CLEAR
Exam Review

VOLUME XXXIV, NUMBER 2 | FALL 2024

A Journal

CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 108 Wind Haven Drive, Suite A, Nicholasville, KY 40356.

Design and composition of this journal have been underwritten by Prometric, a leading provider of technology-enabled testing and assessment solutions to many of the world's most recognized licensing and certification organizations. Supporting more than 8 million test takers annually at testing locations in 180 countries around the world, Prometric has over three decades of experience working with clients of all sizes across a multitude of industry sectors to develop and deploy innovative test development and test administration solutions that address evolving market needs. Learn more about how to grow your testing program at Prometric.com.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $15 per issue to others. Contact CLEAR (859) 309-4733 or cer@clearhq.org for membership information or to purchase journal issues.

Advertisements for CER may be reserved by contacting CLEAR at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page and cost $100 or $200, respectively, per issue.

CLEAR Exam Review

VOLUME XXXIV, NUMBER 2 | FALL 2024

CONTENTS

 ISSN 1076-8025

From the Editors

SARAH WENNIK, MBA, MS
CYNTHIA WOODLEY, Ed.D.

Sarah Wennik, MBA, MS

Cynthia Woodley, Ed.D.

We hope this letter finds you well and thriving in your professional pursuits. As *CLEAR Exam Review (CER)* continues to evolve, we are excited to share that this edition reflects our ongoing transition from a sole focus on exams to a broader exploration of issues relevant to the field of professional regulation.

In this issue, we feature our regular columns: "Abstracts and Updates," "Legal Beat," and "Recent CLEAR *Quick Poll* Results." Each contributor offers valuable insight into various aspects of assessment and credentialing in practice. This month's "Abstracts and Updates" covers a range of recent publications, including a new edition of *Assessment in Nursing Education,* three recent articles addressing quality assurance and the validity of assessments, and updates on ChatGPT-4 and other AI innovations. Additionally, Dr. George Gray examines four articles on diverse topics, including assessment design to mitigate cheating in MOOCs, a new taxonomy for exam restrictions, and the role of exams in competency-based medical education.

In "Legal Beat," Dale Atkinson discusses a license reciprocity case involving a New York-licensed dentist seeking reciprocal licensure in Pennsylvania. The Pennsylvania Dental Board denied the request, citing concerns that New York's dental licensing requirements were not substantially equivalent to Pennsylvania's. The Commonwealth Court, however, reversed the Board's decision. To learn more about the reasoning behind this outcome, delve into "Legal Beat."

Carla Caro presents the results from the two most recent CLEAR membership surveys in the *Quick Poll* column. The polls highlighted in this issue are the "Use of Jurisprudence Exams" and "Pass Rate Changes from Before Covid-19 to the Present." Interestingly, the "Pass Rate Changes" poll saw one of the lowest response rates in recent years. Was it because the topic has been thoroughly explored and discussed, or was everyone simply enjoying a screen-free vacation in July? We invite you to share your thoughts and suggestions for future *Quick Polls* on the CLEAR Regulatory Network Message Board.

Beyond these insightful standard columns, we are pleased to introduce our featured articles: "Ethics & Disciplinary Programs: Mitigating Risk & Minimizing Exposure" and "Practical Issues in Standard Setting Part I: An Approach for Incorporating Policy Considerations with Method." In the first article, Richard Bar, JD, Dr. Christine D. Niero, Ann Witherspoon, and John Zarian, JD, MFin, present a comprehensive 10-part framework outlining best practices for credentialing bodies navigating ethical challenges. While some of these challenges related to exams, this article broadly addresses ethical concerns for credentialing bodies, in line with *CER's* expanded focus to better engage our regulator readers.

From the Editors

Our second article remains aligned with *CER's* traditional focus on examinations, particularly for those with an interest in psychometrics. If you have a favorite standard-setting method—perhaps the Bookmark method—you'll find this article intriguing. Drs. Gregory Cizek and Lorin Mueller introduce a thoughtful amendment to standard setting processes by incorporating "historical bookmark" data. This approach provides participants with information on where previous pass rates would be located in the ordered item booklet, serving as useful guideposts without imposing limitations. We're excited to announce that the exploration of standard setting will continue in the next issue with a follow-up article from these authors.

Thank you for your continued support and dedication to enhancing the regulatory community. We hope this issue both informs and inspires you in your work. As editors, we take great pride in curating content that speaks to the challenges and opportunities you face every day. Our goal is to offer valuable insights that spark new ideas, deepen understanding, and encourage meaningful conversations within our field.

As we continue to expand our scope and evolve *CLEAR Exam Review,* we want you to know that your feedback is instrumental in shaping each issue. We encourage you to share your thoughts on what resonates with you, the challenges you face, and which topics you would like to see explored in future editions. The regulatory landscape is constantly changing, and we are committed to ensuring that *CER* evolves in tandem, providing a platform for discussions that matter most to you.

We invite you to engage with us and the broader regulatory community on the CLEAR Regulatory Network. Your voice is essential in driving the future of professional regulation, and we look forward to hearing from you as we collectively work to strengthen our shared mission of fostering integrity, fairness, and accountability in regulatory practices.

Warm regards,

Sarah Wennik
Editor, *CLEAR Exam Review*

Cynthia Woodley
Editor, *CLEAR Exam Review*

Abstracts and Updates

GEORGE T. GRAY, Ed.D.
Testing and Measurement Consultant

This issue covers a number of publications: the newly published seventh edition of a book on nursing education (featuring assessment methodology) and a number of other examination- and licensure-related topics. Given the number of articles, frequently only the abstract will be represented with no further summary. In these cases, the first word of the write-up will be ABSTRACT. The source for most of the references comes from Google Scholar. In some instances, the full reference may be reviewed online on Google Scholar or as a link embedded in this document.

Finally, although the topic of ChatGPT, featuring the performance of ChatGPT on examinations, was covered in the most recent issue of *CER,* this topic continues to stimulate additional studies and perspectives. Many articles have been published, including a fairly large number of international studies. Progress in this area is proceeding quickly, notably with the release of ChatGPT4; therefore, several new references to ChatGPT are included in this column.

Assessment in Nursing Education

Oermann, M. H., Gaberson, K. B., & De Gagne, J. C. (2024). *Evaluation and Testing in Nursing Education* (7th ed.). Springer Publishing.

The preface of this book states that its intended purpose is to provide a resource for assessing learning for teachers in nursing education programs and health care agencies. The authors note that, although the examples of test items and other assessments are based in the nursing field, the formats of the illustrations are applicable to other health care fields, as well.

The book has five sections and eighteen chapters. Part 1 illustrates Concepts of Assessment: “Assessment and the Educational Process” and “Qualities of Effective Assessment Procedures: Validity, Reliability, and Usability.” Part 2 has seven chapters:

- Planning for Testing
- True-False and Matching
- Multiple-Choice and Multiple-Response
- Short-Answer (Fill-in-the-Blank) and Essay
- Assessment of Higher Level Learning and Clinical Judgment
- Test Construction and Preparation of Students for Next Generation NCLEX® and Certification Examinations (written by Desireé Hensel)
- Assessment of Written Assignments

Part 3 is titled “Test Construction and Analysis.” It includes “Assembling, Administering, and Scoring Tests,” “Testing and Evaluation in Online Courses,” and “Test and Item Analysis: Interpreting Test Results.”

Part 4 includes “Clinical Evaluation,” “Clinical Evaluation Methods,” and “Simulation and Objective Structured Clinical Examination for Assessment.”

Part 5 is titled “Issues Related to Testing and Evaluation in Nursing Education.” Topics are social, ethical, and legal issues; grading; and program evaluation and accreditation.

Abstracts and Updates

Studies Covering Quality Assurance and Validity of Assessments

Amaral, E., & Norcini, J. (2023). Quality assurance in health professions education: Role of accreditation and licensure. *Medical Education, 57*(1), 40-48. https://doi.org/10.1111/medu.14880

ABSTRACT:

"Objective: The aim of this paper is to provide an overview of the major quality assurance strategies, accreditation and licensure, in health professions education. It explores the nature of these regulatory processes using Brazil and the United States as examples because these large systems are at different ends of the developmental continuum. For each, it describes the tensions that arise, offers a critical synthesis of the evidence and maps out future directions."

"Results: Given wide variability among operating medical schools in curricular design, length of study, resources and facilities for clinical training and supervision, the nature of regulatory bodies varies considerably. Nonetheless, they share tensions related purpose and process including quality assurance versus quality improvement, outcomes versus process and continuous versus episodic evaluations and assessments. Clear evidence of effectiveness, especially for accreditation, is scarce and difficult to obtain, particularly as it relates to health outcomes."

"Conclusions: Regulatory processes need to be built around clear definitions of the goals for each stage of professional development, the current movement towards competency-based education and the variable durations of medical education. These changes must motivate revisions in the content and process of programmes for accreditation and licensure, complimentary efforts towards quality of care, and stimulate a significant research effort."

Norcini, J., Grabovsky, I., Barone, M. A., Anderson, M. B., Pandian, R. S., & Mechaber, A. J. (2024). The associations between United States Medical Licensing Examination performance and outcomes of patient care. *Academic Medicine, 99*(3), 325-330. https://doi.org/10.1097/acm.0000000000005480

ABSTRACT:

"Purpose: The United States Medical Licensing Examination (USMLE) comprises a series of assessments required for the licensure of U.S. MD-trained graduates as well as those who are trained internationally. Demonstration of a relationship between these examinations and outcomes of care is desirable for a process seeking to provide patients with safe and effective health care."

"Method: This was a retrospective cohort study of 196,881 hospitalizations in Pennsylvania over a 3-year period… for 5 primary diagnoses: heart failure, acute myocardial infarction, stroke, pneumonia, or chronic obstructive pulmonary disease. The 1,765 attending physicians for these hospitalizations self-identified as family physicians or general internists. A converted score based on USMLE Step 1, Step 2 Clinical Knowledge, and Step 3 scores was available, and the outcome measures were in-hospital mortality and log length of stay (LOS). The research team controlled for characteristics of patients, hospitals, and physicians."

"Results: For in-hospital mortality, the adjusted odds ratio was 0.94 (95% confidence interval [CI] = 0.90, 0.99; $P <$ .02). Each standard deviation increase in the converted score was associated with a 5.51% reduction in the odds of in-hospital mortality. For log LOS, the adjusted estimate was 0.99 (95% CI = 0.98, 0.99; $P <$ .001). Each standard deviation increase in the converted score was associated with a 1.34% reduction in log LOS."

Abstracts and Updates

"Conclusions: Better provider USMLE performance was associated with lower in-hospital mortality and shorter log LOS for patients, although the magnitude of the latter is unlikely to be of practical significance. These findings add to the body of evidence that examines the validity of the USMLE licensure program."

Reisdorff, E. J., Joldersma, K. B., Kraus, C. K., Barton, M. A., Knapp, B. J., Kupas, D. F., Clemency, B. M., & Daya, M. (2024). Internal validity evidence for the American Board of Emergency Medicine Emergency Medical Services Certification Examination. *Prehospital Emergency Care.* https://doi.org/10.1080/10903127.2024.2379872

ABSTRACT:

"Objectives: The American Board of Emergency Medicine (ABEM) Emergency Medical Services Medicine (EMS) subspecialty was approved by the American Board of Medical Specialties on September 23, 2010. Subspecialty certification in EMS was contingent on two key elements—completing Accreditation Council for Graduate Medical Education (ACGME)-accredited EMS training and passing the subspecialty certification examination developed by ABEM. ... Meaningful certification requires rigorous assessment. ...(T)he EMS certification examination sought to embrace the tenets of validity, reliability, and fairness. For… this report, the sources of validity evidence were anchored on the EMS core content, the examination development process, and the association between fellowship training and passing the certification examination."

