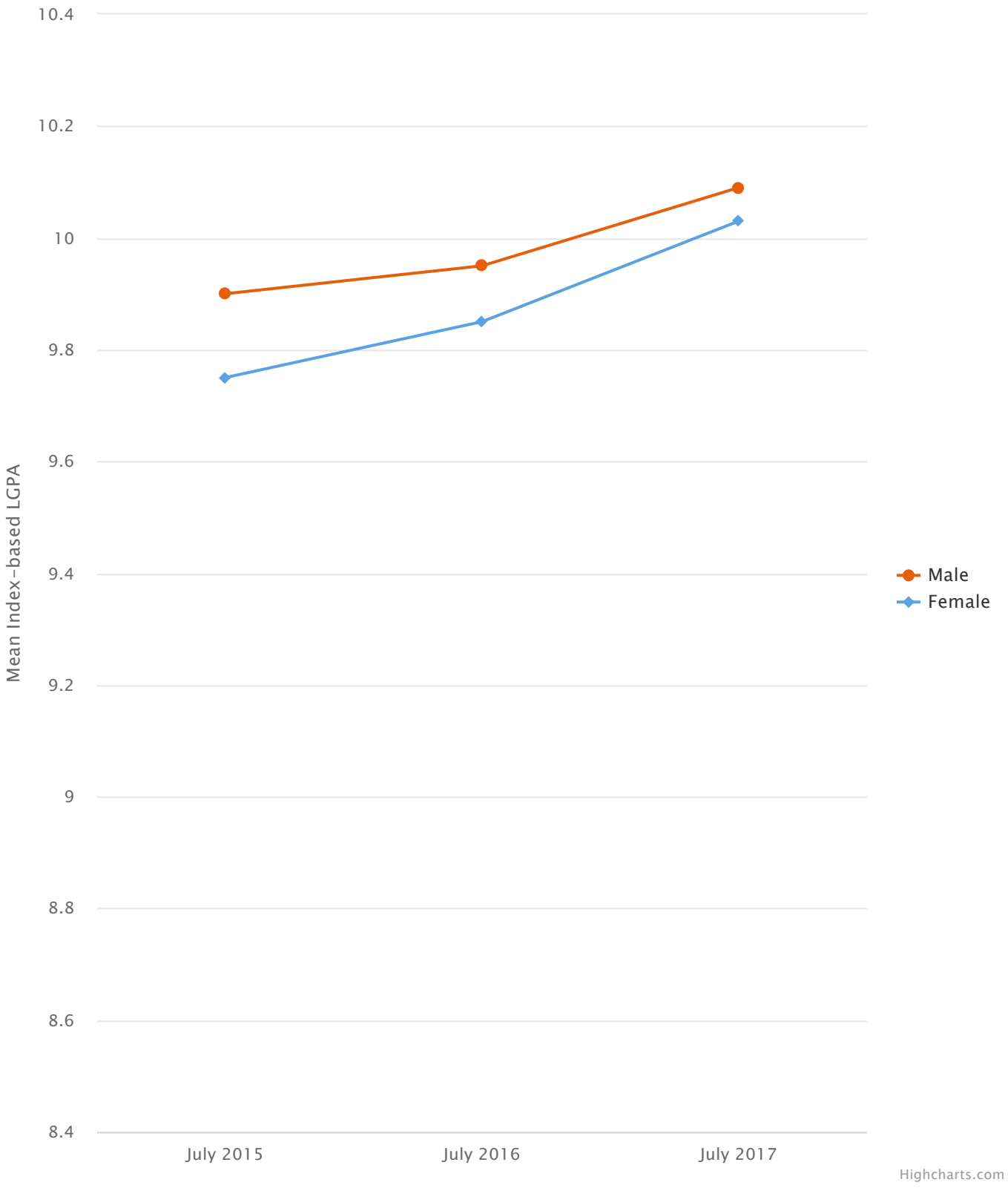
(3e)


3.7
3.6
3.5
3.4
3.3
3.2
3.1
3
2.9
2.8
2.7
Mean 4-point LGPA
July 2015
July 2016
July 2017
Male
Female
Highcharts.com