"Methods: We chose to use validity evidence that included: 1) content validity (based on the EMS core content); 2) response processes (test items required intended cognitive processes); 3) internal structure supported by the internal relationships among items; 4) relations to other variables, specifically the association between examination performance and ACGME-accredited fellowship training; and 5) the consequences of testing."

"Results: There is strong content validity evidence for the EMS examination based on the core content and its detailed development process. The core content and supporting job-task analysis was also used to define the examination blueprint. Internal structure support was evidenced by Cronbach's coefficient alpha, which ranged from 0.82 to 0.92. Physicians who completed ACGME-accredited EMS fellowship training were more likely to pass the EMS certification examination (chi square, $p < 0.0001$; Cramér's, $V = 0.24$). Finally, there were two sources of consequential validity evidence—use of test results to determine certification and use of the resulting certificate."

"Conclusions: There is substantial and varied validity evidence to support the use of the EMS certifying examination in making summative decisions to award certification in EMS. Of note, there was a statistically significant association between ACGME-accredited fellowship training and passing the examination."

ChatGPT-4 and Other AI Updates

Katz, M. K., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions Royal Society A, 382*(2270). https://doi.org/10.1098/rsta.2023.0254

ABSTRACT:

"In this paper, we experimentally evaluate the zero-shot performance of GPT-4 against prior generations of GPT on the entire uniform bar examination (UBE), including not only the multiple-choice multistate bar examination (MBE), but also the open-ended multistate essay exam (MEE) and multistate performance test (MPT) components. On the MBE, GPT-4 significantly outperforms both human test-takers and prior models, demonstrating a 26% increase over ChatGPT and beating humans in five of seven subject areas. On the MEE and MPT, which have not previously

been evaluated by scholars, GPT-4 scores an average of 4.2/6.0 when compared with much lower scores for ChatGPT. Graded across the UBE components, in the manner in which a human test-taker would be, GPT-4 scores approximately 297 points, significantly in excess of the passing threshold for all UBE jurisdictions. These findings document not just the rapid and remarkable advance of large language model performance generally, but also the potential for such models to support the delivery of legal services in society."

Angel, M. C., Rinehart, J. B., Canneson, M. P., & Baldi, P. (2024). Clinical knowledge and reasoning abilities of AI large language models in anesthesiology: a comparative study on the American Board of Anesthesiology Exam. *Anesthesia and Analgesia, 139*(2), 349-356. https://doi.org/10.1213/ane.0000000000006892

ABSTRACT:

"Background: Over the past decade, artificial intelligence (AI) has expanded significantly with increased adoption across various industries, including medicine. Recently, AI-based large language models such as Generative Pretrained Transformer-3 (GPT-3), Bard, and Generative Pretrained Transformer-[4] (GPT-4) have demonstrated remarkable language capabilities. While previous studies have explored their potential in general medical knowledge tasks, here we assess their clinical knowledge and reasoning abilities in a specialized medical context.

"Methods: We studied and compared the performance of all 3 models on both the written and oral portions of the comprehensive and challenging American Board of Anesthesiology (ABA) examination, which evaluates candidates' knowledge and competence in anesthesia practice."

"Results: Our results reveal that only GPT-4 successfully passed the written examination, achieving an accuracy of 78% on the basic section and 80% on the advanced section. In comparison, the less recent or smaller GPT-3 and Bard models scored 58% and 47% on the basic examination, and 50% and 46% on the advanced examination, respectively. Consequently, only GPT-4 was evaluated in the oral examination, with examiners concluding that it had a reasonable possibility of passing the structured oral examination. Additionally, we observe that these models exhibit varying degrees of proficiency across distinct topics, which could serve as an indicator of the relative quality of information contained in the corresponding training datasets. This may also act as a predictor for determining which anesthesiology subspecialty is most likely to witness the earliest integration with AI."

Giannos, P. (2023). Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurology Open, 5*(1). https://doi.org/10.1136/bmjno-2023-000451

ABSTRACT:

"Background: Large language models such as ChatGPT have demonstrated potential as innovative tools for medical education and practice, with studies showing their ability to perform at or near the passing threshold in general medical examinations and standardised admission tests. However, no studies have assessed their performance in the UK medical education context, particularly at a specialty level, and specifically in the field of neurology and neuroscience."

"Methods: We evaluated the performance of ChatGPT in higher specialty training for neurology and neuroscience using 69 questions from the Pool—Specialty Certificate Examination (SCE) Neurology Web Questions bank. The dataset primarily focused on neurology (80%). The questions spanned subtopics such as symptoms and signs,

diagnosis, interpretation and management with some questions addressing specific patient populations. The performance of ChatGPT 3.5 Legacy, ChatGPT 3.5 Default and ChatGPT-4 models was evaluated and compared."

"Results: ChatGPT 3.5 Legacy and ChatGPT 3.5 Default displayed overall accuracies of 42% and 57%, respectively, falling short of the passing threshold of 58% for the 2022 SCE neurology examination. ChatGPT-4, on the other hand, achieved the highest accuracy of 64%, surpassing the passing threshold and outperforming its predecessors across disciplines and subtopics."

"Conclusions: The advancements in ChatGPT-4's performance compared with its predecessors demonstrate the potential for artificial intelligence (AI) models in specialised medical education and practice. However, our findings also highlight the need for ongoing development and collaboration between AI developers and medical experts to ensure the models' relevance and reliability in the rapidly evolving field of medicine."

Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., & Wood, D. A. (2024). Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries. *Review of Accounting Studies, 29*(3), 2318-2349. https://doi.org/10.1007/s11142-024-09833-9

ABSTRACT:

"ChatGPT frequently appears in the media, with many predicting significant disruptions, especially in the fields of accounting and auditing. Yet research has demonstrated relatively poor performance of ChatGPT on student assessment questions. We extend this research to examine whether more recent ChatGPT models and capabilities can pass major accounting certification exams including the Certified Public Accountant (CPA), Certified Management Accountant (CMA), Certified Internal Auditor (CIA), and Enrolled Agent (EA) certification exams. We find that the ChatGPT 3.5 model cannot pass any exam (average score across all assessments of 53.1%). However, with additional enhancements, ChatGPT can pass all sections of each tested exam: moving to the ChatGPT 4 model improved scores by an average of 16.5%, providing 10-shot training improved scores an additional 6.6%, and allowing the model to use reasoning and acting (e.g., allow ChatGPT to use a calculator and other resources) improved scores an additional 8.9%. After all these improvements, ChatGPT passed all exams with an average score of 85.1%. This high performance indicates that ChatGPT has sufficient capabilities to disrupt the accounting and auditing industries, which we discuss in detail. This research provides practical insights for accounting professionals, investors, and stakeholders on how to adapt and mitigate the potential harms of this technology in accounting and auditing firms."

One of the paragraphs in the "Conclusions" section of the article provides more insight into the authors' perspectives on ChatGPT and the accounting industry:

"The results of our study demonstrate that ChatGPT can perform sufficiently well to pass important accounting certifications. This calls into question some of the competitive advantages of the human accountant relative to the machine. To our knowledge, for the first time, AI has performed as well as the majority of human accountants on real-world accounting tasks. This raises important questions about how the machine and accountant will cooperate in the future. We encourage research to help understand where machine and human abilities are best deployed in accounting. We also encourage research that advances the capabilities for machines to perform more accounting work—freeing accountants to innovate and add greater value to their organizations and society."

Abstracts and Updates

Other Assessment Topics

Alexandron, G., Wiltrout, M. E., Berg, A., Gershon, S. K., & Ruipérez-Valiente, J. A. (2022). The effects of assessment design on academic dishonesty, learner engagement and certification rates in MOOCs. *Journal of Computer Assisted Learning, 39,* 141-153. https://doi.org/10.1111/jcal.12733

ABSTRACT:

"Background: Massive Open Online Courses (MOOCs) have touted the idea of democratizing education, but soon enough, this utopian idea collided with the reality of finding sustainable business models. In addition, the promise of harnessing interactive and social web technologies to promote meaningful learning was only partially successful. And finally, studies demonstrated that many learners exploit the anonymity and feedback to earn certificates unethically. Thus, establishing MOOC pedagogical models that balance open access, meaningful learning, and trustworthy assessment remains a challenge that is crucial for the field to achieve its goals."

"Objectives: This study analysed the influence of [a] MOOC assessment model, denoted the Competency Exam (CE), on learner engagement, the level of cheating, and certification rates. At its core, this model separates learning from for-credit assessment, and it was introduced by the MITx Biology course team in 2016."

"Methods: We applied a learning analytics methodology to the clickstream data of the verified learners (*N* = 559) from four consecutive runs of an Introductory Biology MOOC offered through edX. The analysis used novel algorithms for measuring the level of cheating and learner engagement, which were developed in the previous studies."

"Results and Conclusions: On the positive side, the CE model reduced cheating and did not reduce learner engagement with the main learning materials – videos and formative assessment items. On the negative side, it led to procrastination, and certification rates were lower."

"Implications: First, the results shed light on the fundamental connection between incentive design and learner behaviour. Second, the CE provides MOOC designers with an 'analytically verified' model to reduce cheating without compromising on open access. Third, our methodology provides a novel means for measuring cheating and learner engagement in MOOCs."

Dawson, P., Nicola-Richmond, K., & Partridge, H. (2024). Beyond open book versus closed book: A taxonomy of restrictions in online examinations. *Assessment & Evaluation in Higher Education, 49(*2), 262-274. https://doi.org/10.1080/02602938.2023.2209298

ABSTRACT:

"Educators set restrictions in examinations to enable them to assess learning outcomes under particular conditions. The open book versus closed book binary is an example of the sorts of restrictions examiners have traditionally set. In the late 2000s this was expanded to a trinary to include open web examinations. However, the current technology environment, particularly for online examinations, makes this trinary not particularly useful. The web now includes generative artificial intelligence tools, and contract cheating sites, both of which are capable of completing examination questions within the examination period. Closed book, open book and open web no longer offers enough clarity or specificity when communicating examination restrictions. This article proposes a new taxonomy of restrictions for examinations, with a particular focus on online examinations. The taxonomy consists of three dimensions: information, people and tools. ... Five criteria are provided to help examination designers in selecting

restrictions: the learning outcomes being assessed; the feasibility of restrictions; consequential validity; authenticity; and values. Taken together, this taxonomy and the criteria provide ways of thinking about restrictions in examinations that can prompt educators towards examination designs that are more valid and robust against cheating."

Bhanji, F., Naik, V., Skoll, A., Pittini, R., Daniels, V. J., Bacchus, C. M., & Bandiera, G. (2024). Competence by design: The role of high-stakes examinations in a competency-based medical education system. *Perspectives on Medical Education, 13*(1), 68-74. https://doi.org/10.5334/pme.965

ABSTRACT:

"Competency based medical education is developed utilizing a program of assessment that ideally supports learners to reflect on their knowledge and skills, allows them to exercise a growth mindset that prepares them for coaching and eventual lifelong learning, and can support important progression and certification decisions. Examinations can serve as an important anchor to that program of assessment, particularly when considering their strength as an independent, third-party assessment with evidence that they can predict future physician performance and patient outcomes. This paper describes the aims of the Royal College of Physicians and Surgeons of Canada's ("the Royal College") certification examinations, their future role, and how they relate to the Competence by Design model, particularly as the culture of workplace assessment and the evidence for validity evolves. For example, high-stakes examinations are stressful to candidates and focus learners on exam preparation rather than clinical learning opportunities, particularly when they should be developing greater autonomy. In response, the Royal College moved the written examination earlier in training and created an exam quality review, by a specialist uninvolved in development, to review the exam for clarity and relevance. ..."

Gorney, K., Sinharay, S., & Liu, X. (2023). Using item scores and response times in person-fit assessment. *British Journal of Mathematical and Statistical Psychology, 77*(1), 151-168. https://doi.org/10.1111/bmsp.12320

When I am looking for references to include in this column, the name Sandip Sinharay always catches my eye; so, in closing, here is an article for the readers who are psychometricians and may not have seen this publication.

ABSTRACT:

"The use of joint models for item scores and response times is becoming increasingly popular in educational and psychological testing. In this paper, we propose two new person-fit statistics for such models in order to detect aberrant behaviour. The first statistic is computed by combining two existing person-fit statistics: one for the item scores, and one for the item response times. The second statistic is computed directly using the likelihood function of the joint model. Using detailed simulations, we show that the empirical null distributions of the new statistics are very close to the theoretical null distributions, and that the new statistics tend to be more powerful than several existing statistics for item scores and/or response times. A real data example is also provided using data from a licensure examination."

Legal Beat

Do Manikins Have Cheeks?

DALE J. ATKINSON, ESQ.
Managing Member, Atkinson & Atkinson, LLC
http://www.atkinsonfirm.com/home

In most professions, several factors are prerequisites to initial licensure eligibility. These factors include applications, fees, education, experience, examination, and good moral character. As regulatory boards assess applications for licensure of persons already licensed in another jurisdiction, certain factors may be recognized to have been met to allow for efficient eligibility determinations. The basis behind interstate recognition of eligibility factors from state to state is premised upon uniform standards through, for example, organizations that accredit education programs. Furthermore, uniform competence assessments (licensure examinations) are ubiquitous among the professions that require licensure. This interstate recognition of licensure examinations is generally due to the credibility of the developing organization and the validity of the assessment mechanism. Statutes, or more appropriately, rules and regulations, are enacted that recognize education and/or examinations across state lines to contribute to the efficiencies of the licensure of professionals.

This process for licensing applicants who are already licensed in another jurisdiction is referred to as either reciprocity or endorsement. The use of interstate compacts provides another process of recognition and may be the subject of a future article. Reciprocity generally refers to an agreement between jurisdictional entities that essentially agree, "You take ours and we will take yours." More common are endorsement statutes that allow one state to license an applicant who is already licensed in another state if the requirements for licensure in the initial jurisdiction are substantially similar to those of the second jurisdiction. Endorsement statutes allow for some objectivity in determining "substantially equivalent." Consider the following.

In 2017, an applicant (referred to as Applicant) was a licensed dentist in the state of New York. In that same year, the Applicant briefly practiced in Addison, New York, and later purchased a practice from a retiring dentist. That purchased practice had offices in both Watkins Glen, New York, and Elkland, Pennsylvania. The Applicant, who was only licensed in New York, sought licensure in Pennsylvania after acquiring the Elkland office. On May 4, 2021, he filed an application for licensure by endorsement to the State Board of Dentistry of Pennsylvania (Board).

The Pennsylvania endorsement statute stated that a licensing board "shall issue" a license to an applicant if that applicant "holds a current license … from another state, territory or country and the licensing board or licensing commission determines that the state's, territory's or country's requirements are substantially equivalent to or exceed the requirements established in this Commonwealth." By letter dated July 27, 2021, the Board denied the Applicant's application for licensure. It held that New York's dental licensing requirements were not substantially equivalent to Pennsylvania's requirements. The basis behind this decision was that New York did not require dental applicants to pass a clinical examination. The Board determined that the clinical examination was integral to the licensure process in Pennsylvania.

The Applicant filed an appeal of his denial of licensure application. He argued that while New York did not require passage of a clinical examination, it did require applicants to successfully complete a "clinically-based postdoctoral general practice or specialty dental residency program, of at least one year's duration as well as

a 'formal outcome assessment evaluation of the resident's competence to practice dentistry.'" He argued that the New York prerequisites to licensure were at least substantially equivalent to a clinical examination. The appeal was delegated by the Board to a Hearing Officer. A formal hearing was held on October 4, 2021, and the Applicant was the only witness to testify. The Applicant testified to the value of the New York residency requirement based upon his personal experiences. He also testified as to the difficulties he would encounter if he sat for the clinical examination required in the Commonwealth.

Specifically, the Applicant testified to the completion of his residency program at a medical center in Albany, New York. New York law requires such residency programs to require completion of identified procedures, including "two full crowns; two endodontically treated teeth; four restorations, meaning two anterior, two posterior; and one periodontal case." The Applicant contrasted these minimum requirements to the procedures he actually performed during his residency. They included 21 crown preparations, 14 endodontic procedures, 197 restorations, and 18 periodontal cases. He emphasized the breadth of the New York residency content compared to the examination recognized by the Commonwealth, that being the agency of the American Board of Dental Examiners (ADEX). ADEX is the entity that develops examinations used to assess entry-level competence of applicants for licensure as dentists.

The ADEX examination is a two- or three-day test that involves "only" three crown preparations, two restorations, one complete root canal, and one partial root canal. The New York residency programs occur over a one-year time frame. The Applicant's testimony also included the anticipated difficulties of taking the ADEX, where he would be responsible for locating and transporting live patients to the test site. Furthermore, the ADEX has limited spaces for dentists, with over 80% of examination opportunities reserved for dental students and the remaining space for currently licensed dentists. Finally, the Applicant identified that many of the procedures performed on the ADEX involved the use of manikins with plastic teeth rather than live human subjects. These manikins often do not have cheeks, changing the procedure(s) and competence assessments and conclusions.

On December 20, 2021, the Hearing Officer issued an opinion concluding that the New York dental licensing requirements equaled or exceeded those of Pennsylvania. The Hearing Officer found the Applicant's testimony that a residency program is superior to a clinical examination to be persuasive and noted the extensive requirements of the New York dental residency program and the use of live patients. In addition, residency programs involve the development and use of procedures over a period of time that allows for professional maturity. The Hearing Officer also referred to the legislative intent of opening up the licensure process and remedying the time-consuming procedures that may discourage out-of-state professionals from coming to Pennsylvania. On December 27, 2021, the Board issued its notice of intent to review the findings of the Hearing Officer.

On March 16, 2022, the Board issued its final adjudication, holding that the New York licensure requirements for dental applicants were not substantially equivalent to those of Pennsylvania. It found that many of the New York criteria were indeed equivalent, but the written examination and residency were not substantially equivalent to the Pennsylvania written and dental clinical examinations. The Board noted that "the clinical experience of every resident is different, and competency is measured by the subjective opinion of the program director or attending dentist," but the dental clinical examination is based upon objectivity and a requirement that all examinees demonstrate minimum competence. It concluded that Pennsylvania's clinical examination assessed an additional measurement of competency otherwise absent in New York.

The Board rejected the individual experiences and number of procedures performed as a basis for equivalence and interpreted the applicable Pennsylvania statute as permitting a comparison of the text of the New York law with the Pennsylvania law. Finally, the Board rejected the public policy considerations of the Hearing Officer and, ultimately, denied the application for licensure by endorsement. The Applicant appealed this ruling to the Commonwealth Court.

The Commonwealth Court first identified the standard of review as whether substantial evidence supports the ruling of the Board. The Court will only disturb the ruling of the Board if there was an abuse of discretion, the Board exceeded its authority, or the Board misapplied the law. It found no previous judicial decisions interpreting the statute in question.

Turning to the facts of the current matter, the Court identified that New York law required dental applicants to pass a written examination and successfully complete a "clinically-based postdoctoral general practice or specialty dental program, of at least one year's duration in a hospital or dental facility accredited for teaching purposes … ." These residencies must provide a formal outcome assessment addressing competence, including a notarized statement(s) by the director and/or attending dentist attesting to the successful completion of, at minimum, certain identified procedures.

The Pennsylvania statute delegates authority to the Board to license applicants "after examination," and the promulgated regulations provide that applicants "shall pass the National Board Dental Examination (written examination) and the dental clinical examination administered by one of five testing agencies." As noted by the Court, neither the statute nor regulations specify the requirements of any dental clinical examination. The Court next cites the endorsement statute that addresses substantial equivalence and gives the Board discretion in making such determinations. Substantial equivalence requires the Board to make this determination by viewing the criteria "as a whole" when deciding whether New York law is comparable to or exceeds the education, examination, and experience required by Pennsylvania law.

The Court found that the New York licensing requirements were substantially equivalent to or exceeded the Pennsylvania requirements. It held that the Pennsylvania clinical examination and the New York residency program served the same purpose and were functionally interchangeable. In fact, and through amendments to the statute initiated in 2002, New York eliminated its clinical examination requirement in 2006 and replaced it with the residency program. The Court held that the Board failed to undertake a plain-language reading of Pennsylvania law when making its decision. The Board suggested that "no residency program, no matter how stringent, could substitute for a clinical examination." This narrow and conclusory view ignores the authority of the Board to determine an applicant's competence separate and apart from substantial equivalency.

While the Court agreed with the Board that it consider the requirements of the applicant's state rather than the applicant's individual experiences, the Court noted that the Board violated its own rule. It did so by looking beyond the Pennsylvania law to justify the decision on non-equivalence. The "critical qualities of a dental clinical examination on which the Board relied, grading criteria and objectivity, do not appear in Pennsylvania's licensing statute and regulations." The law does not specify the grading criteria, nor does it reference controls to address objectivity. "Just as New York delegates competency assessments to residency programs, Pennsylvania delegates those assessments to testing agencies."

The Court found no support for the conclusions of the Board under the applicable law and the factual circumstances of the current matter. Thus, it reversed the Board's March 16, 2022, letter denying the Applicant's licensure and remanded the matter to the Board for further consideration, taking into account this opinion.

This case represents an important judicial decision addressing both equivalence under an endorsement statute as well as statutory construction. The absence of statutory or regulatory language identifying the need for objectivity and certain grading standards opened the door to a conclusion that a post-doctorate residency program may be an equivalent means of assessing competence.

Reference

Haentges v. State Board of Dentistry, 2023 Pa. Commw. LEXIS 216

Recent CLEAR Quick Poll Results

CARLA M. CARO, MA, ICE-CCP.
Program Director, Credentialing & Career Services
ACT

The CLEAR Examination Resources and Advisory Committee (ERAC) periodically issues *Quick Poll* surveys. These *Quick Polls* are not designed as scientific studies but rather are intended to gather snapshot information regarding current issues within the regulatory community. This article discusses the results of two recent *Quick Polls* administered in May and July of 2024.

Use of Jurisprudence Exams (May 2024)

Questions

- Does your organization include a jurisprudence exam as part of the licensure/registration/certification process?
  - ☐ Yes
  - ☐ No *[Those selecting this option were exited out of the survey.]*
- What type of jurisprudence exam does your organization use?

	Open book	Closed book
Formative Provides feedback and encourages ongoing learning		
Summative Assesses knowledge level against a standard		

Select all that apply for each type of exam listed at the top of the two columns.

- How often are your licensees/registrants/certificants required to take the jurisprudence exam?
  Select all that apply.
  - ☐ Upon initial license/registration/certification
  - ☐ Upon renewal
- How often is the jurisprudence exam updated?
  - ☐ Regularly *(Specify time period of update.)*
  - ☐ As needed
  - ☐ Never
- Compared to your expectations as a regulator, what is your evaluation of the cadence for updating the local jurisprudence examination?
  - ☐ Just right
  - ☐ Could be better

Number of responses: 194

Recent CLEAR Quick Poll Results

Results

One hundred ninety-four respondents answered the poll. Of these, 60% (n = 116) indicated that they include jurisprudence exams as part of licensure, registration, and/or certification, and 40% (n = 78) indicated that they do not include jurisprudence exams.