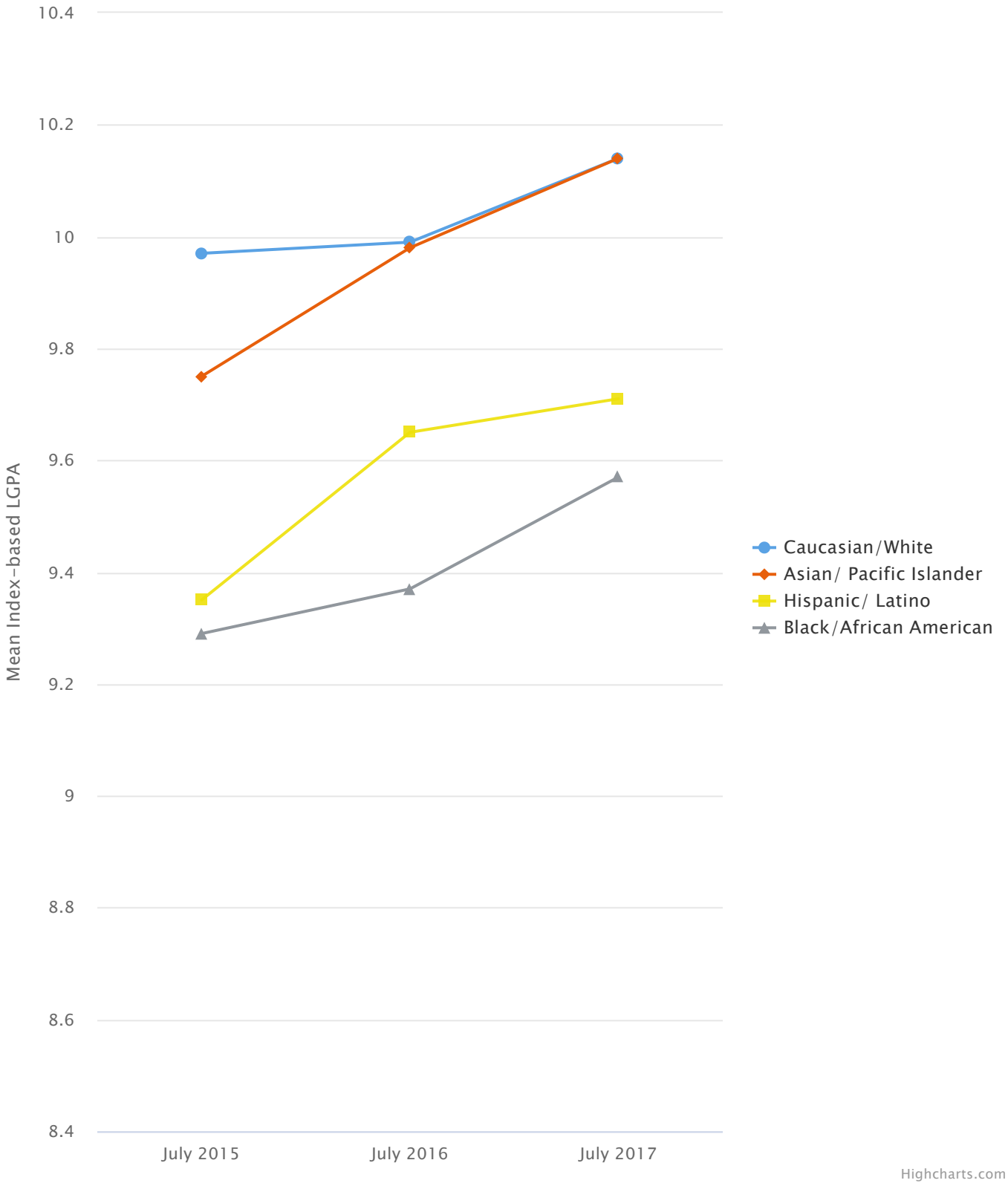
(3f)


3.7
3.6
3.5
3.4
3.3
3.2
3.1
3
2.9
2.8
2.7
Mean 4-point LGPA
July 2015
July 2016
July 2017
Caucasian/White
Asian/ Pacific Islander
Hispanic/ Latino
Black/African American
Highcharts.com

(3g)


10.4
10.2
10
9.8
9.6
9.4
9.2
9
8.8
8.6
8.4
Mean Index-based LGPA
July 2015
July 2016
July 2017
Male
Female
Highcharts.com

(3h)



The Relationship Between Bar Exam Performance and Candidate Background Characteristics

UGPA, LSAT score, 4-point LGPA, and index-based LGPA each had a relatively strong, statistically significant positive relationship with bar exam score, where a positive relationship indicates that an

increase in background characteristic is associated with an increase in bar exam score. UGPA had the weakest relationship with bar exam score, and index-based LGPA had the strongest relationship, followed by LSAT score and then 4-point LGPA. Each relationship would be considered moderately strong to strong.[7]

One way of illustrating the relationship between the background characteristics and bar exam performance is to show how candidates at different levels of background characteristics performed on the bar exam. Figure 4 shows mean bar exam scores for candidates grouped by the three background characteristics (UGPA, LSAT score, or LGPA). For example, Figure 4a plots mean bar exam scores for candidates with UGPAs below 2.50 on the far left, then candidates with UGPAs between 2.50 and 2.69, and so on. Candidates with UGPAs below 2.50 had mean bar exam scaled scores between 254 and 267 depending on the year, and candidates with UGPAs above 3.89 (far right) had mean bar exam scaled scores between 303 and 316. Mean bar exam scaled scores increased as UGPAs increased from left to right across the figure, showing a moderately strong positive relationship.

Mean bar exam scaled scores also increased as LSAT scores and LGPAs increased (Figures 4b, 4c, and 4d). Another way of summarizing the positive relationships in these figures is that they illustrate that candidates with lower UGPAs, LSAT scores, or LGPAs tended to score lower on the bar exam, and those with higher UGPAs, LSAT scores, or LGPAs tended to score higher. The upward shift in the three lines corresponding to the three different years of scores indicates that as the years progressed, bar exam scores increased.

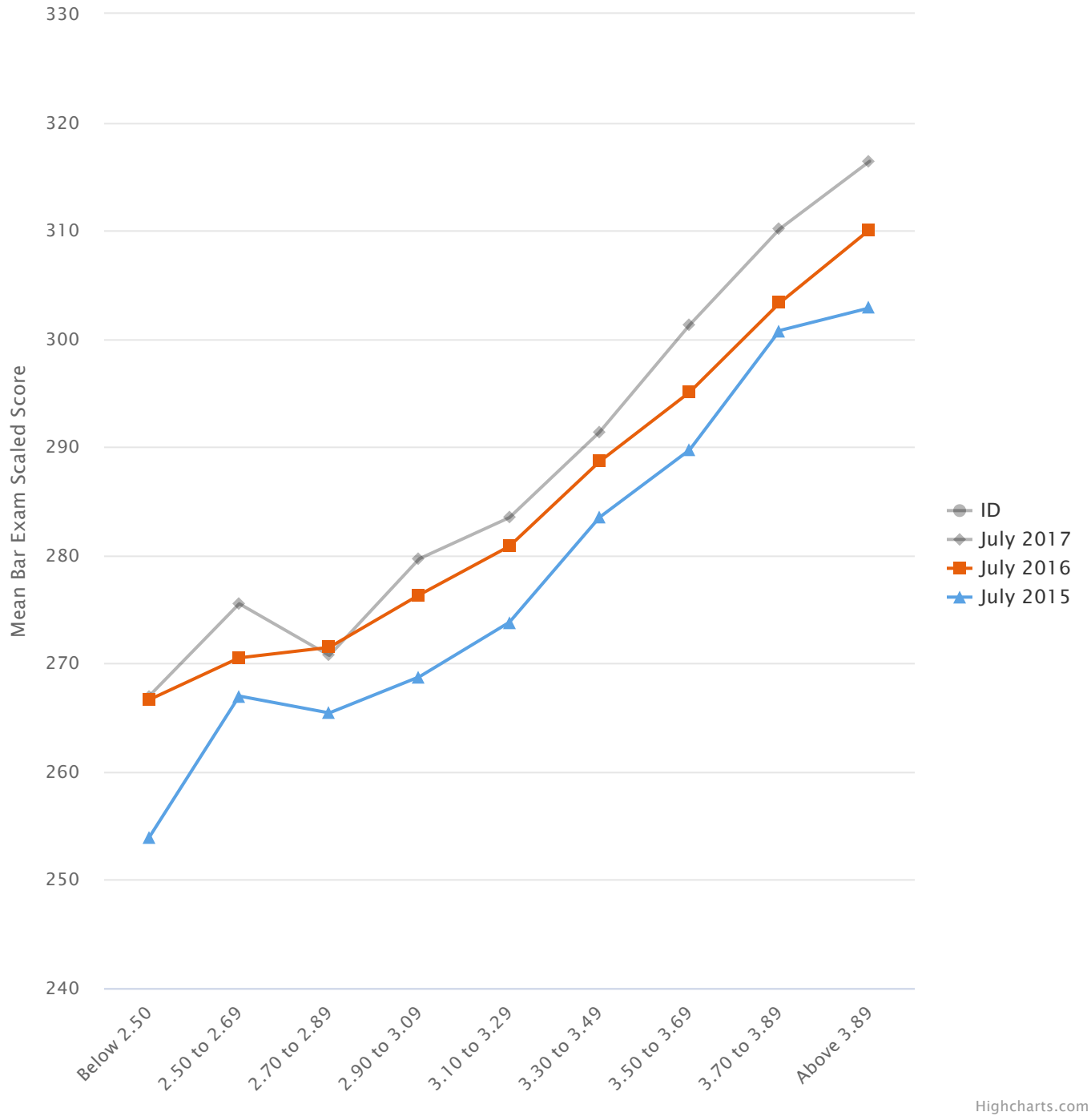
This study illustrated that for a bar exam like the one in New York, which already used the MBE and one MPT question on its exam prior to UBE adoption, the effects of UBE adoption were at most small and likely positive.

Figure 4 also illustrates that the differences in lines relating the candidate background characteristics and bar exam scores across the three years were relatively similar for the three background variables shown in 4a, 4b, and 4c. What is notable is that the differences in the lines almost disappeared in 4d, which relates index-based LGPAs to bar exam scores. The index-based LGPAs can be considered LGPAs adjusted