Of the 116 respondents who use jurisprudence exams, between 76 and 78 respondents answered a series of follow-up questions. First, respondents were asked about the types of jurisprudence exams they offer; multiple responses were permitted. The most common approach to jurisprudence exams (34% of respondents to this question, n = 26) was to require a closed-book summative exam that assesses examinees' knowledge level against a standard. Twenty-eight percent (n = 21) used summative open-book exams. A further 16% (n = 12) offered both open-book formative exams and an open-book summative exam to provide feedback and encourage ongoing learning. Other models included using both formative and summative closed-book exams (12%, n = 9), using formative open-book exams only (8%, n = 6), and using formative open-book and summative closed-book exams (3%, n = 2).

Figure 1. Organization Includes a Jurisprudence Exam as Part of the Licensure/Registration/ Certification Process

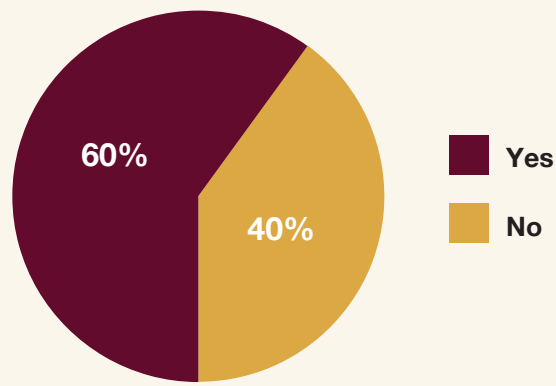
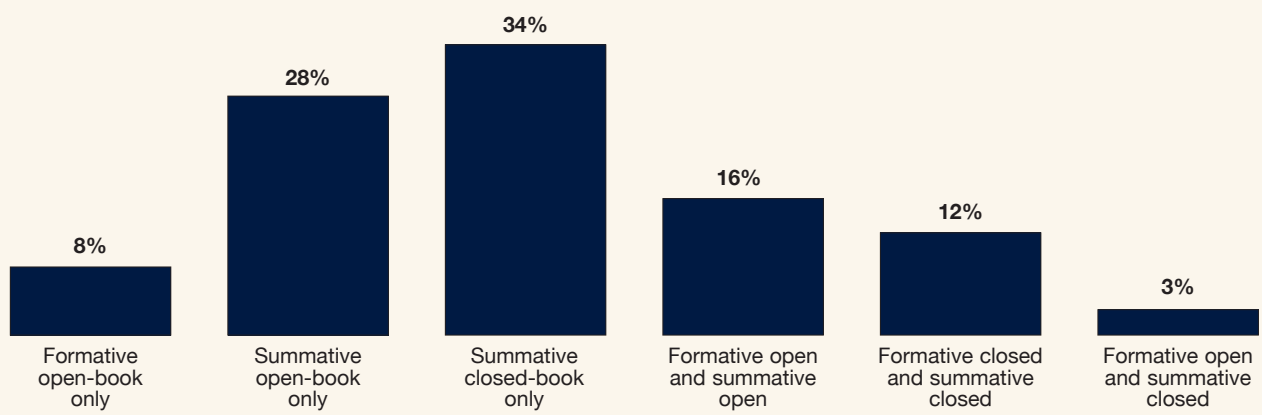


Figure 2. Type(s) of Jurisprudence Exam Given



Recent CLEAR Quick Poll Results

Respondents were asked to identify when licensees, registrants, or certificants were required to take the jurisprudence exam; respondents could select all options that applied. Almost all respondents (99%) required the jurisprudence exam upon initial licensure, registration, or certification; 12% required the jurisprudence exam upon renewal, and 14% indicated some other time when they required the jurisprudence exam. Write-in responses included the exam being required pre-licensure, upon reinstatement after different numbers of years in inactive status, when seeking recognition in a new or different jurisdiction, or as part of a cycle of continuing education.

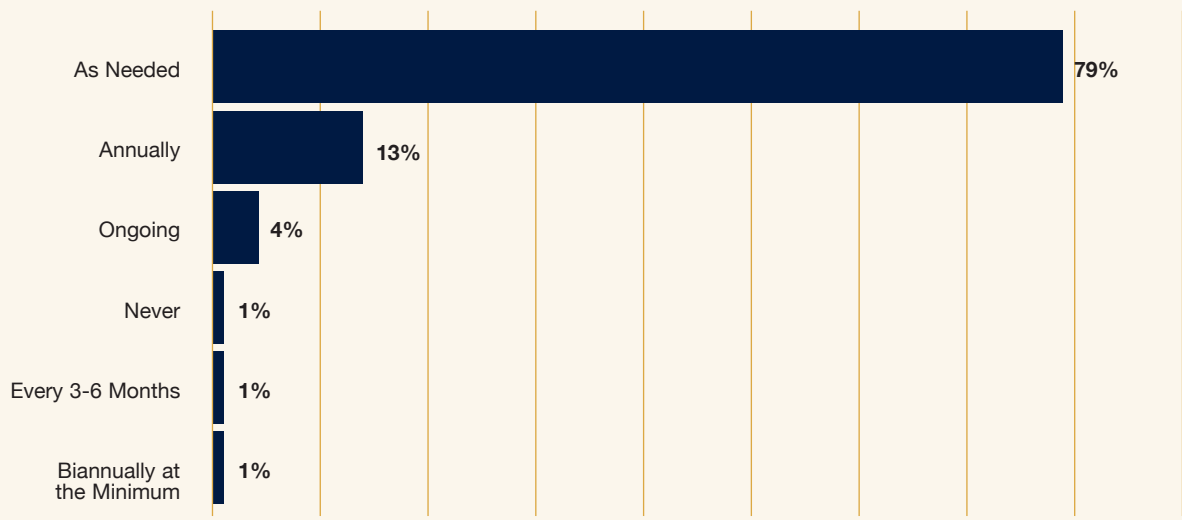
Table 1. Jurisprudence Exam Frequency Requirements

Response option	Number of respondents	Percent
Upon initial license/registration/certification	77	99%
Upon renewal	9	12%
Other	11	14%

Note. N = 78. Multiple responses permitted—total sums to more than 100%.

Seventy-six respondents answered the question about how frequently they updated their jurisprudence exams. These exams are most typically updated on an "As Needed" basis (79%, n = 60). Reasons for updating as needed included when there are changes in laws, rules, statutes, policies, or bylaws and after psychometric review flagging poorly performing items. Ten respondents (13%) indicated their exams are updated regularly on an annual basis, with new questions or topics added. Updates may be followed by standard setting for a new cut score. Smaller numbers of respondents indicated they updated the exam on an ongoing basis (n = 3), biannually (n = 1), or never (n = 1).

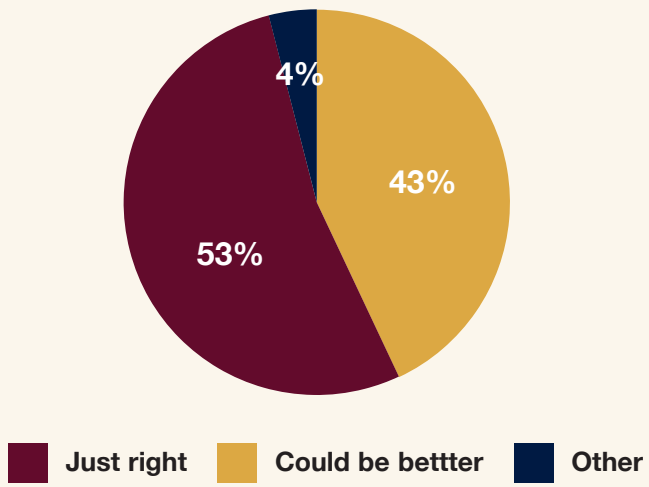
Figure 3. Frequency of Jurisprudence Exam Updates



Recent CLEAR Quick Poll Results

Respondents were also asked, “Compared to your expectations as a regulator, what is your evaluation of the cadence for updating the local jurisprudence examination?” Seventy-eight individuals answered this question. Just over half (53%, n = 41) indicated the cadence was just right; 44% (n = 34) indicated that the cadence could be better; and 4% (n = 3) wrote in some other response. (Note: Percentages do not sum to 100% due to rounding.) Write-in responses mentioned the difficulty of writing exam questions related to jurisprudence, challenges with keeping up to date with changes in legislation and regulation, and descriptions of how regulatory updates are provided to their licensees.

Figure 4. Satisfaction with Cadence of Jurisprudence Exam Updates



Recent CLEAR Quick Poll Results

Pass Rates Changes from Before Covid-19 to the Present (July 2024)

Questions

Introduction: This Quick Poll is exploring how pass rates may have changed from before the Covid-19 pandemic to the height of the pandemic, and from before the pandemic to the current time. Responses will help illuminate whether any changes during the pandemic have become the "new normal."

If your board or organization offers more than one licensure or certification exam, answer for the program with the largest number of candidates in 2019.

- Indicate if pass rates for the exam changed from before the pandemic (that is, from 2019) to the height of the Covid-19 pandemic. For purposes of this survey, the height of the pandemic is defined as **March 2020 through September 2022.**

  Compared to 2019, did the pass rate change **during the height of the Covid-19 pandemic?**

  - ☐ Decreased/Pass rate was lower
  - ☐ Increased/Pass rate was higher
  - ☐ Pass rate stayed the same
  - ☐ Not applicable

- Indicate if pass rates have changed from before the pandemic to the current time (that is, your most recent administration).

  Compared to 2019, has the **most recent pass rate** changed?

  - ☐ Decreased/Pass rate is lower
  - ☐ Increased/Pass rate is higher
  - ☐ Pass rate is the same
  - ☐ Not applicable

- To what do you attribute any changes, if applicable? *Select all that apply.*

  - ☐ Changes in exam content and specifications
  - ☐ Changes in exam delivery method
  - ☐ Changes in candidate population, including growing or shrinking numbers
  - ☐ Changes in eligibility requirements
  - ☐ Changes in candidate preparation
  - ☐ Changes in candidate access to the exam related to geographic borders
  - ☐ Not applicable, no change in pass rate
  - ☐ Other *(Please specify.)*

Recent CLEAR Quick Poll Results

- What is your role/what type of organization do you represent?
  - ☐ Regulator of a single licensing board or organization
  - ☐ Umbrella organization regulating multiple professions
  - ☐ Certification body, organization, or board
  - ☐ Consultant, testing company, or vendor
  - ☐ Other *(Please specify.)*
- Where are you located?
  - ☐ Canada
  - ☐ United States
  - ☐ Europe *(Specify country.)*
  - ☐ Australia
  - ☐ New Zealand
  - ☐ Other *(Specify country.)*

Number of responses: 27

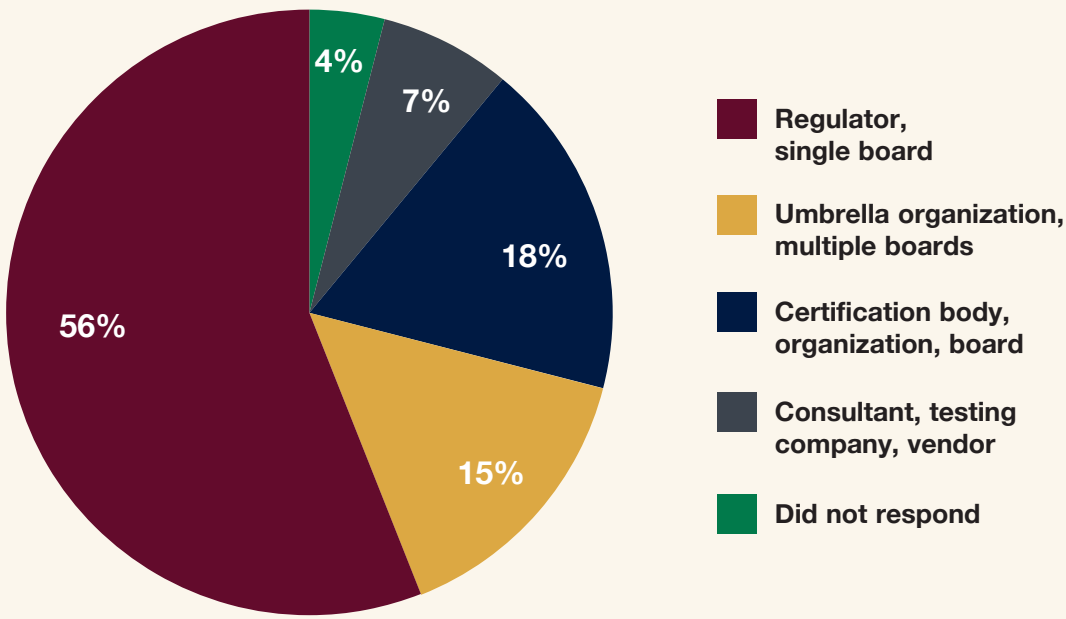
Results

Only 27 respondents completed this *Quick Poll*—one of the smallest numbers in recent years. This begs two questions: First, has the topic of pass rates and the pandemic been explored to its fullest, and are regulatory boards and the credentialing community no longer interested in or engaged with this issue? Second, did administering a *Quick Poll* in July, when many individuals may be on summer break, contribute to the low number of responses?