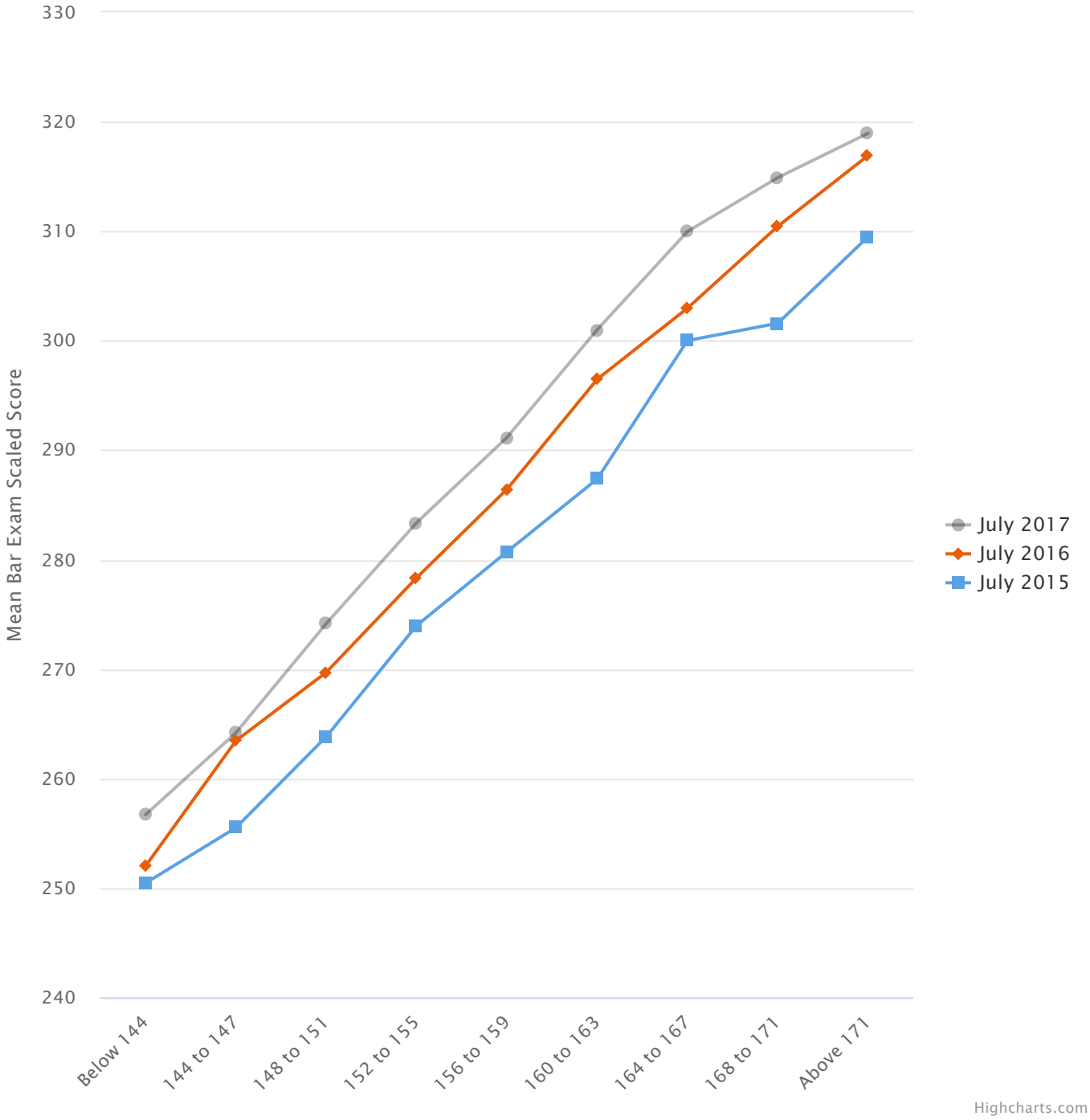
for differences in law-school-level UGPAs and LSAT scores. Thus, the increase in mean bar exam scores across the three years noted earlier and seen in the upward shift in the lines across the three years in figures 4a, 4b, and 4c was mostly removed when using index-based LGPAs. The lines are particularly close for the bar exam score of 266 where New York sets its passing score. The full report and appendices provide additional analyses that further reinforce that most of the increases observed in average bar exam scores across the three years could be accounted for by a combination of UGPAs, LSAT scores, and LGPAs. In other words, improvement in the UGPAs and LSAT scores of candidates on entry to law school and their subsequent performance in law school (LGPAs) accounted for most of the improvement in bar exam scores across the three July exams.

Figure 4. Correlations between mean UGPA, LSAT score, and LGPA (4-point and index-based) and mean bar exam scaled score, July administrations, 2015-2017

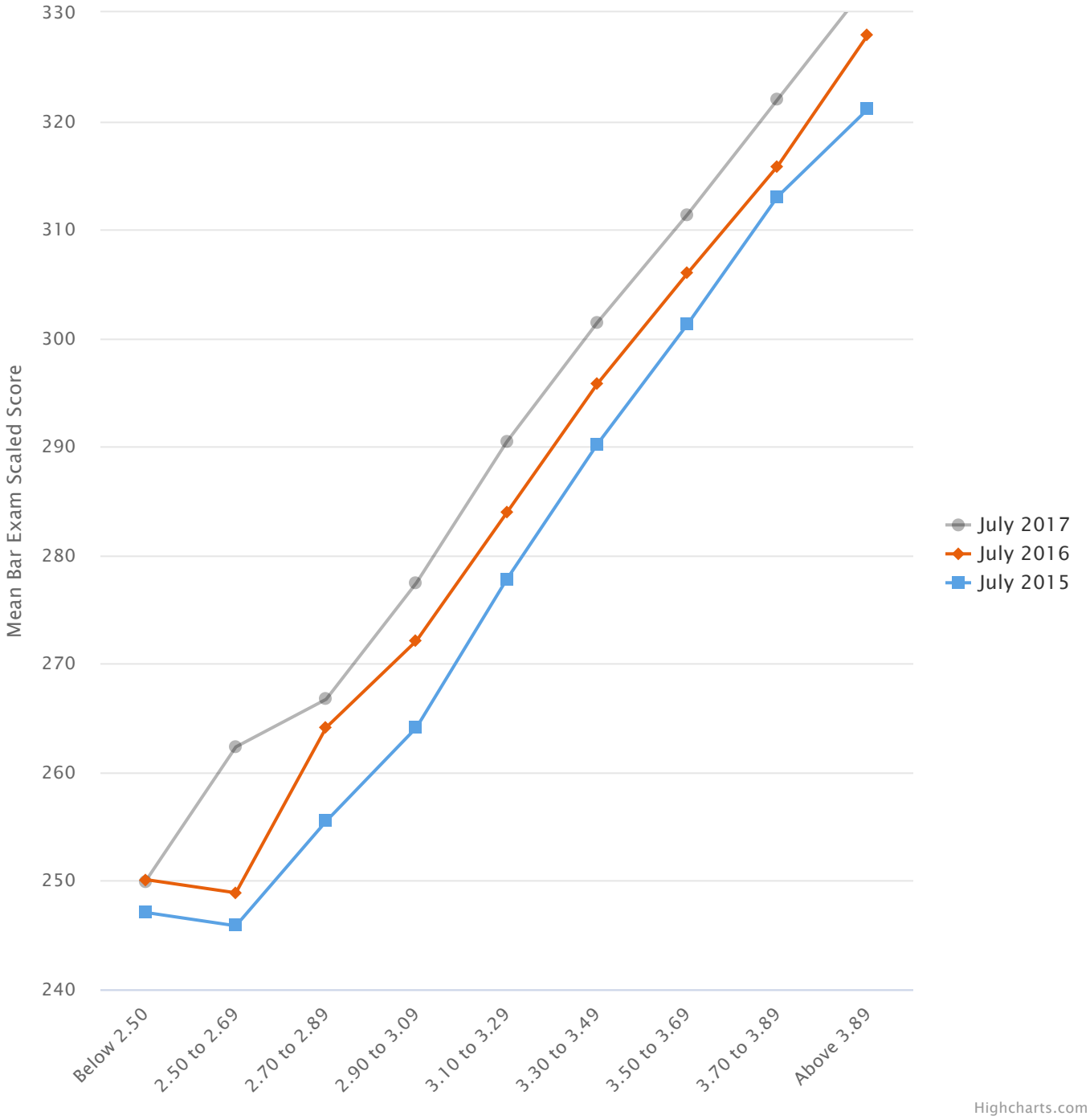
(4a)


330
320
310
300
290
280
270
260
250
240
Mean Bar Exam Scaled Score
Below 2.50
2.50 to 2.69
2.70 to 2.89
2.90 to 3.09
3.10 to 3.29
3.30 to 3.49
3.50 to 3.69
3.70 to 3.89
Above 3.89
ID
July 2017
July 2016
July 2015
Highcharts.com

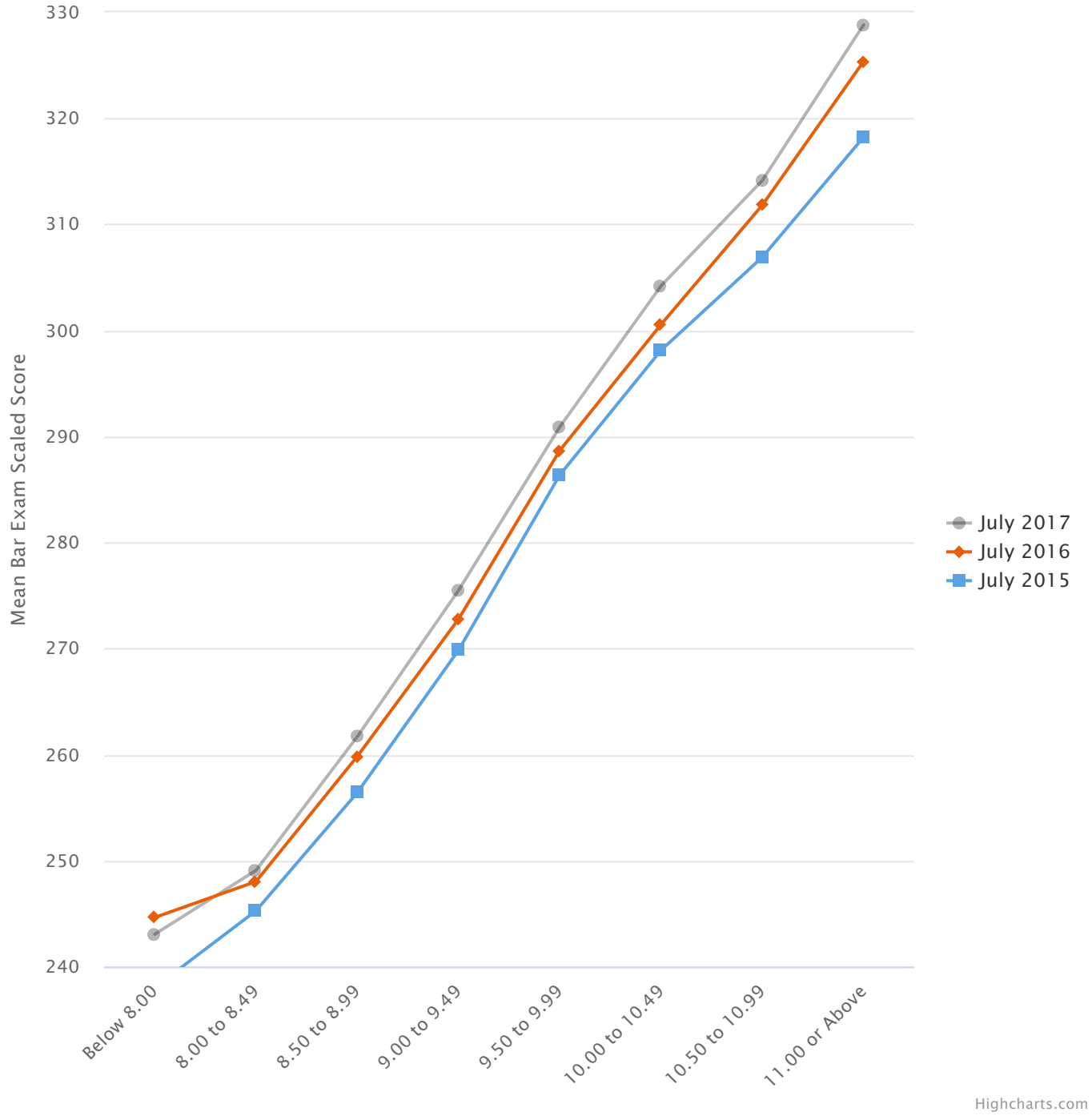
(4b)