Of the respondents, more than half (56%, n = 15) were regulators from a single licensing board or organization. Five (18%) were members of a certification body, organization, or board. Four (15%) were from an umbrella organization or central agency regulating multiple boards or professions; two (7%) were consultants, testing companies, or vendors; and one did not respond.

Recent CLEAR Quick Poll Results

Figure 5. Pass Rate *Quick Poll* Respondent Affiliation



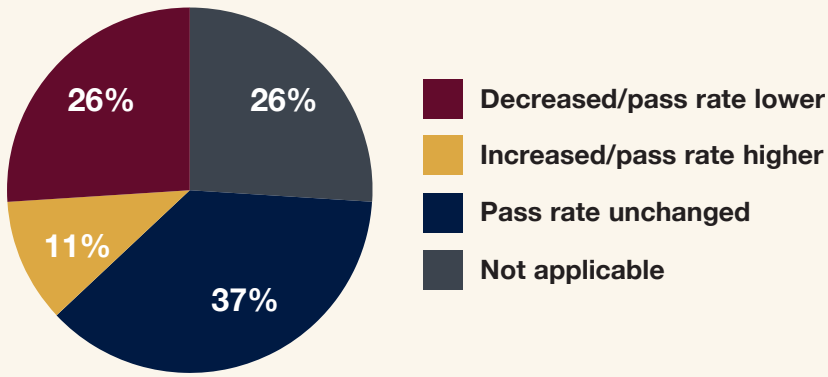
All respondents were from North America, with 14 from the United States and 12 from Canada.

Table 2. Regional Distribution of Respondents

Region	Number of respondents	Percent
United States	14	51.9%
Canada	12	44.4%
Did not respond	1	3.7%
Total	**27**	**100.0%**

Recent CLEAR Quick Poll Results

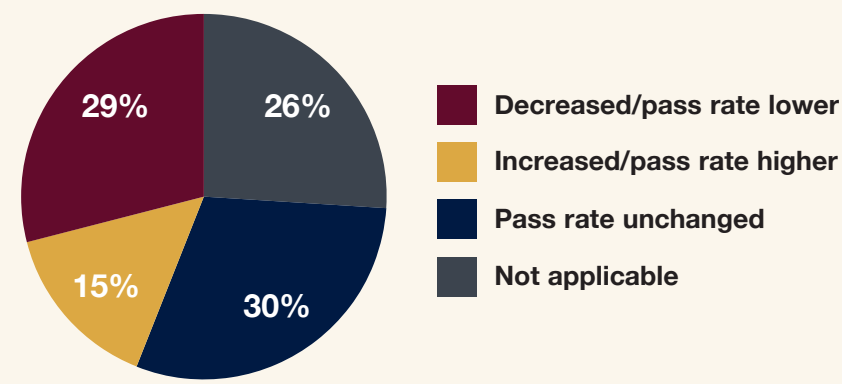
Figure 6. Pass Rate Change at the Height of the Pandemic



Pass rates changed in a variety of ways from before the pandemic to the height of the pandemic (defined as March 2020 through September 2022 for purposes of the poll). Just over one-fourth indicated that the pass rate decreased, 11% indicated that the pass rate increased, and 37% indicated that the pass rate remained unchanged. A further 26% indicated that this did not apply to them.

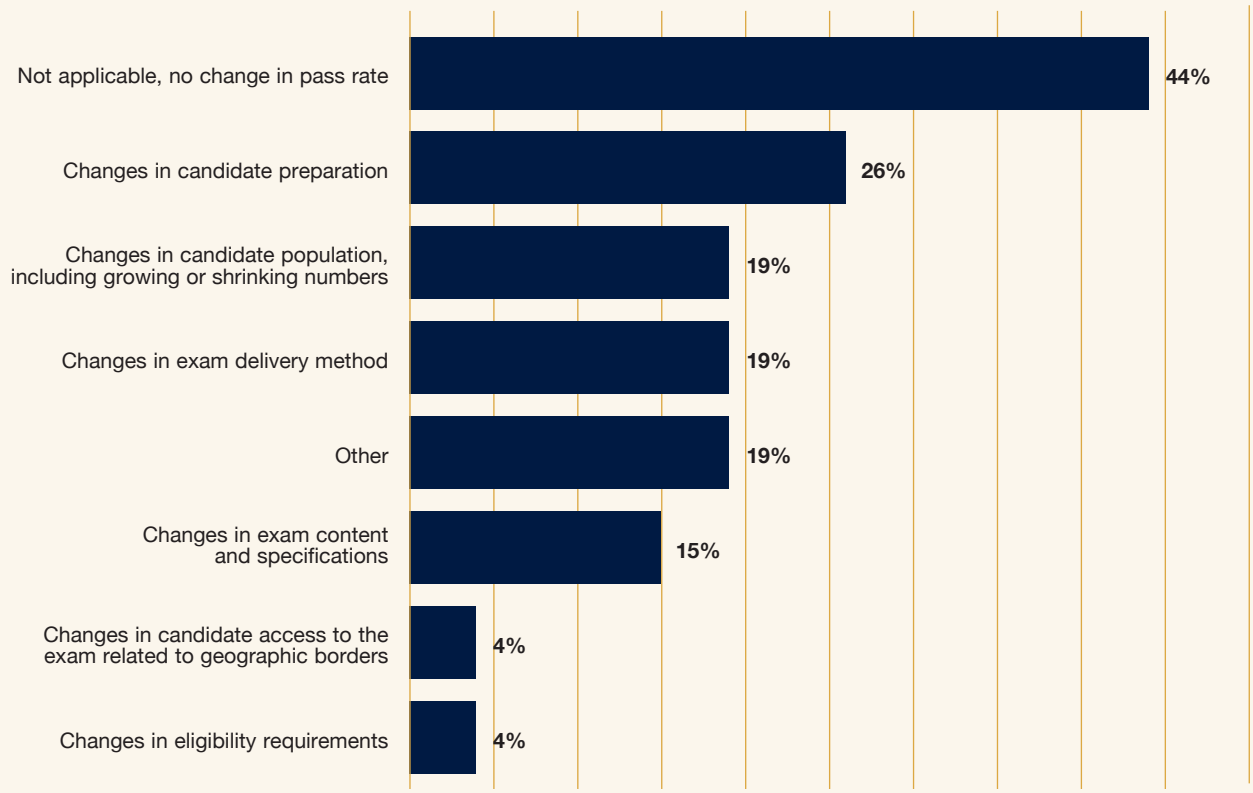
Results were somewhat similar when respondents were asked if the pass rate changed from before the pandemic to the present. Twenty-nine percent indicated that the current pass rate is lower now than before the pandemic; 15% indicated that the current pass rate is higher now than before the pandemic; and 30% indicated that the pass rate is unchanged—so slightly more have either a lower or higher pass rate now than before the pandemic. The same 26% indicated that this did not apply to them.

Figure 7. Pass Rate Change Compared to the Present



Recent CLEAR Quick Poll Results

Figure 8. Reason(s) for Change in Pass Rate



Interestingly, when offered various options to explain the reasons for changes in the pass rate, the same reason or reasons were provided regardless of whether pass rates increased or decreased, or when pass rates changed relative to before the pandemic. Of the respondents who saw a change in the pass rate, the most commonly cited reason was that there were changes in candidate preparation (26%, n = 7). Five respondents each (19%) selected changes in candidate population, including growing or shrinking numbers; changes in exam delivery methods; or some reason other than the provided options as the reason for pass rate changes. However, changes in the exam content and specifications (15%, n = 4) were more likely to be associated with a decrease rather than an increase in the pass rate. This factor may or may not be related to the pandemic and might be more related to typical exam update processes. Whether or not changes to exam content might have reflected new material related to the pandemic was not explored.

Of the respondents who provided other explanations for changes in the pass rate, all but one reported lower pass rates at the height of the pandemic and from before the pandemic to the present. Their written responses were as follows:

- Decreases in pass rate.
  - Anti-testing opinions in the media, test optional for admissions and in-school.
  - Changes in how curriculum is taught, which have impacted learning like modality and reduction of clinical hours.
  - Changes in quality/delivery of education during pandemic.
  - Questioning if education, practicum have prepared students for the exam.
- Increase in pass rate at the present; unchanged pass rate during the pandemic.
  - Changes in pre-license course delivery method.

Respondents from the U.S. and Canada had very similar patterns when describing whether and how pass rates changed during the pandemic, and from before the pandemic to the present, with two exceptions. Three U.S. respondents reported that pass rates decreased during the height of the pandemic but are higher now than before the pandemic; two other U.S. respondents reported that pass rates increased during the height of the pandemic but were either lower or the same as those before the pandemic. No Canadian organization reported either of these patterns—when pass rates decreased, increased, or stayed the same from before the pandemic to its height, this change continues to the present.

When reviewing the results of this *Quick Poll,* it is important to keep in mind that the number of respondents was relatively low, so it may not be possible to generalize whether these data represent a new normal in pass rates.

PROMETRIC

FUTURES START HERE.®

Develop your assessments faster, delivering with unlimited reach and scale, while empowering your candidates to reach their full potential - all from a single platform with top-rated security, powerful integrations, and 30 years of industry leading expertise.

Develop

Prepare

Assess

Ascend

Finetune Generate

The ultimate authoring tool for assessment items and content.

Increase the productivity, efficiency, and creativity of authors, enabling them to create larger, more diverse item banks in a fraction of the time, while securely harnessing the power of Large Language Models (LLMs) patent pending AI technology with Finetune Generate®.

Prometric Boost

Personalized exam preparation designed to identify and close knowledge gaps.

Increase knowledge retention by offering personalized practice exams that dynamically adjust to help candidates maximize their potential for success, leveraging the latest advancements in AI, while digitizing published neuroscience research to drive engagement.

Multi-Modality Delivery

A new standard in remote assessment, paired with traditional in-center testing, built to provide flexible, equitable, and secure assessments.

Provide a consistent testing experience regardless of in-center or remote modalities. We've worked with experts from across the globe, we created a new approach to using AI in remote assessments. Leveraging advanced AI to flag behavioral anomalies such as facial, object, and screen detection, paired with options ranging from fully automated AI, to AI-assisted live proctoring, all to provide candidates a seamless testing experience regardless of testing location and proctoring method.

Achievement Ecosystem

Open architecture designed to integrate with providers spanning the entire assessment journey.

Suit the individual needs of your program by seamlessly integrating from our library of industry partnerships and integrations, or bringing your own. With options for performance based-testing, skills-based labs, digital badging, and more, we'll help tailor your unique program along every step of the exam development and delivery process.