330
320
310
300
290
280
270
260
250
240
Mean Bar Exam Scaled Score
Below 144
144 to 147
148 to 151
152 to 155
156 to 159
160 to 163
164 to 167
168 to 171
Above 171
July 2017
July 2016
July 2015
Highcharts.com

(4c)


330
320
310
300
290
280
270
260
250
240
Mean Bar Exam Scaled Score
Below 2.50
2.50 to 2.69
2.70 to 2.89
2.90 to 3.09
3.10 to 3.29
3.30 to 3.49
3.50 to 3.69
3.70 to 3.89
Above 3.89
July 2017
July 2016
July 2015
Highcharts.com

(4d)



Conclusions

Mean bar exam scores and pass rates on the bar exam in New York increased, on average, after UBE adoption, and the improvement in performance was explained in large part by improvements in the UGPAs, LSAT scores, and LGPAs of candidates taking the New York bar exam. The improvement in bar exam performance in New York after UBE adoption was likely not attributable to the UBE. In addition, the UBE did not have sustained or adverse effects on candidates in New York compared to the prior bar exam. Differences in pass rates and average bar exam scores on the UBE observed across groups defined by gender and race/ethnicity were also observed prior to UBE adoption in New York.

This article provides only a glimpse of the results from the study that NCBE conducted for the NYSBLE. The full report addresses the topics described here in more detail and addresses other topics, including the performance of repeat test takers who did not pass the bar exam on their first attempt and MBE performance in New York compared to all other jurisdictions. This study illustrated that for a bar exam like the one in New York, which already used the MBE and one MPT question on its exam prior to UBE adoption, the effects of UBE adoption were at most small and likely positive.

Notes

1. The press release, Executive Summary, full report, and appendices are available on the New York State Board of Law Examiners' website: Impact of Adoption of the Uniform Bar Examination in New York, https://www.nybarexam.org/UBEReport.html. *(Go back)*
2. Results for each topic were further analyzed by gender and race/ethnicity. For a more complete description of the data and extensive analysis of the samples of data, see the full report and associated appendices at https://www.nybarexam.org/UBEReport.html. *(Go back)*
3. In New York, roughly 30% of candidates in July and 40% of candidates in February received their legal education outside the United States. *(Go back)*
4. School-level UGPAs and LSAT scores were used to adjust LGPAs such that index-based LGPAs as a group from a more selective school (based on UGPAs and LSATs at that school) would be higher than index-based LGPAs from a less selective school (based on UGPAs and LSATs at that school). Otherwise put, if two candidates from different law schools have the same LGPA, the candidate from the more selective school would generally have the higher index-based LGPA. (The index referred to is one that was computed for this scaling process based on each candidate's LSAT score and UGPA.) Index-based LGPAs were placed on a scale that ranged from roughly 1 to 15 with most values typically near 10 and a range typically falling between roughly 7 and 13. *(Go back)*
5. Average bar exam scores were 283.04 in July 2015, 286.85 in July 2016, and 290.98 in July 2017. *(Go back)*
6. The increase in mean bar exam score but decrease in pass rate may seem counterintuitive but has to do with shifts in the distributions of scores between July 2015 and July 2016 for the Black/African American group. (See Figure 4.2.29, Distributions of July Bar Exam Scores, Black/African American Candidates, New York State Board of Law Examiners Sample, on page 150 of the full report.) *(Go back)*
7. Correlations between UGPAs and bar exam scaled scores were 0.46, 0.40, and 0.45 in July 2015, 2016, and 2017, respectively. The corresponding correlations between LSAT scores and bar exam scores were 0.56, 0.57, and 0.57. The 4-point LGPA correlations were 0.65, 0.61, and 0.61, while the index-based LGPA correlations were 0.76, 0.75, and 0.75. (See Table 3.7.1, Correlations among UGPA, LSAT Scores, LGPA, MBE, Written Scores and Bar Exam Scores, School-Based Sample, on page 109 of the full report for additional details.) *(Go back)*

Andew A. Mroch, PhD, is Senior Research Psychometrician for the National Conference of Bar Examiners.

Mark A. Albanese, PhD, is the Director of Testing and Research for the National Conference of Bar Examiners.

Contact us to request a pdf file of the original article as it appeared in the print edition.

Membership Type Count as of	12/31/2010	12/31/2011	12/31/2012	12/31/2013	12/30/2014	12/31/2015
Start of year	14,254	14,607	14,711	14,876	14,893	14,888
End of year	14,607	14,711	14,876	14,893	14,888	14,784
Net gain (loss)	353	104	165	17	-5	-104
% gain or loss	2.42%	0.71%	1.11%	0.11%	-0.03%	-0.70%
State % increase (loss) in pop	0.53%	0.53%	0.84%	1.05%	1.23%	1.49%
total lawyers needed for pop	14,330	14,684	14,835	15,032	15,076	15,110
lawyers above (below) need	277	27	41	-139	-188	-326

Membership Type Count as of	12/31/2016	12/31/2017	12/31/2018	12/31/2019	12/31/2020	5/3/2021
Start of year	14,784	14,851	14,802	14,802	14,903	14,622
End of year	14,851	14,802	14,802	14,903	14,622	
Net gain (loss)	67	-49	0	101	-281	
% gain or loss	0.45%	-0.33%	0.00%	0.68%	-1.92%	
State % increase (loss) in pop	1.52%	1.24%	0.92%	0.68%	0.38%	0.68%
total lawyers needed for pop	15,009	15,035	14,938	14,903	14,960	
lawyers above (below) need	-158	-233	-136	0	-338	-1,172

There is a great need for civil legal services for low and moderate income people in Oregon that is not adequately met by the existing legal services delivery network.