Learn how a Prometric partnership can benefit your assessment program.

LEARN MORE >

Ethics & Disciplinary Programs: Mitigating Risk & Minimizing Exposure

RICHARD BAR, JD, President, GKG Law, P.C.

CHRISTINE D. NIERO, Ph.D., Vice President, Professional Testing, Inc.

ANN WITHERSPOON, Compliance Officer, Society of Certified Senior Advisors (SCSA)

JOHN ZARIAN, JD, MFin, General Counsel, CCO Certification (CCO)

With the increased visibility of credentials, especially those that are high stakes, designing and administering an effective ethics and disciplinary program is of key importance to assuring fairness, due process, and legal compliance. Credentialing bodies routinely develop and manage such programs as a condition of accreditation and sound business practices.

However, managing an ethics and disciplinary program gives rise to a number of unique or heightened risks. How can credentialing organizations mitigate these associated risks? A good place to start is to consider which program assets are most valuable, examine program vulnerabilities, and determine which areas need protection. Understanding the risks associated with administering ethics and disciplinary programs will enable credentialing personnel to establish proper frameworks that minimize exposure to their organizations.

In this article, we identify 10 areas that should be of particular concern to any credentialing body as it seeks to mitigate the risks associated with an ethics and disciplinary program.

1. Intellectual Property

Ethics and disciplinary matters involving candidates and certificants are often associated with the theft or attempted theft of intellectual property. These matters strike at the heart of a credentialing body's most critical assets: (i) credentials; (ii) trademarks and certification marks; (iii) exam question databases; and (iv) examination forms.

A credentialing body can mitigate its risk by ensuring that intellectual property is positioned for maximum protection in the event of an attempted or actual breach. If intellectual property is not properly protected, the organization could lose significant value. Moreover, proper protection will strengthen any administrative or legal action taken against infringers.

As a threshold matter, require your volunteers and vendors to sign strong confidentiality agreements, especially those who have access to exams. There should be no exceptions. You and your vendors should also have strong exam security protocols. Evaluate the protocols regularly and compare them to best practices and technological advances. Credentialing bodies should give limited access to their exam questions on a need-to-know basis only.

As a general rule, exam questions and exams should be registered with the U.S. Copyright Office (or the equivalent in other jurisdictions). If you register your exam questions and exams, you will be in a much better position in the event of an infringement, including the ability to sue in federal court and the availability of statutory damages. In addition, certification marks and other trademarks should be registered with the U.S. Patent and Trademark Office. The registration process can be expensive; however, it will provide valuable enforcement rights and remedies that may not otherwise exist.

These best practices will support the administration of an ethics and disciplinary program, minimizing the organization's intellectual property risk.

2. Contractual Agreements

Contracts are an excellent tool for preventing conflicts and mitigating the risk of administering an ethics and disciplinary program. They furnish a clear record and, in the event of a dispute, provide very strong legal protection for the organization. In the press of work, it may be tempting simply to recycle shopworn agreement templates or forms. However, the time invested in drafting strong agreements will pay great dividends in risk mitigation and serves as an organizational tool to think through potential risks.

In reference to ethics and discipline, there are many potential terms that can be included in a program participant agreement. For example, the candidates and certificants can agree upfront that they consent to the ethics and disciplinary program and any potential sanctions imposed thereby. Some organizations even have certificants agree in advance that their names may be posted online in the event of a revocation or other sanctions. If the organization's processes involve video recordings, the participant agreement can address privacy and similar issues.

With the advent of forensic auditing and the increased use of forensic evidence in detecting irregularities, the question arises whether the organization may take action based on forensic evidence alone. Consider obtaining consent for this in the participant agreement.

In recent years, most organizations have moved from paper agreements to agreements that are "signed" virtually online. This framework gives the organization far greater ability to make changes to the agreement, and program participants can be required to sign the agreement every time they take an exam or otherwise access their account, capturing changes to the agreement.

Importantly, while digitally mediated agreements with electronic signatures are binding, they do involve special considerations. For example, the contract itself must be easily accessible and readable (e.g., via hyperlink), and affirmative consent must be clearly manifested. In addition, credentialing bodies should maintain back-end records and version control so they can "prove" the terms of the agreement in the event of a dispute.

In sum, a valid contract evidencing consent is a powerful legal tool. Accordingly, binding legal agreements are a critical component of risk mitigation for any testing organization.

3. Litigation Holds

Commercial disputes can spark suddenly, but they can also escalate slowly until they reach a boiling point. These situations are fraught with risk, in part, because the very prospect of litigation triggers legal duties that may have an impact on a subsequent lawsuit.

Under the law, a party is required to preserve evidence when it knows or should know that the evidence is likely to be relevant in pending or future litigation. The breach of this duty can result in sanctions, ranging in severity and potentially including the entry of judgment against a party.

To illustrate, consider an exchange of internal emails relating to a dispute with an outside party. The communications clearly mention the potential for litigation; however, the matter goes dormant for nearly a year,

and the emails are deleted in the ordinary course of business. If a lawsuit is then filed, the deletion of emails by the organization may be deemed "spoliation" and result in substantial penalties up to and including the dismissal of legal claims.

To mitigate such risks, organizations should adopt a "litigation hold" policy. When litigation is *reasonably* anticipated (i.e., when you know there is a credible threat that the dispute will lead to litigation), the organization should have a plan for taking "reasonable actions" to preserve information that is relevant to the dispute. This may involve the identification of key records and custodians, partial suspension of standard deletion protocols, and communication of a litigation hold notice to some or all employees. Of course, these actions should be documented.

If the organization acts reasonably and in good faith in these cases, even if documents are inadvertently lost or destroyed, it can avoid subsequent sanctions for spoliation of evidence.

4. Reputation Management

Reputation management is an important part of mitigating risk for any credentialing body. Having thorough procedures in place related to identifying non-compliance with requirements reduces the liability of the credentialing body and is essential for maintaining trust, credibility, and respect within the industry or profession. These considerations apply with special force to managing an ethics and disciplinary program given the potential impact on program participants.

Ensure you clearly communicate all standards and rules and that all parties understand and agree that they must abide by the requirements of your certifications. Require people to specifically agree that they will not bring the organization or the credential into disrepute. Make it clear that the credentialing body has decision-making power related to reputation.

If warranted for your industry or profession, credentialing bodies should invest in monitoring procedures and/or platforms to aid in awareness of any nefarious activity. Wrongful use reviews can be implemented to ensure the correct use of certification marks online and in business practices.

In general, organizations should maintain transparency, consistency, and impartiality in all matters to highlight their commitment to the integrity of credentials and the organization. All interested parties should know the certification body's requirements and procedures related to the review and enforcement of behavior standards.

Finally, credentialing bodies should do what they say they are going to do. People will pay attention to how you are enforcing your code of ethics, following your stated procedures, providing due process, and handling any negative situations. Ultimately, this inures to the broad benefit and credibility of an ethics and disciplinary program.

5. Investigative Process

An organization's administrative procedures are the foundation of due process and fairness in an ethics and disciplinary program, including any investigation. Administrative policies and procedures should be clearly written and communicated, and those procedures should be followed. Staff members and committee members should refer to those policies frequently and do their best to interpret and apply the procedures consistently. Adherence to these practices will reduce the risks inherent in deviating from an organization's written policies and procedures.

To be sure, there is a balance to be struck, but to mitigate risk, organizations should err on the side of increased transparency and communication. Documentation is an important part of the process. Even decisions not to open an investigation should be documented because they reflect the judicious application of a credentialing body's process.

The importance of clear, consistent, and complete documentation of all steps in an investigative process cannot be overstated relative to the mitigation of risk to a credentialing body. Have reliable systems in place to track and store information, dates, communications, status updates, and documentation. Keep detailed records and document all relevant information.

Even during investigation and screening, consider having a team of personnel make decisions based on consensus. It is much more difficult to challenge the decision of a team than that of a single individual (who may be accused of bias). When decisions are made concerning the potential revocation of credentials, involve outside subject matter experts; their business judgment will carry significant weight. In general, allow for appeals to be made liberally.

6. Due Process

Whereas credentialing bodies are subject matter experts and understand what acceptable behavior is, courts understand process and often focus on fairness. Consequently, courts generally do not like to get involved in the merits of ethics committee matters, but they will step in when they feel that a credentialing body's rules and procedures are unfair or are not being followed. Courts are much more likely to defer to administrative procedures that have the hallmarks of fairness.

Thus, an organization's ethics and disciplinary programs must pay close attention to fairness and due process, which is the minimum needed so that all parties to an ethics investigation are treated fairly. Credentialing bodies should evaluate their processes and ensure that policies and procedures are fair, reasonable, and aligned with best practices. Rules and procedures should be reviewed regularly.

As suggested above, you should include a statement in the application and the program participant agreement confirming that the applicant agrees to comply with the code of ethics and the rules and procedures of the ethics and disciplinary program.

Generally, credentialing bodies should refrain from acting on anonymous complaints unless they can be separately verified using reliable sources, such as other people, newspapers, court records, and so on. An ethics committee should initiate action based on verifiable information. A standard complaint form should be utilized to aid in this process.

Furthermore, rules and procedures should include timelines (for example, giving a party 30 days to respond). These may be stated as guidelines, or subject to exceptions, but including timelines demonstrates an organization's commitment to due process.

In addition, ethics and disciplinary programs should treat similarly situated people in similar ways. A good database can help track ethics matters so that decisions are consistent with prior comparable situations. There should be meaningful distinctions for outcomes to differ from past precedents.

7. Self-Reported Information Versus Third-Party Verification

Credentialing bodies have a due diligence duty to determine the appropriate requirements related to disclosure of information for the holders of their credentials. Assessing the risk level, regulatory requirements, compliance standards, costs, and the nature of the information you are dealing with will help you determine how to structure any disclosure and background review procedures. This process, in turn, impacts the management of an ethics and disciplinary program.

In certain industries, self-reported information may be appropriate and be all that is necessary. In such cases, make it easy for people to update self-reported disclosure information on an ongoing basis so that, if something changes, they can quickly communicate that to you. In other industries or situations, in addition to self-reported data, it is prudent to include some form of third-party verification of candidate history, such as third-party background checks to verify a candidate's employment, education, credentials, and criminal history.

Credentialing bodies should understand privacy and other laws related to third-party verifications to ensure they comply with all regulations. Depending on the purpose of a background check, including whether your credential is required for working in a particular industry or profession, some jurisdictions may not allow the use of certain types of background screening. Consider using an accredited third-party vendor that understands your industry and profession's requirements and all applicable state laws surrounding third-party verification, as regulations can vary widely in what you can and cannot do related to the permissible purpose of a consumer check.

As appropriate, credentialing bodies should analyze the costs of any third-party reviews and know their available resources to determine the appropriate level of outside verification for the organization. Understand what you are getting for your money for things like criminal background checks. A strong background check needs to include layers of information at the local, state, and national levels, and there can be a wide range of additional costs added to checks at the county and state levels for information lookup and review.

In addition, decide whether you will pass any third-party verification costs on to candidates, or if the costs will be included as part of your application and/or renewal fees. If you will be covering the fees, understand the variability of certain types of background checks depending on a candidate's location, and determine your fees accordingly.

Finally, have a plan on how to handle third-party review delays, as these can occur when dealing with state and local jurisdiction procedures and limits.

8. When to Take Action

As noted above, ethics committees should not commence action based on anonymous complaints alone unless the information supporting the complaints is independently verified. Ethics committees should refrain from making findings until after they have considered all evidence and completed their investigations. However, committees may take preliminary action pending the outcome of their investigations (such as an interim suspension) if the rules and procedures allow this for the benefit of public protection.