EXECUTIVE SUMMARY

This report, commissioned by the Oregon State Bar, examines the civil legal needs of low and moderate income Oregonians. The survey was also sponsored by the Oregon Judicial Department and the Office of Governor John Kitzhaber, M.D. The primary source of data used in this study is a legal needs survey of 1,011 low and moderate income persons conducted with the assistance of Portland State University throughout Oregon during the fall and winter of 1999-2000. Additional information was provided by judges, lawyers, social service workers, community leaders and legal services providers through focus groups, interviews and surveys.

Summary of Findings from Judges, Lawyers, Social and Legal Services Providers

- There is a great need for civil legal services for low and moderate income people in Oregon that is not adequately met by the existing legal services delivery network.

- More services are needed in the area of family law, particularly in child custody and domestic violence cases. Part of that need can be met by providing advice and other limited services short of full representation. Court representation is needed in cases where the opposing party is represented or there is an imbalance of power.

- Housing advocacy to increase the quantity and quality of housing for low income people, reduce the incidence of unlawful discrimination, enforce the residential landlord tenant act and provide sufficient self-help information to assert defenses in eviction actions is a priority need that is insufficiently unmet.

- Employment law issues such as collection of wages, wrongful discharge, discrimination, and unsafe working conditions are an important emerging area of unmet legal need.

- The unmet need for services is not limited to the foregoing substantive areas, but includes a wide range of other issues discussed in this report.

- There is a need to provide targeted services to particular client groups who often encounter unique substantive legal issues or face special barriers to access to the legal system, such as the disabled, the elderly, farm workers, immigrants, Native Americans, the non-English speaking, and youth.

- There is a significant unmet need for outreach, community education and access to easily used, high-quality self-help materials.

- A full range of legal assistance should be available to low and moderate income Oregonians, including community education, outreach, advice, transactional assistance, direct representation of individuals in court, multi-party and class litigation, lobbying and administrative advocacy. These services should be available to all, without regard to legal status or remote geographical location.

Summary of Findings: Oregon Legal Needs Survey of Low and Moderate Income Oregonians

- The highest needs for legal assistance arise in housing, public services, family, employment and consumer cases.

- Other areas of high need for particular population groups include elder abuse, education, farm worker statutory, and immigration issues.

- Lower income people obtain legal assistance for their problems less than 20% of the time. 9.6% of all cases are handled by legal aid attorneys, 4.3% are handled by the private bar on a *pro bono* or reduced fee basis, and 3.8% are handled for full fees.

- Particular population groups examined in the study have unique legal needs that often require specialized services or approaches.

People obtaining representation have a much more favorable view of the legal system and are satisfied with the outcome of the case 75% of the time when represented by a legal services lawyer.

- Most people who experience a legal need and don't obtain representation feel very negatively about the legal system and about 75% are dissatisfied with the outcome of the case.

- People obtaining representation have a much more favorable view of the legal system and are satisfied with the outcome of the case 75% of the time when represented by a legal services lawyer.

- Lack of legal information, ignorance of resources and remedies, unavailability of convenient services and fear of retaliation are the most significant factors causing lower income Oregonians not to seek legal representation when they have a legal problem.

Capacity of Existing Services to Meet Needs of the Low and Moderate Income

A network of existing resources currently addresses the civil legal needs of low and moderate income Oregonians. Legal services are provided at no cost by basic and specialized legal services entities. Private lawyers also provide free, or *pro bono*, services through a range of programs, and assist with low cost representation through the Modest Means Program of the Oregon State Bar. Unrepresented litigants are assisted by court staff, social and educational institutions, the Oregon State Bar's Tel-Law program, libraries and the legal services programs. Agencies of the state assist with resolution of some legal problems of lower income Oregonians.

Six legal services programs comprise the basic legal services network in the state, Legal Aid Services of Oregon (LASO)(12 field offices); Oregon Law Center (OLC)(four field offices); Center for Nonprofit Legal Services (Medford); Marion-Polk Legal Aid Services (MPLAS); Lane County Legal Aid Services (LCLAS); and Lane County Law and Advocacy Center. Among the field offices are three that serve special populations, the LASO Native American Program and the Farm Worker Programs of LASO and OLC. Farm worker attorneys from both programs also work at office sites throughout the state.

Among the specialized providers in the nonprofit legal services network are the Oregon Advocacy Center, St. Andrew Legal Clinic, St. Matthew Legal Clinic, Juvenile Rights Project, Immigration Counseling Service, Catholic Charities Immigration Program, Lutheran Family Services, SOAR, Jewish Family Services, Law School Clinics, and the Fair Housing Council of Oregon.

This system is augmented by the efforts of private lawyers working on a *pro bono* or reduced fee basis through the Modest Means Program of the Oregon State Bar. Staff of the Oregon Judicial Department play a key role in assisting unrepresented parties through formal courthouse facilitator programs, conciliation services and other informal help. The Attorney General, through the Division of Child Support, and the county district attorneys assist in establishing paternity and in collecting and modifying child support obligations. The Justice Department also works effectively on consumer fraud issues. The Bureau of Labor and Industries enforces wage and discrimination laws.