Should the subject of an ethics complaint not respond to the complaint or otherwise not participate in the investigation, the ethics committee should continue the investigation and may consider the allegations set forth in the complaint to be true if the allegations are reasonable and the rules and procedures permit doing so. Furthermore, an ethics committee should not stop its investigation if an accused party surrenders the credential. It is important to have final decisions on all investigations documented in case the accused

resurfaces in the future. However, the committee should have the discretion to terminate an investigation if the complainant ceases to participate.

Finally, the credentialing board should monitor compliance with issued sanctions as part of its ethics and disciplinary program. If an individual's credential has been suspended or revoked, the organization should take steps to ensure that the accused is not continuing to use the credential.

9. Sanctions and Limits of Enforcement

Organizations may identify a range of potential sanctions to be imposed in ethics and disciplinary matters. These sanctions can range from revocation of a credential (permanently or for a fixed period) to suspension (either for a fixed period or until a condition such as retraining is satisfied), to a reprimand or formal warning. However, it should be recognized that any sanction may carry a stigma because sanctions are generally regarded as punitive consequences for a breach of the code of ethics or as "penalties" for misconduct.

For this reason, it is valuable to distinguish between ethics and disciplinary sanctions and the invalidation of examination results. The better practice is to outline these two concepts in your policies and procedures. Thus, your policies and agreements may stipulate that exams can be invalidated even without direct evidence of misconduct, based solely on forensic evidence, whenever the organization determines that it lacks confidence in the validity of one or more exams. This decision can be made on technical and practical considerations alone and may result in a companion ethics and disciplinary matter, depending on the circumstances. Differentiating between ethics and disciplinary sanctions and exam invalidation allows for a more clearly defined process and gives the organization a wider range of options in complex cases.

As noted above, when sanctions are imposed, ethics committees must treat comparably situated parties similarly. Organizations will be subject to intense scrutiny if their findings and sanctions differ meaningfully from similar past situations. This is especially true when the credential is a requirement to work.

Upon completion of an investigation, ethics committees should issue timely, rational, and historically consistent decisions. The rules and procedures should allow ethics committees to communicate decisions on the organizations' websites and to interested parties, including employers, relevant organizations, and state agencies.

10. Mitigating Risk in High Stakes Certifications

The administration of ethics and disciplinary programs creates unique risks for credentialing bodies. This article has sought to provide strategies for identifying, avoiding, and reducing those inherent risks and the related exposure for organizations.

In a modern business environment, the identification and reduction of risk is an important mindset, and risk mitigation is a critical element of success. This is certainly true when it comes to potential claims arising from the administration of ethics and disciplinary programs.

Ultimately, the best course of action is to be aware of the associated risks and not hesitate to involve senior management in appropriate cases. Indeed, organizations should build heightened or elevated reviews into their procedures to be triggered under proper circumstances. In some cases, the early involvement of additional personnel, senior management, or the board of directors may be the best way to deescalate situations and mitigate risk.

Practical Issues in Standard Setting Part I: An Approach for Incorporating Policy Considerations with Method

GREGORY J. CIZEK, Ph.D.
Guy B. Phillips Distinguished Professor Emeritus
University of North Carolina-Chapel Hill

LORIN MUELLER, Ph.D.
Managing Director of Assessment
Federation of State Boards of Physical Therapy

The term *standard setting* is used to refer to the process of establishing one or more performance levels or "cut scores"—the scores required to pass an examination. The activity of standard setting is perhaps the most consequential and visible component of the test development, administration, and reporting process. A variety of methods exist for setting standards (see, e.g., Cizek & Bunch, 2006, for descriptions of some of the most common approaches) with the Angoff (1971) and Bookmark (2001) approaches being among the most common of what Jaeger (1989) called *test-centered* methods.

Background

The process of implementing a test-centered standard setting procedure commonly involves a group of highly-qualified participants—also referred to as panelists, judges, or subject matter experts (SMEs)—who are asked to provide judgments about the items or tasks used in a testing program. The judgments are made regarding the estimated performance of a *just-qualified candidate* (JQC), where the JQC is conceptualized as the hypothetical examinee who possesses just enough of the knowledge, skills, and abilities (KSAs) necessary to pass the examination (e.g., the KSAs necessary for safe and effective entry-level practice in a profession).

When an Angoff variation is used, panelists make judgments for each item in a test in the form of the proportion of JQCs they believe would answer each item correctly. When a Bookmark approach is used, panelists identify the item in a sequence of test items ordered from least to most difficult, where the probability of a JQC answering an item correctly drops below a pre-established value—the *response probability* (RP) value, often set at .50 or .67. Although standard setting panelists are often diverse as to the regions, practice specializations, sex, ethnicities, or other characteristics deemed to be relevant to their participation, it is equally important that they are well-qualified with respect to their familiarity with the candidate population, the job demands of the field, and their expertise and experience in the area for which the performance standards will be set.

At this point, it is relevant to note that the subject matter experts (SMEs) selected as standard setting panelists do not actually *set* the cut score for a test. It is perhaps more accurate to refer to the panelists as "standard recommenders" than "standard setters." Rather, it is the entity responsible for the credentialing program that has the authority to establish the cut score that will be used on an examination. That entity—typically the board of directors of the credentialing agency—may well accept the passing score recommendation put forth by the standard setting panelists, but they may also choose to modify or adjust the recommendation.

The credentialing board has the final, legitimate authority to set any cut score that will be used along with distinct characteristics and perspectives that differ from the SMEs who serve on standard setting panels. Whereas the SMEs who serve as standard setting panelists are recruited because of their high degree of content area knowledge and skill, a credentialing board can include both members who are experts in a field and members with other relevant areas of expertise, such as law, business and finance, or consumer considerations. In contrast

to standard setting panelists—who are typically directed to focus primarily on item/task content demands when making their judgments and *not* to consider policy implications—credentialing boards must also consider the perspectives from various stakeholder groups (e.g., jurisdictions, training programs), the organization's mission (e.g., public protection), the relative costs of false-negative and false-positive credentialing decisions, and other risk-mitigation issues alongside the content experts' recommendations.

The Problem

In many contemporary standard setting applications, there is at least some attempt to keep the content-grounded and policy-based standard setting perspectives distinct. In the common, two-step process, a group of SMEs first derives a cut score recommendation based on their judgments about test content. Next, the policy body takes that recommendation into account within the constellation of other factors they must consider in making a final cut score determination. Contemporary standard setting has blurred these roles somewhat. For example, SMEs are routinely provided with impact data regarding the likely pass/fail percentages that would result from the application of the group's overall recommended cut score. Such information is intended to inform their deliberations and result in more "reasonable" recommendations than content-only referenced judgments might suggest. Similarly, credentialing boards entertain both policy considerations and the content-based recommendations of SME panels.

A dilemma arises when panelists (appropriately) make their content-grounded judgments about the likely performance of JQCs on the various items and tasks that comprise a test but end up recommending (even after iterations or "rounds" of review and discussion) a cut score that the entity responsible for actually establishing the operational cut score deems to be unacceptable based on the various policy considerations. It is surely an awkward situation facing the credentialing entity: they want to honor the efforts and expertise of the SMEs they empaneled to make a recommendation, but they do not wish to accept a cut score recommendation that they believe is inappropriate; neither do they want to make an ad hoc, arbitrary adjustment.

In a separate article, we will address defensible options that could be considered for making an adjustment to a panel's recommended cut score. However, in this article, we address what we believe is arguably preferable: implementing a process that potentially avoids the necessity of making an adjustment.

The Context

In addition to the foundational works describing content-referenced standard setting procedures, such as the Angoff, Bookmark, and other approaches, psychometric literature contains descriptions of methods that rely relatively more on policymakers' perspectives and relatively less on content judgments. The first generation of such approaches is evident in methodologies such as those proposed by Beuk (1984), Hofstee (1983), and de Gruijter (1985) that attempt to reconcile content-based recommendations with practical concerns.

More recently, McClarty et al. (2013) developed and evaluated what they called an evidence-based standard setting method; Haertel (2002) proposed what he termed the "briefing book" method. Both methods provide ancillary information to a panel of policymakers who either directly identify a passing score (or range of acceptable passing scores) or provide policy guidance to a subsequent content-specialist panel. The ancillary information may consist of pass rates on a related measure (e.g., an in-training examination), comparisons of pass rates of relevant groups (e.g., first-time test takers, foreign-trained test takers), or other applicable data (e.g., GPAs, clinical ratings, or in-service performance of contrasting groups). Such approaches may be desirable when a policy-making body has strong beliefs about the performance standard; when it is desired to prioritize policy perspectives over content judgments; and when dependable, valid ancillary information is readily available.

One specific dilemma that sometimes arises is when a panel's recommended cut score would have an impact on pass rates that is deemed to be either overly stringent or overly lenient by the credentialing entity. This situation can be addressed with modest success during the rounds of the standard setting workshop itself. Between rounds of ratings, facilitators can provide estimates of preliminary impact, informing panelists of the projected pass/fail rates should panelists' recommendations be implemented and whether those rates deviate substantially from what has been historically observed.

There are at least four concerns related to such an approach. First, after encouraging panelists to focus exclusively on the subject matter, workshop facilitators are placed in the potentially difficult position of encouraging the SMEs to depart somewhat from exclusively content-based judgments. Second, some panelists may become anchored to their initial rating and feel the need to defend their judgment rather than consider new information and adjust accordingly, especially if their rating represented one of the extreme values in the initial round. Third, in many standard setting methodologies, it is not intuitive to panelists exactly *how* to revise their judgments in a subsequent round of ratings to shift their recommended cut score in a given direction. Finally, in at least some situations, panelists might ask: If the credentialing entity has a certain cut score range in mind, why should they have to guess what that range is, as opposed to being provided with such policy guidance upfront?

The Historical Bookmark Approach

In response to that reasonable question, we say, "Amen."

As opposed to cajoling panelists to move a cut score in a given direction during a standard setting workshop, it seems preferable to provide panelists with whatever relevant information exists to guide their effort. Similar to the contemporary, evidence-based approaches cited earlier, at least one kind of valid, ancillary information that can be provided to panelists is policy guidance related to acceptable pass/fail rates.

We have implemented what we tentatively refer to as a Historical Bookmark approach to standard setting for four different health-related credentialing contexts. In the following paragraphs, we provide a brief description of the approach and results from an application of the method.

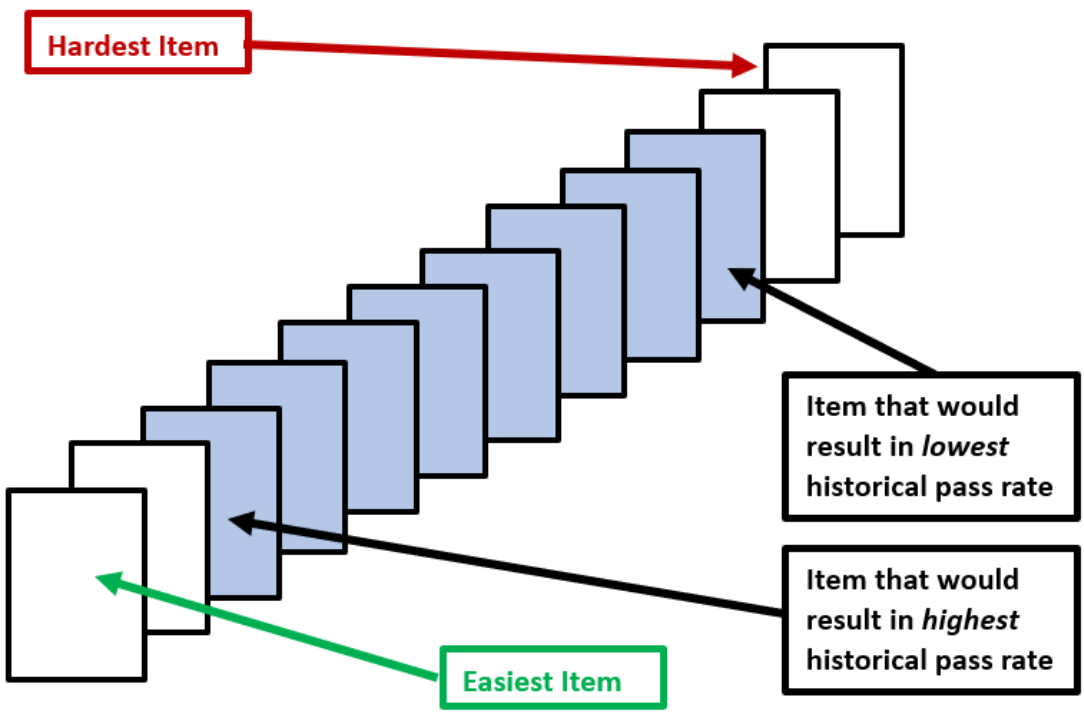
In each implementation of the *Historical Bookmark* approach, a standard setting study was deemed desirable because there had been modifications to the tests' content outlines to comport with the results of regular job analysis surveys. Because the modifications were judged to be modest and no other substantial changes impacted the field (e.g., regulatory changes, training program changes, licensure requirements, or candidate population shifts), the credentialing entity held the *a priori* view that pass/fail rates on the new assessments should be fairly consistent with historically observed rates. Further, given the significant shift in pass rates seen in many testing programs following the COVID-19 pandemic, the organization wanted to manage the risk of further lowering pass rates (without sufficient content-based justifications) by providing more information upfront in the process.

In each case, a Bookmark standard setting method was identified as a suitable choice based on assessment item/task formats, administration mode, ease of use for panelists, and other factors. A 2.5-day standard setting workshop was conducted in June 2023. A total of 18 panelists participated; panelists were recruited to be diverse and representative of background characteristics deemed to be important to the profession. All core elements of a Bookmark methodology were included. For example, panelists took and self-scored a test form; training and practice in the method were provided; panelists developed descriptions of the JQC; three rounds of ratings, review, discussion, and feedback (normative, impact) were included; and evaluation surveys were administered at key junctures.

A traditional ordered item booklet (OIB) was assembled, but with two novel elements. The first novel element followed from the credentialing entity's examination of its pass/fail rates for the past decade to determine the highest and lowest pass/fail rates over that period. (Any reasonable time interval could be chosen that reflects a period where pass rates could be considered relevant to the current assessment program.) The OIB was marked to indicate the item in the booklet that, if selected by a panelist, would match the highest historical pass rate; a second item was marked later in the booklet that corresponded to the lowest historical pass rate. Figure 1 provides a simplified illustration of how this information was presented to panelists.

During the standard setting workshop, these indications were explained to panelists, and they were informed as to the policy rationale for these "guideposts." Importantly, panelists were also explicitly instructed *not* to consider the marked pages as limits. That is, after their extensive discussions to develop a specific description of the hypothetical JQC, panelists were directed to make their individual cut score recommendations based on how a JQC would be expected to perform based on the content of the assessment items/tasks, not limited to the range of the indicated OIB pages. If, while reviewing the items and tasks in the OIB, a panelist judged that the appropriate bookmark was outside the range, the panelist should make that recommendation.[1]

Figure 1. Simplified Illustration of "Guideposts" in OIB



[1] Although we report here on the incorporation of historical data to guide standard setting in the context of a bookmark procedure, we are currently working to develop an analogous approach for use when an Angoff approach or similar item-based method is selected.

The second novel element involved supplementing the OIB around the range of the marked guidepost pages (i.e., by adding items just before, inside of, and immediately after the range). In designing the Historical Bookmark approach, it was determined that items within the range should reflect a broad spectrum of test characteristics. For example, the range should include items from all content categories spanning diverse body systems and patient characteristics and incorporate items with associated graphics, video stimuli, and other surface item characteristics. Supplementing the OIB in this way allowed the facilitators to make the OIB the same length as an operational form by replacing unscored pretest items with scored items.

One additional benefit of supplementing the marked OIB range was that it provided panelists with a greater breadth of content within the range of likely JQC performance. Another was that it provided increased flexibility in terms of panelists' bookmark placements and helped to avoid situations that can arise in traditional Bookmark applications where a one-page movement of a bookmark can result in a substantial cut score change (when the scale locations of the adjacent items are far apart) or no cut score change (when the scale locations of the adjacent items are identical or nearly so). This second modification of the typical Bookmark approach resulted in a 28-page range between the indicated guidepost items. This yielded a range of approximately 15% of the OIB, which was judged to have proven reasonable leeway for the SMEs to adjust the passing standard while providing enough content representation for making fine-grained distinctions.

Results

Two data sources provide relevant information to assess the results of our modification of the bookmark standard setting procedure: descriptive statistics on panelists' ratings and qualitative information gathered from workshop surveys.

Table 1 provides the round-by-round means and standard deviations of panelists' individual cut score recommendations. As can be seen in the table, and as is typical of most standard setting procedures, panelists' overall judgments generally stabilized across rounds and decreased in variability.

Table 1. Means, Standard Deviations, and Estimated Cut Score by Round

Round	Mean rating across panelists	Standard deviation of panelists' ratings	Estimated cut score (θ)
1	109.2	11.8	0.88
2	105.2	7.2	0.86
3 (Final)	105.6	4.7	0.86

Qualitative information also provided supportive evidence. Table 2 provides selected results[2] taken from two of the evaluation surveys; results for items A–D are from a readiness survey administered following training; items E–H were administered as part of the final workshop evaluation. Overall, panelists reported a high degree of understanding of the purpose of the standard setting meeting and the historical bookmarking procedure. At the conclusion of the workshop, they reported being confident in their work, their use of the historical bookmarking procedure, and the cut score recommendation resulting from the procedure.

[2] Results from survey items not directly relevant to the Historical Bookmark procedure (e.g., time allocated for discussion, meeting logistics, etc.) were equally positive but have been omitted here.

Table 2. Selected Evaluation Results from Standard Setting Workshop

Item	Item content and response options	Frequency
A	Clarity of meeting purpose	
	Very clear	15
	Clear	3
	Somewhat clear	0
	Not clear	0
B	Description of process	
	Very clear	13
	Clear	5
	Somewhat clear	0
	Not clear	0
C	Confidence in understanding of Bookmark task	
	Very Confident	12
	Confident	4
	Somewhat Confident	2
	Not Confident	0
D	The purpose of highest and lowest indicators in OIB is:	
	to ensure bookmarks are not placed outside that area	0
	to indicate items potentially confusing to a JQC	1
	to provide guidance on the reasonableness of potential cut scores	17
	to ramp up the cut score and ensure that fewer candidates pass	0
E	Overall comfort with bookmark judgment task	
	Very Comfortable	16
	Comfortable	2
	Somewhat Comfortable	0
	Not Comfortable	0
F	Overall confidence using bookmark method	
	Very Confident	18
	Confident	0
	Somewhat Confident	0
	Not Confident	0
G	Overall confidence that final result is appropriate and defensible	
	Very Confident	13
	Confident	4
	Somewhat Confident	1
	Not Confident	0
H	Opinion of group's recommended passing standard	
	About Right	16
	Too High	0
	Too Low	2

Table 3 provides a sample of responses provided to a final, open-ended survey question that permitted panelists to comment on any aspect of the workshop. Only comments bearing on the standard setting procedure itself and results are presented in the table; i.e., responses related to meeting logistics, food service, and other unrelated aspects have been omitted. Of note, only one participant (Respondent G) indicated that they would have liked to make a first bookmark placement without knowledge of the historical guide rails. Panelists did not seem to be limited by the guardrails however, as six panelists placed a bookmark beyond the last page of the guardrails.

Table 3. Open-ended Responses for Final Workshop Evaluation

Respondent	Responses to Final Open-Ended Question
A	"Very informative process. The inclusion of all the data in the build up to the final judgements was very necessary."
B	"Great process. I think knowing a bit more about the actual process prior to the weekend may have made the process smoother."
C	"I felt like this weekend flowed really well."
D	"The whole process was exceptional, and all of the presenters and participants were top notch. Great weekend and 5 star all the way around! Thanks!!"
E	"I think the passing standard is about right based on the data we received with the suggested content that will be added."
F	"I would recommend that it be moved to a higher number question as the cut off score, but not by much."
G	"I appreciate the information we were given between each round of taking the OIB. I do wish that the first round could have been completed independent of the pre-set bookmarks of former passing scores. I feel that this instantly influenced my rating, although my score did vary over the subsequent rounds."
H	"I feel the standard should be closer to [omitted] as the expectations for entry level are increasing."
I	"Thank you."
J	"Excellent job"
K	"Thank you for guiding us through this process patiently. This was a very pleasant experience and a learning experience. This was a better process than the standard setting committees I have been on with [omitted]."
L	"This was a valuable process to participate in."

Note. Responses are in answer to the following prompt: "Please use the space below to provide any additional comments or suggestions about the standard setting process."

Conclusions

Standard setting often involves balancing the desires of SMEs to be diligent in recommending a cut score that they can confidently endorse as signifying competence against an appropriate, fair challenge posed by a comprehensive standardized examination of qualified candidates and the resulting impact on the candidate population. Standard setting facilitators are well familiar with the arduous task of having to walk a panel back from an unrealistic, aspirational standard to one that more reasonably comports with the abilities of the minimally qualified candidate. The Historical Bookmark Method is one way to provide panelists with more information upfront in a manner that isn't overly restrictive or manipulative.

Based on our experience with the methodology described here, panelists appreciated having some guidance on the region of the OIB within which to focus most of their deliberations. The guidance did not prevent them from making recommendations outside of that range, and they reported that they were confident in the resulting standard. Considering recent changes in pass rates seen among many testing programs, the Historical Bookmark Method provides a sound option for organizations wishing to mitigate the risk of a standard setting process diverging greatly from acceptable standards without a sufficient content-based rationale.

Author Bios

Gregory J. Cizek is Guy B. Phillips Distinguished Professor Emeritus at the University of North Carolina-Chapel Hill. His scholarly interests include validity, standard setting, and test security. He is author or editor of over 12 books and 300 articles, chapters, and presentations, including Setting Performance Standards (2001, 2012), Cheating on Tests (1999), and Validity: An Integrated Framework for Test Score Meaning and Use (2020). He has held service and leadership positions with the National Assessment Governing Board and the American Educational Research Association (AERA) and is past President of the National Council on Measurement in Education (NCME). Dr. Cizek has managed national licensure and certification programs and has served as an elected member of a local board of education. He began his career as an elementary school teacher.

Email: cizek@unc.edu

Lorin Mueller, Ph.D., is the Managing Director of Assessment at the Federation of State Boards of Physical Therapy, with expertise in psychometrics and test development. Dr. Mueller has contributed his measurement expertise to such efforts as the development of high stakes tests for the selection of advanced mathematicians, medical personnel, elementary school teachers, air-traffic controllers, and human-resource professionals. Dr. Mueller previously worked at the American Institutes for Research (AIR) as a Principal Research Scientist, has served on several test development advisory boards, and is a former president of the Personnel Testing Council-Metro Washington and a SIOP Fellow.

Email: lmueller@fsbpt.org

Reference List

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). American Council on Education.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*(2), 147–152.

Cizek, G. J., & Bunch, M. (2006). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests.* SAGE Publications.

Clauser, B. E., Kane, M., & Clauser, J. C. (2019). Examining the precision of cut scores within a generalizability theory framework: A closer look at the item effect. *Journal of Educational Measurement, 57*(2), 216–229.

de Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement, 22*(4), 263–269.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 16–22.

Hofstee, W. K. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109–127). Jossey-Bass.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher, 42*(2), 78–88.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Lawrence Erlbaum Associates.