Key Findings Regarding Existing Services

- The current legal services delivery system cannot meet the critical legal needs of lower income Oregonians without additional funding.

- The current legal services delivery system is meeting the legal needs of low income people in 53,650 (or 17.8%) of the 301,944 cases a year that require a lawyer's assistance. The unmet need is estimated to be about 250,000 cases a year.

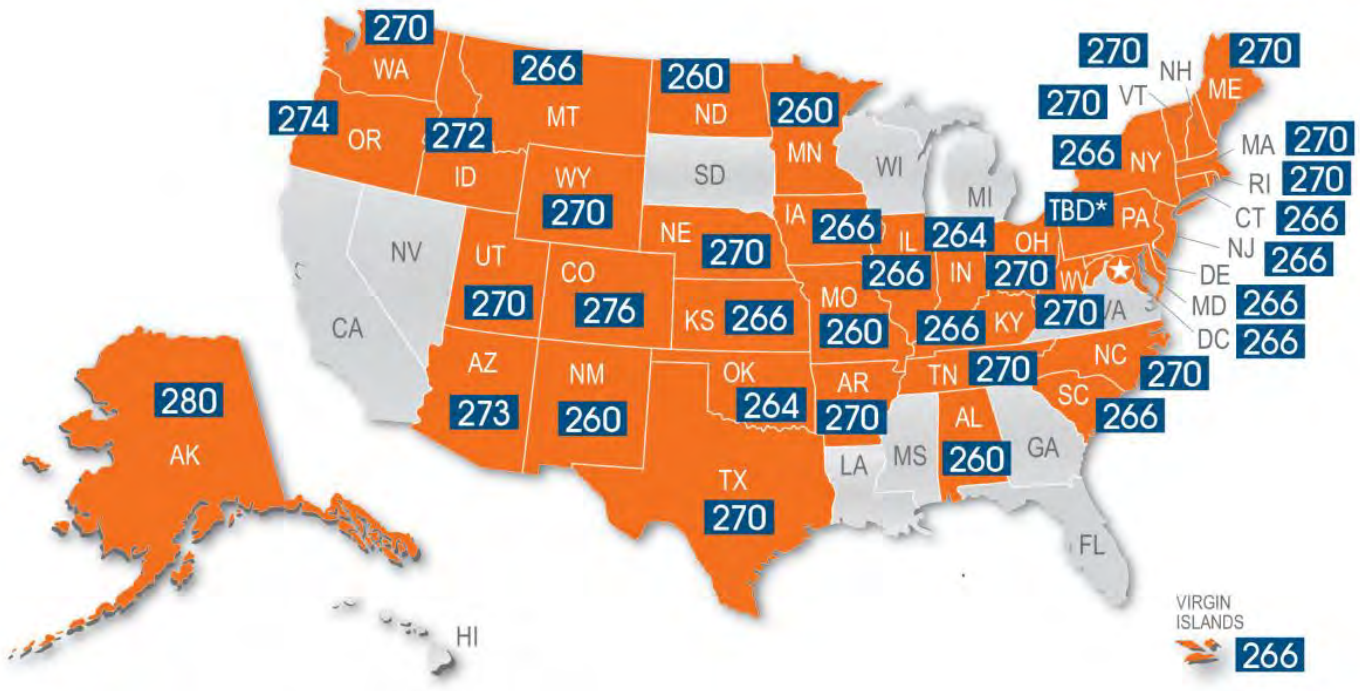
(https://www.ncbex.org/)

NCBE MEDIA KIT (/MEDIA-KIT/) | THE BAR EXAMINER WEBSITE > (HTTPS://THEBAREXAMINER.ORG/)

MINIMUM SCORES

Minimum Passing UBE Score by Jurisdiction

This map shows UBE jurisdictions in orange and lists the minimum passing score for each jurisdiction. The same information is displayed in tabular format below the map. **Note that North Carolina, Oregon, and Washington temporarily lowered their minimum passing scores for the July 2020 exam to 268, 266, and 266, respectively, due to the COVID-19 pandemic. Visit July 2020 Bar Exam: Jurisdiction Information (/ncbe-covid-19-updates/july-2020-bar-exam-jurisdiction-information/) for more information. North Carolina's lowered minimum passing score also applies to the February 2021 exam; Washington's lowered minimum passing score also applies to the February and July 2021 exams.**



Minimum Passing UBE Score*	Jurisdiction

(https://www.ncbex.org/)

260	Alabama, Minnesota, Missouri, New Mexico, North Dakota
264	Indiana, Oklahoma
266	Connecticut, District of Columbia, Illinois, Iowa, Kansas, Kentucky, Maryland, Montana, New Jersey, New York, South Carolina, Virgin Islands
270	Arkansas, Maine, Massachusetts, Nebraska, New Hampshire, North Carolina, Ohio, Rhode Island, Tennessee, Texas, Utah, Vermont, Washington, West Virginia, Wyoming
272	Idaho
273	Arizona
274	Oregon
276	Colorado
280	Alaska

Since jurisdiction rules and policies change, you are strongly advised to consult the jurisdiction's bar admission agency (http://www.ncbex.org/exams/ube/) directly for the most current information.

**The minimum passing score in Pennsylvania has not yet been determined.*